# The testing effect applied to procedural skills

## *Using an echocardiogram simulator*

Stephan Mooibroek
s1536729
January, 2014

**Master Project Thesis**
Human-Machine Communication
University of Groningen, The Netherlands

Internal supervisor:
Dr. Fokie Cnossen (Artificial Intelligence, University of Groningen)

External supervisor:
Dr. R.A. Tio (Department of Cardiology, University Medical Center Groningen)

## Abstract

This study focuses on long-term retention of knowledge and skills required for making transthoracic echoes. Declarative knowledge and procedural skills are interconnected for many skills. Before making an echocardiogram, the practitioner needs to collect from declarative memory what echocardiogram is suitable to visualize what he is looking for. Before being able to diagnose a patient based on the echocardiogram, the practitioner needs the skill to make the appropriate echocardiogram. The aim of this study was to find the best way to achieve long-term retention of declarative knowledge and procedural skills by applying the testing effect and spacing effect. The testing effect explains that declarative memory can be improved when subjects retrieve information rather than restudying the material an equal amount of time; i.e. retesting is more effective than restudying (e.g. Abbott, 1909; Gates, 1917; Agarwal, Karpicke, Kang, Roediger III, & McDermott, 2008; Zaromb & Roediger III, 2010). The spacing effect explains that a repetition will be most beneficial 'if the material had been in storage long enough as to be just on the verge of being forgotten' (e.g. McGeoch, 1943; Banaji & Crowder, 1989; Pashler, Rohrer, Cepeda & Carpenter, 2006). The testing effect and spacing effect have been repeatedly shown on declarative knowledge; both effects have not been studied often in combination with procedural skills. Medical students have been trained on the basic anatomy and function of the heart (declarative knowledge) and received a theoretical introduction to transthoracic echocardiography. Additionally subjects were trained on making heart echoes (procedural skills). Groups with or without an interim test between the training and the final test have been compared on both their declarative performance and their procedural performance. As expected, interim testing has a beneficial effect on the long-term retention of declarative knowledge. This effect goes beyond declarative performance, as the groups that took an interim procedural test outperformed the groups that did not take an interim test. It seems the testing effect can be generalized to procedural skills.

# Acknowledgements

It has been a journey... I was offered a job at the end of 2012. Even though I did not finish my thesis yet, I decided to go for the job and in 2013 tried to combine writing my thesis with a fulltime job. (Obviously) writing took longer than hoped and expected. Now I can finally say I am very proud I am done.

First of all I would like to thank my supervisors Dr. F. Cnossen and Dr. R.A. Tio. Thank you for giving me the opportunity to start with this project. I would like to thank both for your support during the project and I really appreciate the trust you gave me. Besides that I again want to thank Dr. F. Cnossen for her support on a personal level as well.

I would like to thank all members of the Skill Center; their support and flexibility made it possible to do all the measurements and gather the data. I furthermore want to thank my parents and friends for supporting me throughout the process. Lennart Stapelkamp, the many discussions and coffee moments were both very nice and really helped me. Erik van Dijk, thanks for the time and effort you put in providing feedback on the statistical analysis and the results section. Special thanks to my best friend Sjouke Piersma; the many long talks have been very useful and your example of (finally) finishing your PhD inspired me to finish this process.

Last but not least I want to give a special thanks to my girlfriend Sandra al Saifi. You know it has been a struggle for and I now realise it has been a struggle for you as well. You have always given me unconditional support and trust; I cannot thank you enough for your role in this success!

# Table of Contents

# Introduction

An echocardiogram is a test that uses sound waves to create a moving picture of the heart. There are several reasons why a doctor could decide to perform an echocardiogram, such as assessing the heart function or checking for diseases. In order to complete a standardized examination the echocardiographer is expected to identify normal and abnormal structures and assess heart functions (Bose, et al., 2009).

There are two types of echocardiograms: transthoracic (TTE) and transesophageal (TEE). TTE is a non-invasive method through the thorax often used to assess left ventricular function (Hillis & Bloomfield, 2005). The TEE is an invasive method where the probe is manoeuvred through the oesophagus. For the TTE a transducer (or probe) is placed on the body and manoeuvred in the required position to obtain an image. Making clear echoes can be challenging by the fact that lungs, ribs and/or other body tissue may interfere.

The present study focuses on long-term retention of the skills and knowledge required for making a TTE echo. Optimal training is crucial for the performance in life threatening situations. The faster and more accurate an echo can be made, and the better someone is at interpreting that echo, the less risk there is for the patient. Currently in order to train intensivists in performing and interpreting an echocardiogram, intensivists go through a one-day training. This one-day training consists of a theoretical instruction, followed by a practice session where the skill of making an echo is practiced under supervision on an echocardiogram simulator. This training is concluded with a short multiple-choice test on the anatomy and function of the heart and on interpreting echoes.

Acquiring and maintaining a skill can be done on a simulator. According to Issenberg, McGaghie, Petrusa, Gordon and Scalese (2005) research 'on the use and effectiveness of simulation technology in medical education is scattered, inconsistent and varies widely in methodological rigor and substantive focus'. However, the use of simulator in acquiring and maintaining skills seems promising. Boet et al (2011) conducted a study where subjects performed a cricothyroidotomy (emergency airway puncture) on a simulator. Results showed that after one training session with the simulator subjects' skills improved both on short-term (tested on the same day) and long-term (tested after 6 months and after 12 months).

Performing an echo requires both mastering the skill to manipulate the probe, knowing what to do and how to interpret an echo (The Cardiac Society of Australia and New Zealand, 2009). There are different types of knowledge humans can master, and while multiple distinctions could be made, one in particular is relevant: the difference between declarative knowledge and procedural skills; i.e. knowing 'what' versus knowing 'how' respectively. According to Anderson, Fincham and Douglas (1997) in initial problem solving one explicitly refers to examples, either from instructions or memory. After getting more practice, a set of rules forms to solve that specific problem. After forming this set of rules it is no longer necessary to access declarative information consciously; i.e. this knowledge transitions from a declarative to a procedural form.

In many skills, and also in the domain of making an echocardiogram, declarative knowledge and procedural skills are interconnected. Before making an echocardiogram, the practitioner needs to know what he is looking for and what echocardiogram is suitable. In practice, before being able to diagnose a patient based on the echocardiogram, the practitioner needs the skill to make the appropriate echocardiogram. In the present study a robust effect on learning declarative knowledge, namely the testing effect (which will be discussed in the theoretical background), has been applied on procedural knowledge. In order to test declarative knowledge and procedural skills separately, a distinction between declarative knowledge and procedural skills will be made. This distinction will be further discussed in the methods section.

To verify if knowledge is acquired (either via a simulator or in a realistic setting) tests can be taken. In many situations tests are taken infrequently (Roediger III & Karpicke, 2006). Typically, newly acquired knowledge is evaluated by a test at the end of a period in exam weeks. There are other ways of learning and testing, rather than a period of classical instructions and self-study with a final test, which might be more beneficial for long-term retention of knowledge and skills.

The aim of this study is to find the best way to achieve long-term retention of declarative knowledge and procedural skills by applying the so-called testing effect and spacing effect. Before explaining both the testing effect and the spacing effect in the theoretical background, first a general introduction into learning and memory will be given. By applying the testing effect and spacing effect the present study explored another way of achieving best long-term retention for both declarative knowledge and procedural skills.

# Theoretical Background

In the experiment described in this thesis subjects' performance over time (starting without any pre-knowledge) will be measured. Over time it is possible to determine learning effects and over time retention effects can be determined as well. This section provides an introduction to the subject of learning and some of the underlying mechanisms. Furthermore it provided an introduction to the involved memory systems. Finally, the two learning effects applied in this study, the testing effect and the spacing effect will be explained.

## *Learning*

Humans start learning from a very young age. Learning does not always occur in exactly the same fashion; there is a difference between what we learn incidentally as an unconscious process, and what we learn intentionally. An example of incidental learning would be social learning of children participating in for instance sports. While the social skills are not the main goal of those activities, children do encounter incidental learning due to interactions (National Research Council, 2000). Intentional learning involves cognitive processes that *intentionally* aim for learning, rather than having it as an *incidental* result. While many activities have some form of incidental outcome, only few cognitive processes are done in such a way that it has a learning goal (Bereiter & Scardamalia, 1989). Examples that demonstrate intentional learning are for instance studying for a test or actively following a lecture. Learning can be described as the ability to acquire or reinforce/modify existing knowledge, skills, behaviour. There are many potential bases for human learning, such as education or training, but also for personal development. Human learning may be goal-oriented and may be aided by motivation. Humans can often be viewed as actively seeking new information; i.e. humans are goal-directed (National Research Council, 2000). There are multiple domains in which learning can occur; Benjamin Bloom (1956) suggested three domains of learning; this study focuses on two of those domains, namely the cognitive and psychomotor domain:

1. Cognitive – Factual knowledge and intellectual skills such as a mathematical calculation, etc.
2. Psychomotor – Dancing, riding a bicycle, making an echo, etc.
3. Affective – Feeling emotions, etc.

Learning has to be seen as a (on-going) process; it does not stop once a set of isolated facts has been acquired. It is a continuous process that ultimately makes relatively permanent changes in the organism (Schacter, Gilbert, & Wegner, 2011).

### Declarative learning

Learning typically shows an exponential learning curve. Ebbinghaus (1885) published the book Memory: A Contribution to Experimental Psychology. In this book he shared his findings regarding the processes of learning and forgetting. Ebbinghaus (1885) described the learning curve as follows: when learning, the steepest increase in learning occurs after the first repetition and then quickly evens out. Adding additional repetitions only adds little extra knowledge; i.e. the learning curve is exponential (Heathcote, Brown, & Mewhort, 2000), see Figure 1:

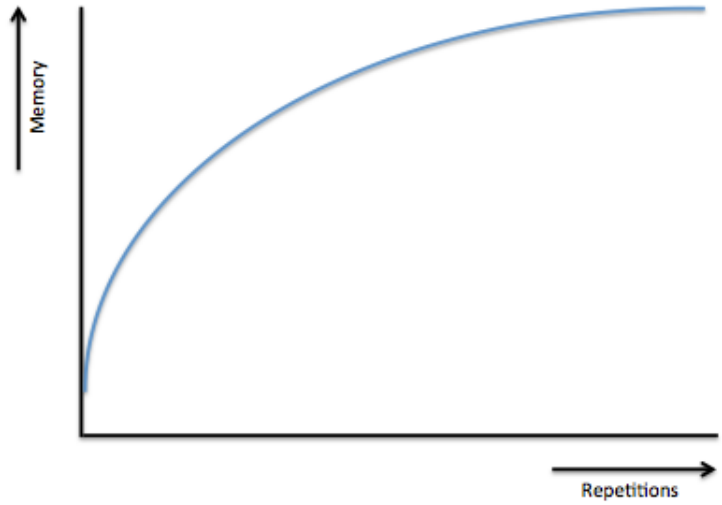



**Figure 1 - Ebbinghaus' (1885) exponential learning curve.**

Taking into account these findings from Ebbinghaus (1885), how can the process of learning be improved? One way is to add extra rehearsals in a spaced manner, another possibility is by adding testing; both options will be discussed further on in the theoretical background.

Once people have learned new information, at a set point in time they start forgetting again; knowledge shows some form of decay over time. It has been demonstrated that there is an important influence of time on forgetting in the working memory; this influence is a time-based decay mechanism (Portrat, Barrouillet, & Camos, 2008). The decay theory states that over time memory traces erode and therefore information cannot be properly retrieved anymore, i.e. over time people forget things (Berman, Jonides, & Lewis, 2009). Additionally distractor tasks have a high influence on recall and recall accuracy; i.e. attention is required for accurate recall (Oberauer & Lewandowsky, 2008). Again Ebbinghaus (1885) already demonstrated this principle; one of the findings he documented is what is now often referred to as the Ebbinghaus forgetting curve, see Figure 2:

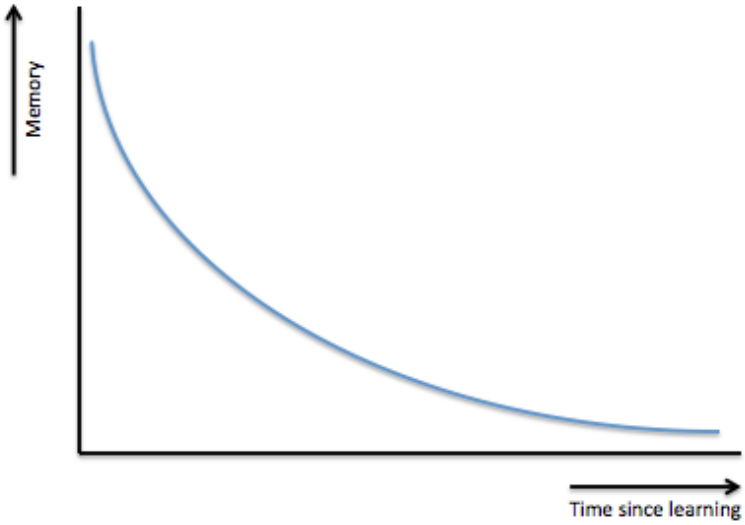



**Figure 2 - Ebbinghaus' (1885) exponential forgetting curve.**

Ebbinghaus' (1885) experiments showed that already after 20 minutes a significant amount (40%) of the acquired knowledge is forgotten. After six days people merely retain roughly 30% of the acquired knowledge. This exponential curve reaches a negative plateau after a while. The forgetting curve as described above is

based on a test in which meaningless syllables were used as stimuli; it does not take into account the quality and relevance of the knowledge. More recent studies provide evidence for the exponential nature of this curve. Averell & Heathcote (2011) showed an exponential function fitted empirical data best measuring cued recall and stem completion.

When further looking at the example of the forgetting curve, the serial position of an item has influence on recall (Ebbinghaus, 1885). When asked to recall a list of items in any order (free recall), the serial position curve is characterized as follows: The last words of a list are mostly retrieved (recency effect) since they are still in short-term memory. From all previous items, the first items from a list are typically more often retrieved than the items halfway a list (primacy effect) (Murdock Jr, 1962); this is mainly due to the fact that items at the beginning of a list are normally rehearsed more often and might be already transferred to long-term memory, see Figure 3. To overcome any effect of serial position, the declarative test questions have been randomized per test in this study.

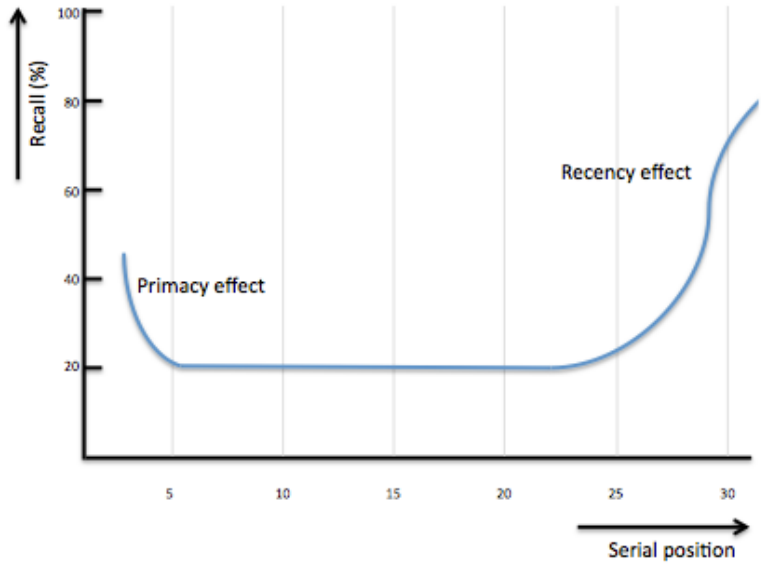



**Figure 3 - When subjects are asked to recall learned items, the last items are recalled best (recency effect), followed by the first items on the list (primacy effect).**

Abovementioned provides a basic understanding of learning related processes. The theory explains that final memory performance can be improved in an exponential way by adding rehearsals, normally in the form of studying. What also has been described is that over time humans forget their knowledge, showing the opposite exponential trend as for the learning curve. What the theory above does not describe is if other forms of rehearsing, rather than restudying the same stimuli, have a similar effect on learning and retention. Another form of rehearsing which might be more beneficial will be discussed at a later stage. Both the learning curve and forgetting curve explain the process regarding declarative knowledge; how these curves apply for procedural skills is not explained. The hypothesis is that similar graphs apply for skills, but with the longer-stretched acquisition and retention periods.

## *The transition to procedural skills*

Knowledge can be divided into declarative knowledge and procedural knowledge (Squire, 2004). Declarative knowledge is explicit, can be verbalized and is related to (novel) events and facts. Declarative knowledge is obtained in a conscious state. Declarative knowledge can be acquired suddenly, by for instance learning word lists for a school test, by a teacher telling what the capital of the USA is, actively memorising a telephone number, or when someone tells you a fact (Anderson, 1976). Declarative knowledge can be formed and stored already after just one encounter but may degrade quickly afterwards (Ferman, Olshtain, Schechtman, & Kami, 2009). Procedural knowledge is about knowing how to do something, e.g. riding a bicycle. Procedural knowledge may not be verbalized, however it can be applied in an unconscious manner. It often takes more practice and time before acquiring a procedure. In essence the distinction between declarative knowledge and procedural skills can be described as knowing what versus knowing how respectively.

The process of procedural learning is somewhat different compared to declarative learning where a single encounter may be sufficient to have learning. Dreyfus and Dreyfus (1980) state that once students become more skilled, they depend less on abstract principles and more on concrete experience. In their work they concluded that any skill training procedure must be based on some model of skill acquisition, so that at each stage the appropriate issues can be addressed and the appropriate methods can be chosen to improve learning. Normally students go through five different developmental stages: Novice, Advanced beginner, Competent, Proficient and Expert. The scheme below from Kirkpatrick and MacKinnon (2012) clearly describes the different stages (see Figure 4):

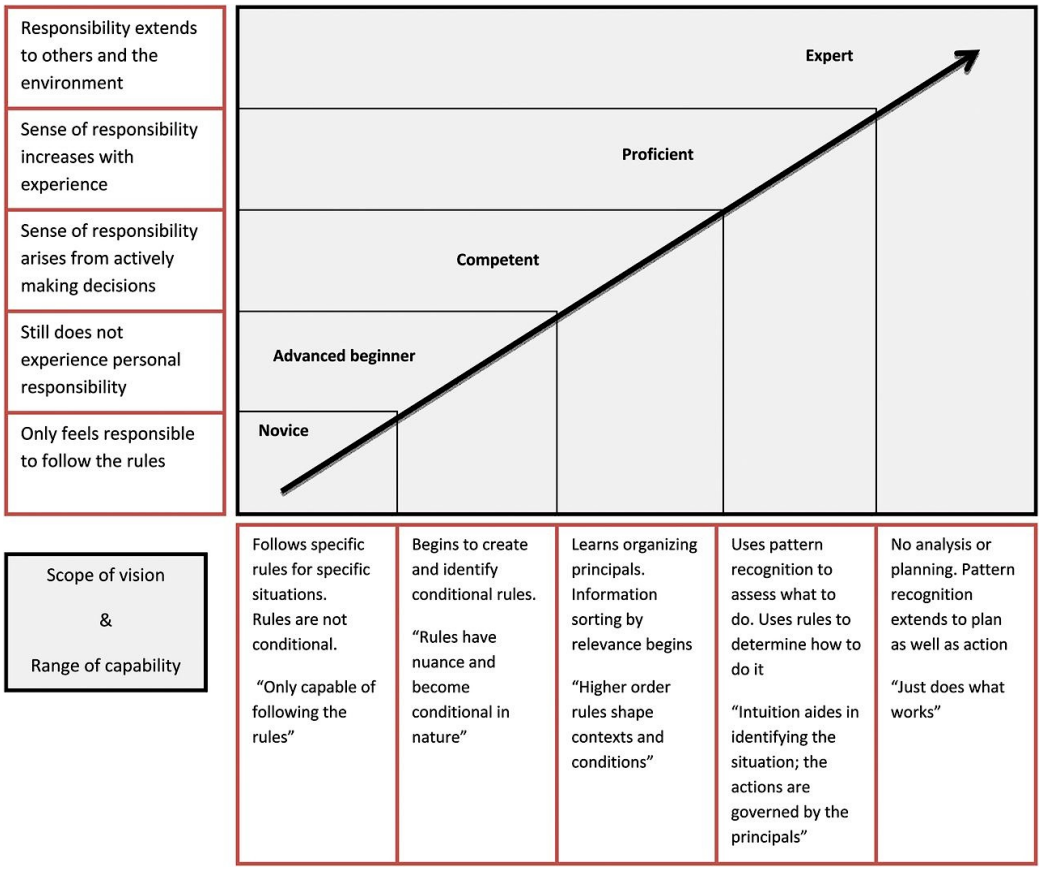



**Figure 4 – Overview of the five stages of skill acquisition. Reprinted from 'Technology-enhanced learning in anaesthesia and educational theory' by K. Kirkpatrick & R.J. MacKinnon, 2012, *Continuing Education in Anaesthesia, Critical Care & Pain*, 12 (5), 263-267.**

The typical process of skill acquisition stretches over a long period of time. Where subjects in this study had to develop their skills within a three-hour training course with roughly an hour of practice, they could not go through all described skill acquisition stages (see Figure 4). Most probably subjects are probably able to develop a set of basic rules. By practicing and using the trainer's feedback, some subjects may have started to develop their set of rules and making this information their own. The process of mastering skills typically takes more time than the planned time during the training sessions, especially when as complex as making echoes.

In the way declarative learning and procedural learning are described, it might seem these systems are isolated, distinct systems. From a historic perspective there are multiple isolated theories about learning, memory and the underlying mechanisms, examples are Pavlov's classical conditioning, or Hull's reinforced stimulus-response learning. All proposed systems are distinct and explain learning and memory from a single perspective (McDonald, Devan, & Hong, 2004). However there is not just one system responsible for learning and memory. So even though at first glance the memory systems seem autonomous and independent, these systems interact in numerous ways (Squire & Zola, 1996). An example showing the interaction is 'proceduralization' of declarative knowledge, where declarative knowledge transforms into procedural knowledge; this process will be described by the example of learning a (second) language. Anderson (1983) describes three different phases in the transition process from declarative to procedural: In the first phase (*cognitive*) information is being stored as distinct rules: 'walked' = 'walk' + '–ed'. This knowledge cannot be used yet in a sentence and other words may not be formed yet, rules are very explicit. In the second phase (*associative*) the isolated facts learned in phase 1 are moulded into more efficient production rules: 'walked' and 'showed' show a similarity which can be covered in a rule: Generating the past tense is done by adding '-ed' to the verb. However there is a risk in phase 2, since the newly generated rule does not take into account for instance irregular verbs. Especially during this phase many errors may occur. Rules are applied more often in this phase as part of the process of gaining experience. In the final phase (*autonomous*) rules become implicit procedures and the learner may add nuances, such as it only applies to a subset of verbs. This process can also work the other way, when parts of procedural knowledge can be recollected as declarative knowledge due to experience (Ferman, Olshtain, Schechtman, & Kami, 2009).

## Skill acquisition and ACT-R

Skill acquisition is a topic often described in relation with ACT-R (e.g. Taatgen & Anderson, 2002; Taatgen, Huss, Dickison & Anderson, 2008). ACT-R is a cognitive architecture developed mainly by John Robert Anderson. ACT-R can be used for simulating and better understanding human cognition (Anderson, 1983). By creating cognitive models, a better understanding of skill acquisition and the transfer of skills can be created. ACT-R uses two different long-term memories: the declarative memory stores facts and experiences and is basically passive. The procedural memory contains condition–action patterns and productions, which goals, results of memory retrieval, and perceptual input onto actions (Taatgen, Huss, Dickison, & Anderson, 2008). In ACT-R skills are represented by productions. Learning in ACT-R goes via instructions in declarative memory, interpreted by productions, carried out by other productions (Taatgen, Huss, Dickison, & Anderson, 2008). In the ACT-R architecture knowledge starts as declarative information and procedural knowledge can be learned when making inferences from factual knowledge that already exists. The advantage of productions is that operators do not have to be retrieved and tested from declarative memory.

Productions have a utility score that reflects how useful they are in a specific situation. If multiple productions meet the preconditions, the one with the highest utility value is selected. Learning within the ACT-R architecture is mostly achieved by a mechanism called production compilation (Taatgen & Anderson, 2002). If two productions fire sequentially, a new production is formed in the procedural memory. Over time compilations produce a considerable speedup and a reduction in errors. Compiled productions will not be selected directly after creation yet. Productions are selected on their utility value and new productions start

with a zero utility value. Hence compiled productions cannot compete with other, more often used, productions having higher utility values. Each time a certain compilation of productions is recreated, the utility value will be increased (Taatgen & Anderson, 2002). Finally, the compiled production will be chosen over the single ones. New productions will be introduced slowly due to this mechanism n, consistent with the idea that skill acquisition is slow. The learning speed is controlled by a learning parameter that determines how fast the utility of the new production converges with the utility of the old production.

An important mechanism that enables learning is the transfer of knowledge (e.g. Perkins & Salomon, 1992; Taatgen, 2013). Humans have the ability to reuse acquired knowledge in different, but similar situations; i.e. humans can transfer learning. Transfer of knowledge occurs when acquired knowledge in one context has an influence on learning in another context or with different, but related, materials (Perkins & Salomon, 1992). There are two scenarios that describe the transfer scenarios. The first scenario explains it requires less effort to learn new, similar systems once someone has already acquired a set way of working. An illustration of this would be learning how to use a text-editor. Once students have learned how to use a certain type of text-editor, learning how to use subsequent text-editors is easier. The more shared elements, the easier, since more knowledge can be transferred (Singley & Anderson, 1985). A second example involves learning how to drive a bus. One can imagine that learning how to drive a bus is much easier once one already knows how to drive a truck. The knowledge of how to drive a truck can be transferred to learning how to drive a bus, making it easier to acquire that skill. The transfer of learning is one of the bases for education. What is usually learned in classrooms and from books differs from real-life situations. However what is learned in classrooms does help, because once this knowledge is acquired, it is easier to deal with real-life situations. To finish the learning process from an educational perspective, the transfer of learning has to take place (Perkins & Salomon, 1992).

Productions have a fairly high complexity, making it challenging to explain how skills are learned and represented in the brain. Additionally productions are usually highly task-specific, making it hard to characterize how skills are interrelated. While already explaining skill acquisition and transfer in detail, the ACT-R architecture was not able to fully describe the acquisition and transfer of skills yet. Taatgen (2013) argues that rather than on a semantic level, the transfer of knowledge mostly occurs on a syntactic level. Singley and Anderson's (1985) experiment of learning text editors again illustrates this; learning a new text editor is easier if someone have already mastered a different editor. Taatgen (2013) proposes the primitive elements theory, a theory in which production rules are broken down into their smallest possible elements called primitive information processing elements (PRIMs). The primitive elements theory proposed by Taatgen (2013) splits the basic information processing units into both task-specific information and task-general information processing patterns. Most elements are task-general and control the flow of information on a syntactic level in the mind. The separation between the task-specific and task-general elements enables reusing general components of a skill for multiple tasks.

Multiple PRIMs are required to complete a task; developing cognitive skills means being able to combine these primitive elements into larger units of which some are still independent of context. Often-recurring combinations are built, combinations that can be used for many different tasks. These combinations form the basis for transfer. Training for tasks means increasing the amount of available sets of operators. Finally productions are still built that are task-specific with the benefit that the components of those productions are based on more generic smaller elements. Taatgen (2013) has shown the primitive elements model provides a better data fit compared to older models. An example of an older model is the identical productions model proposed by Singley and Anderson (1985). In this model they propose production rules as the elements of transfer. The measure of potential transfer is the identical productions between two. The measure for transfer in the model proposed by Taatgen (2013) is the amount of smaller units, a combination of primitive elements that can be used for other tasks. So rather than having entire production rules that could be transferred, smaller

units can be transferred.

The theory described so far provides a broad view on learning of both declarative knowledge and provides a view on how productions are created that form the basis of procedural skills. There is a difference in learning declarative knowledge and learning procedural skills. Declarative learning is more straightforward, and declarative facts can be stored in memory directly; when acquiring procedural skills there is a transition phase in which declarative facts become procedural and do not have to be recollected consciously anymore. The required time for acquiring skills typically is longer as well, since one encounter typically is insufficient to master a skill. Consequently a stronger declarative effect in comparison with the procedural effect might be found.

## *Knowledge and Memory*

There are multiple memory systems; the most important split is between declarative memory (for facts and events) and nondeclarative memory (e.g. skills or habits) (Squire & Zola, 1996). When looking at declarative knowledge, we also refer to declarative memory; consequently for procedural knowledge, we refer to procedural memory. Declarative memory is representational; it provides a way to model the external world, and as a model of the world that is either true or false. Declarative memory is primarily located in brain systems involving the hippocampus (National Research Council, 2000). Linked to procedural knowledge, there is a form of nondeclarative memory. Nondeclarative is neither true nor false. Nondeclarative memory is typically expressed through performance rather than through conscious recollection. Nondeclarative (or procedural) memory is mainly located in brain systems involving the striatum (National Research Council, 2000). Looking at the overview below, both declarative and non-declarative memory can be further split (Squire, 2004), see Figure 5:

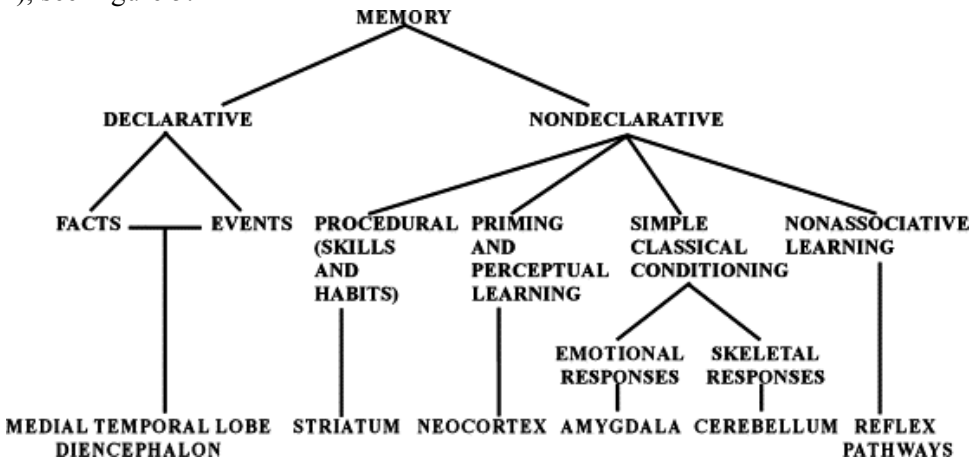



**Figure 5 - Taxonomy of the different memory systems (long-term). Adapted from 'Memory systems of the brain: A brief history and current perspective' by L.R. Squire, 2004, *Neurobiology of Learning and Memory*, 82, p. 173.**

In this study we focus on declarative memory with declarative facts and on a part of nondeclarative memory, namely procedural skills. As discussed learning is not a combination of isolated facts, neither is memory. Memory cannot be seen as a single entity or phenomenon that simply occurs somewhere in the brain in a fixed position; memory rather is a complex something (Squire, 1992). The idea of multiple learning and memory systems arose in 1957, after doing experiments with patients after unilateral removals for temporal lobe epilepsy. While some forms of memory were negatively affected after surgery, early memories and technical skills were still intact. Neither did the surgery affect patients' personality of general intelligence (Scoville & Milner, 1957). Based on these findings a strong idea arose that not all types of knowledge simply exist in one place in the brain. Multiple brain areas seemed to be responsible for different types of knowledge. Furthermore these memory systems operate in parallel (Squire, 2004).

## *Medical Education*

The context of this study is education. In education there has been a shift from knowledge acquisition (both declarative and procedural), towards learning in a way that understanding is the main goal. Students still have to learn a collection of facts from both textbooks and lectures and are often tested on those. While learning these disconnected facts is still important, when looking at experts' knowledge these facts are organized and connected and support understanding. Also this form of contextual, connected knowledge allows to be transferred to other contexts (National Research Council, 2000).

Medical education has undergone changes as well. In healthcare services, there is a continuous pursuit to improve the overall quality. In order to do so the quality of practitioners should be on a level as high as possible. This puts a stress on medical education, since the urge to deliver high quality students gets even more attention than before (Ziv, Small, & Wolpe, 2000). Therefore medical schools are undergoing a shift in the way students are being taught. As patients are still less open to students 'practicing' on them, patient safety and quality are gaining a higher priority than bedside training and education. As a consequence medical education has been altered in multiple ways, such as restructured curricula and an increased share of self-directed learning. These alternations however do not overcome the lack of connection between the classroom and the clinical environment (Okuda, et al., 2009). Simulation-based medical education is proposed to fulfil these needs. In order to help students develop good technical skills before practicing on real patients, simulator-based training is becoming widely used all over Europe (Maiss, Naegel, & Hochberger, 2011). Simulators offer multiple benefits. Since students have more time to practice their skills, less practice on real patients is required. As a consequence fewer errors are likely to occur, improving patient safety and wellbeing. Simulator interfaces are already on such a quality level, that they can provide learners with visual 'patient' reactions as a response to the learner's actions. (Kunkler, 2006). It has been demonstrated that medical students remember more of a skill after performing a procedure rather than merely reading about it (Croley & Rothenberg, 2007). Simulators enable repeated practice before actually practicing on a real patient. Simulator-based medical education has proven to be beneficial in comparison with traditional education, such as lectures, books, articles, and web-based resources (Shakil, Mahmood, & Matyal, 2012). Cantù and Penagini (2012) did a systematic review on the use of computer simulators in digestive endoscopy domain in which it shows that subjects learn on simulators with relatively short learning curves.

In the present study, students have been trained on making echoes using an echocardiogram simulator. Where in practice only limited time for quality training and practice is available, by using high quality simulators, the lack of time for training could be compensated for (Shakil, Mahmood, & Matyal, 2012). By using a simulator, students can familiarize themselves with the different views, create and understanding for the translation between three dimension images and two dimension images, manipulating the probe etc. Neelankavil et al. (2012) conducted a study on the effectiveness of simulator-based echocardiography training of noncardiologists in congenital heart diseases to determine the most effective training method, hypothesizing that simulation-based training would outperform other training methods. Neelankavil et al. (2012) compared a control group, which received a lecture-based video training versus the simulation group, which received training on a TTE simulator. The simulation group significantly outperformed the control group on all criteria after the first training session. Besides objective results showing the efficacy of TTE simulator training, additionally students highly appreciate practicing on a simulator (Platts, Humphries, Burstow, Anderson, Forshaw, & Scalia, 2012). Forty trainee sonographers were trained in a simulator workshop on making the apical two-chamber (AP2C) view and on imaging the superior vena cava (SVC) using the same CAE VIMEDIX$^{TM}$ ultrasound simulator as used in this study. After training participants assessed on its utility. Subjects were very positive about the usefulness of the simulator regarding identifying the SVC and obtaining the AP2C view.

## *Testing effect*

A growing number of studies have shown that testing itself can enhance learning and improve long-term retention beyond the effect of spending the same amount of time on studying (e.g. Agarwal, Karpicke, Kang, Roediger III, & McDermott, 2008; Zaromb & Roediger III, 2010). It has been repeatedly shown that declarative memory can be improved when subjects retrieve information rather than restudying the material an equal amount of time; i.e. retesting is more effective than restudying. This benefit of testing over studying is called the testing effect. The testing effect was first described in the early 1900s (e.g. Abbott, 1909; Gates, 1917).

Gates' study (1917) is a classic study on the testing effect. In this study, children from four grades (1st, 4th, 6th and 8th) had to study from the book Who's Who in America. More specifically the subjects had the study lists of nonsense syllables and brief biographies. The subjects were instructed to read the material. At some point subjects had to stop reading and to mentally retrieve whatever they could from their reading. Subjects received a written test immediately after learning and were tested again after 3-4 hours. If subjects could not recall the materials during the test, they could check the material again. Actually by giving subjects this opportunity, it creates a negative influence on controlling the experiment since the amount of time restudying varied per subject. The general conclusion from Gates' study for both types of material was that a combination of study and self-testing produces better memory than merely restudying the material. Furthermore this study suggests that before self-testing can facilitate learning a certain amount of studying is required.

Most recent studies on the testing effect have been conducted on verbal materials such as word lists (Kang, 2010; Roediger III & Karpicke, 2006). Roediger III & Karpicke (2006) illustrate an example of what modern studies are generally like. Their experiment consisted of two phases: the first phase consisted of four seven-minute periods. The difference between groups was that during these periods subjects were either asked to study a prose passage or to take a recall test. For the recall test subjects were instructed to write down as much of the material from a passage regardless of the order of wording on a test sheet with the title of the passage. The second phase of the experiment occurred after a five-minute, two-day, or one-week retention interval. In the second phase, subjects had to recall passages from the first phase. Results showed that while on the short-term interval (five-minute) restudying proved to be more beneficial than testing, for the longer intervals (two days and one week) retesting showed a better retention of their knowledge. To higher the plausibility of outcomes, multiple testing effect studies were performed in simulated classrooms rather than laboratory settings among which Roediger III and Karpicke's study (2006).

Being able to retrieve discrete facts (e.g. words) does not directly demonstrate a better understanding of the subject that can be credited to testing (Daniel & Poole, 2009). Zaromb and Roediger III (2010) proposed the question whether the testing effect applies to learning and retention of the conceptual organization of study materials. To answer their question, Zaromb and Roediger III (2009) conducted two experiments with categorized word lists. The first experiment showed the beneficial effect of repeated testing rather than repeated studying applied when free recalling word lists, i.e. the testing effect applied. The main purpose of Zaromb and Roediger's second experiment was conducted to further examine the effects of testing on learning and retention of wordlists representing different taxonomic categories. In order to do so they compared delayed recall performance, measured by:

- Total word recall;
- Category recall (Rc)
- Words per category recall (Rw/c);
- Organization, measured by clustering (ARC)

Subjects in all four conditions took final tests (both free and category cued recall) one day after initial training. Results show that testing can improve organization of recall—or category clustering—in delayed free recall. By applying the testing effect, subjects not merely remember items better; the organisation of items improves as well.

Since most studies on the testing effect were conducted involving verbal learning, from a theoretical standpoint Kang (2010) hypothesized whether the type of stimuli is responsible for the beneficial effects of testing. Only a few examples of studies using other stimuli are available; two studies were performed using abstract visuospatial information (Carpenter & Pashler, 2007; Kang, 2010). As could be expected, both studies showed that having a test after initial studying enhances retention of declarative knowledge. More important, these studies showed that the testing effect could be generalized to visuospatial information as well. The way subjects are tested has an influence on the magnitude of the testing effect. Where restudying or taking a multiple choice test both enhanced final recall in comparison with no activity, the testing effect can be amplified by using an initial short answer test rather than a multiple-choice test (Butler & Roediger III, 2007). When using a multiple-choice test, giving feedback either directly or delayed, enhances the amount of correct answers and reduces the amount of intrusions (Butler & Roediger III, 2008).

Concluding the testing effect has shown to be a robust effect on different forms of declarative knowledge. Roediger and Karpicke (2006) state in their general review of the testing effect that the testing effect can be generalized from psychology laboratories to classroom settings (with educationally relevant materials).

## The testing effect and skills

So far the testing effect has been discussed in relation to declarative knowledge. The present study is aimed at finding if the testing effect can be applied on acquiring skills. In order to obtain and maintain skills one needs to practice. Berden et al. (1993) showed that three or at least six monthly refresher courses are required to maintain resuscitation (reviving from unconsciousness) skills. Could this effect be achieved without reinstructing subjects? Kromann, Bohnstedt, Jensen, and Ringsted (2010) showed that testing as a final activity in a skills course increases the learning outcome compared to spending similar time of practicing. In order to test this two groups were trained differently: a group with 4 hours of practice versus a group with 3,5 hours of practice combined with a 0,5 hour skills test. Subjects were asked to participate in a retention assessment half a year later. The results were not significant (P=0.06), however the results suggest the testing effect might apply on skills.

An earlier and similar study conducted by the same group (Kromann, Jensen, & Ringsted, 2009) does show a significant difference between groups that merely practiced and groups that took a test as well. The objective of this study was to determine if learning outcome could be increased by adding testing as the final activity in a skills course rather than spending an equal amount of time on practising the skill. Results show that learning outcomes were significantly higher in the intervention group in comparison with the control group. It is important to remark that this study showed effects on retention intervals of two weeks. The present study is aiming at longer retention intervals. Additionally the training and tests were all planned on the same day, without applying any form of a distributed learning model. By spacing the tests over time, stronger effects might occur.

## *Spacing effect*

The way people learn differs per person; where some like to study everything at once, others prefer a more distributed manner. When learning is done in a distributed manner, memory is enhanced beyond learning in a massed manner (Vlach, Sandhofer, & Kornell, 2008). This robust phenomenon is called the spacing effect.

In Jost's Law, a law about forgetting, the spacing effect is described as: "If two associations are of equal strength but of different age, a new repetition has a greater value for the older one" (McGeoch, 1943). When multiple facts have been studied and are not forgotten yet, it is most beneficial to repeat the one with the oldest previous encounter. In more recent literature similar descriptions are given; Banaji and Crowder (1989) explain the spacing effect in a way that a repetition will be most beneficial 'if the material had been in storage long enough as to be just on the verge of being forgotten'. The spacing effect explains that humans learn or remember something more easily when studying in a spaced manner (Cull, 2000), i.e. distributed learning is more effective than so-called 'massed learning'. By adding extra rehearsals, the final memory performance can be improved (see Figure 6):

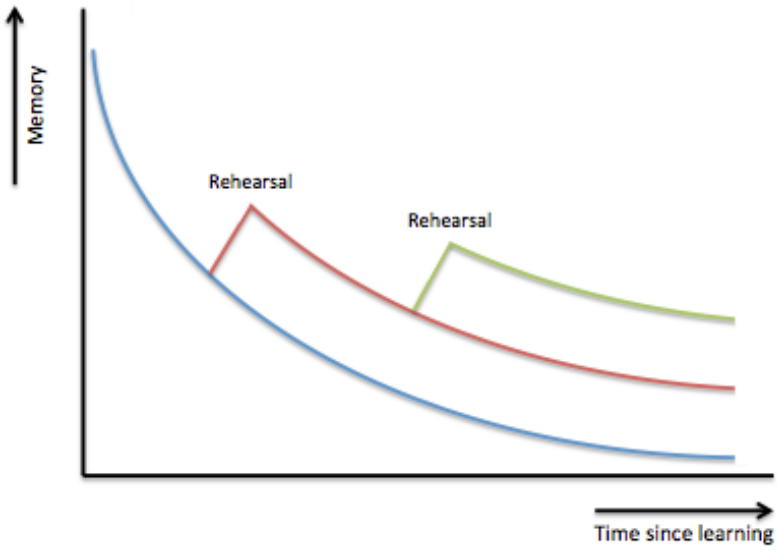



**Figure 6 - Adding rehearsals improves the retention.**

The spacing in itself is not arbitrary. In a learning situation with restudy moment, there are two intervals: the interval between the first time of studying and the restudy moment, the so-called interstudy interval (II) and the interval between the restudy moment and the (final) test, the so-called retention interval (RI). With an increasing retention interval, the interstudy interval should increase as well. Where in the past optimal intervals with a ratio of around 1:1 (interstudy interval:retention interval) were assumed (Crowder, 1976), Pashler, Rohrer, Cepeda, and Carpenter (2006) suggest optimal spacing between the first time study and the restudy moment should be a fraction of the final retention interval, somewhere between 10%-20%. In a short experiment the optimal interstudy interval for a 1-week retention interval was 1 day and for a 50-week retention interval, the optimal interstudy interval showed to be 3 weeks. Following those findings, Pashler et al. (2006) conclude that by spacing the restudy moment between 10%-20% of the final retention interval, an optimal interstudy – and retention interval is obtained (see Figure 7). They add that when using intervals, longer than optimal spacing is less harmful to final retention than spacing shorter than optimal.

The magnitude of the retention interval should be altered based on the type of test. By controlling the retention interval based on the type of test, the magnitude of the spacing effect can be influenced on positively. Where for instance in free recall tests increasing the retention interval will tend to reduce the

advantage of spacing items, in cued recall tests the spacing effect should get stronger and stronger over time (Delaney, Verkoeijen, & Spirgel, 2010).


**Figure 7 - Spacing effect and intervals**

In order to apply the testing effect in such a way that potential effects can be attributed to the testing effect, it is important to be aware of the consistency of the experimental conditions. In a general review of the spacing effect, Delaney, Verkoeijen, and Spirgel (2010) state that research often fails to control encoding strategy in spacing experiments. This could lead to different magnitudes of the spacing effect, since participants might adopt increasingly better study strategies across lists. Averaging across multiple lists, even when the order is counterbalanced, can produce misleading estimates of the true effect size. The effect will be there, but by controlling the conditions, the magnitude is more consistent across subjects (Delaney, Verkoeijen, & Spirgel, 2010).

## *Interaction between both effects*

Where both the testing effect and spacing effect are beneficial for learning and retention, a combination enhances learning even more beyond the individual effects (Izawa, 1992). Most studies on a combination of the testing effect and spacing effect were conducted with verbal materials (e.g. Izawa, 1992; Cull, 2000). Similar patterns as with verbal experiments can be found with visual stimuli (Carpenter & DeLosh, 2005). Carpenter and DeLosh (2005) determined whether the testing and spacing effects could be generalized to name-learning situations. In order to do so, they conducted three experiments. One experiment sequentially presented paired face-name items for 6 seconds each. Paced by subjects, both test and study items were repeated three additional times. The results showed that where both the testing effect and spacing effect are beneficial individually, a combination resulted in the best memory performance. This finding is in line with earlier research (e.g. Cull, 2000).

An important implication from the study of Carpenter and DeLosh (2005) is that unless tests are spaced at non-immediate intervals, memory may not even benefit beyond that of additional study. In practice this means that when taking interim tests, these tests should not be taken directly after learning study materials. By choosing a non-immediate interval, better long-term results can be achieved.

The theory described so far provides an introduction to learning both in the declarative and procedural way. It explains how we acquire knowledge, how we retain knowledge and what could be done to improve the process of learning and retention. We have seen that applying the testing effect and the spacing effect can enhance learning performance, however it is unsure whether these effects can be generalized to procedural skills. The hypothesis is that it is possible, for the main reason that retesting requires subjects to actively recollect their acquired skills and by giving feedback on their performance an additional learning repetition is added. The present study aims at improving learning performance and retention by applying the testing effect and the spacing effect to learn if these effects can be generalized to acquiring and retaining procedural skills.

# Practical Background

## Echocardiography

In order to diagnose cardiovascular problems, guide treatment decisions, monitor changes and determine the need for additional tests, a doctor can decide to use echocardiography. To create moving images of the heart, echocardiography uses high-frequency sound waves. The images obtained by echocardiography can be used for the diagnosis and management of cardiovascular disease. Echocardiography provides helpful basic information such as the size and shape of the heart and information about pumping capacity, and the location and extent of any tissue damage. Additionally it can record blood flow as well using Doppler ultrasound techniques.

There are two types of echocardiography: transthoracic and transesophageal. Transthoracic echocardiography (TTE) is a non-invasive method in which a probe or transducer is paced on the chest. TTE enables an accurate and quick assessment of the heart. The probe can be manipulated in different ways: Positioning, tilt on short axis, tilt on long axis, rotation and a combination of prior. In case TTE is insufficient to get a clear and precise image of the heart, a doctor could decide to do TEE (transesophageal echocardiogram). A specialized probe is placed into the patient's oesophagus. This type of echocardiography is invasive and sustains higher chance for complications. In case of emergency, performing a TEE might not always be possible. In this study subjects are introduced to TTE, the different windows and views subjects had to remember and use will be further discussed.

## Windows and views

There are multiple imaging windows; in total there are four main imaging windows:

1. Parasternal
2. Apical
3. Subcostal
4. Suprasternal.

All echoes included in the test were located in the parasternal window and apical window; see Figure 8, numbers 1 and 2 respectively. In the parasternal window, subjects were required to make the parasternal short axis (PSAX) view and the parasternal long axis (PLAX) view. The long axis simply had to be made in one way; while the short axis can be made on different levels, for the procedural test in this study, subjects were required to make the PSAX on the aortic level. From the apical window, all chamber views (two/three/four/five) can be obtained. The following overview (see Figure 9 A-F) shows examples of all six echocardiograms. To find additional information on the views and how to obtain them can be found in Appendix 1 - Views and instructions on how to obtain them.

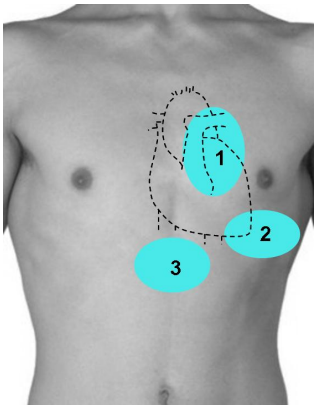
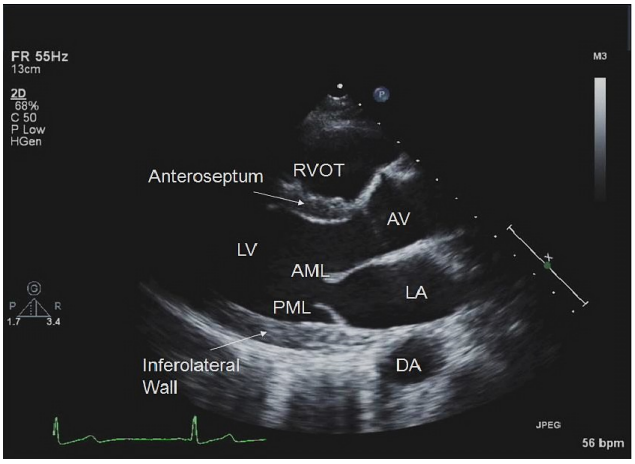
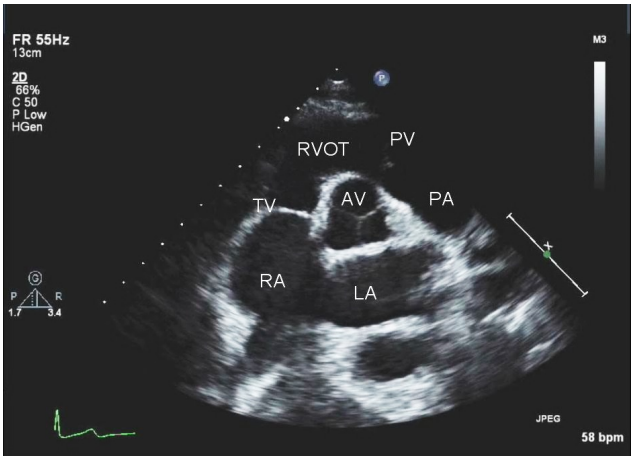



**Figure 8 – Echocardiographical Windows**

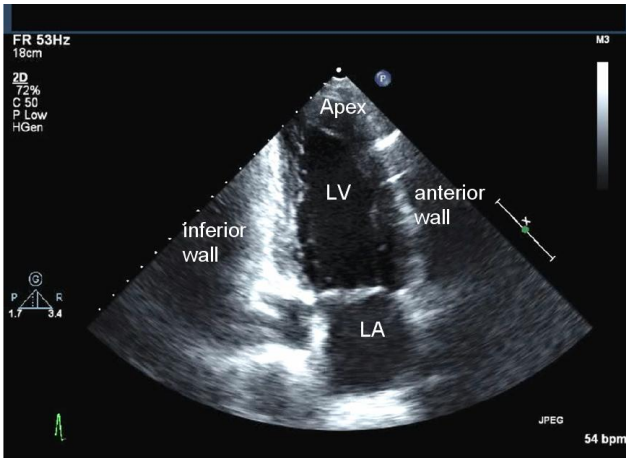
**A - Long axis**



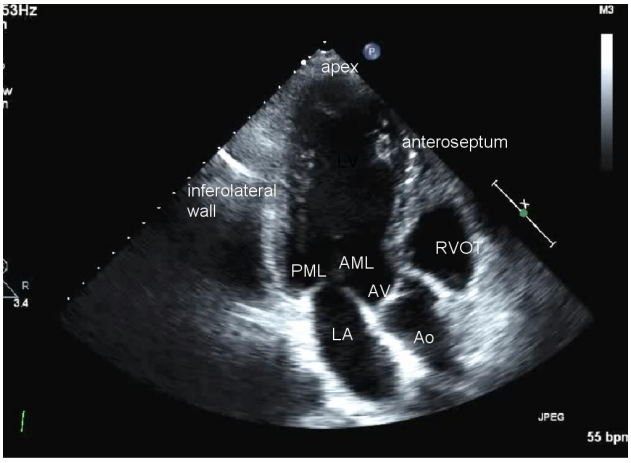

**B - Short axis**



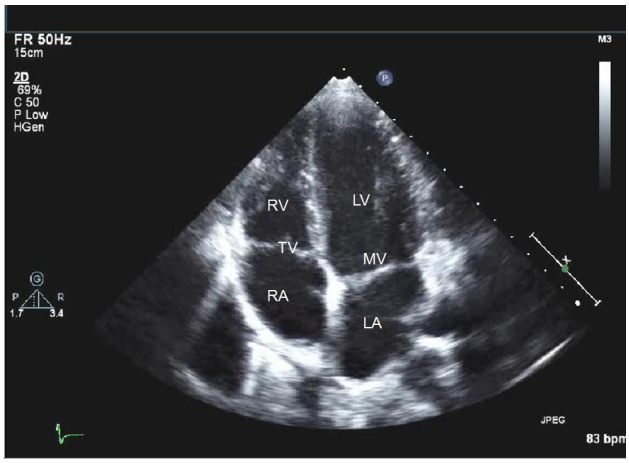

**C - 2 chamber**




**D - 3 chamber**




**E - 4 chamber**




**F - 5 chamber**

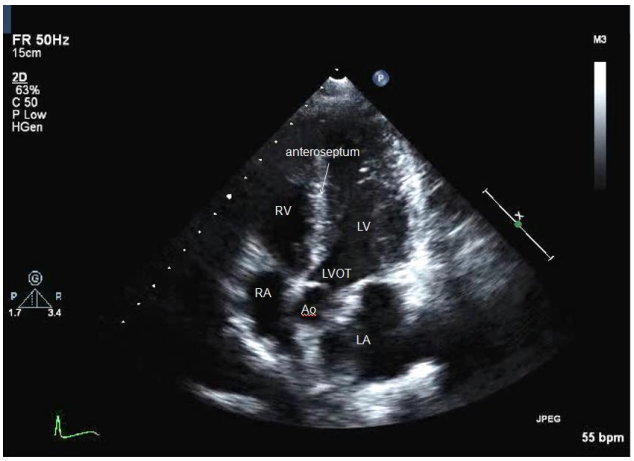



**Figure 9 (A-F) - Examples of the six echoes used in this study.**

# Methods

The present study was conducted in the University Medical Center Groningen's (UMCG) Skills Center. In order to answer the research questions, four groups of medical students have been trained and tested on both their declarative knowledge and procedural skills. The full experimental design will be discussed in this section.

## *Experimental Design*

### Subjects

Medical students (2nd and 3rd year) from the University of Groningen have been recruited for this study. Subjects have been recruited via an e-mail with an attached information letter about the planning, the timing and the (dis)advantages of voluntarily participating in this experiment (see Appendix 2 – Information letter students). Initially 43 students subscribed for the study; for the final test 35 subjects came back. Subjects were assigned to groups based on their enrol date. This was done to rule out the effect of the training becoming more efficient through time.

### Groups

All subjects have attended an initial training in the form of video instructions. Using video rather than live instructions has no influence on the testing effect (Butler and Roediger, 2007). Prior to the training, a declarative pre-test was given to test the subjects' initial declarative knowledge. After the video instructions, subjects practiced on the echocardiogram simulator. After the training and practicing, subjects got both a declarative test and procedural test. For an overview of the training, see Figure 10:


**Figure 10 - Set up training day.**

The first group merely attended the training (including all tests) and did the final tests (both declarative and procedural) after approximately 3 months. The second group got an additional declarative test. The third group got an extra procedural test to find if the testing effect can be directly applied on procedural skills. The final group was retested on both their declarative knowledge and procedural skills. Finally, all groups did both a declarative and procedural test at the end of the testing period. For a complete overview, see Figure 11. The rationale behind these groups was as follows:

**Group 1:**
Control group to measure performance without manipulations.

**Group 2:**
Does the testing apply on this type of declarative knowledge?
*Hypothesis: Outperforms groups 1 & 3 on declarative test.*

**Group 3:**
Does this group benefit from an interim test on procedural skills?
*Hypothesis: Outperforms groups 1 & 2 on procedural test.*

**Group 4:**
Does this group benefit from better declarative knowledge on the performance on the procedural test?
*Hypothesis: Outperforms group 1 & 3 on the declarative test. Outperforms group 1 & 2 on the procedural test. Might outperform group 3 on the procedural test due to benefits from enhanced declarative knowledge.*

# *Training*

## Theoretical session

All training was completed in the Skills Center at the University Medical Center in Groningen. To assure consistency in the training, subjects have been instructed via a video (length 00:21:45). This video consisted of:

- A brief explanation of the basic anatomy and function of the heart;
- An introduction to echocardiography;
- Explanation about the probe/transducer and how to use it;
- Imaging windows (five), further focusing on the parasternal window and apical window;
- Positioning of the patient;
- Short discussion of the views relevant for this study;
- Overview of all views with labels in structures;

## Practical session

All subjects were given the same amount of time to practice the skill. Since groups sizes differed from two to four subjects; the time subjects were allowed to practice was cut off; i.e. groups half the size were allowed half the time to practice. During the training sessions the researcher supported subjects gain understanding on the relation between probe placement/manipulation and the acquired cut plane on the simulator. This was done by either demonstrating on the simulator, on a model of the human heart or through sheets with extra information about the different views and how the plane cuts trough the heart.

Subjects were instructed to start with the parasternal long axis (PLAX) and use that as a starting point to rotate the probe into the parasternal short axis (PSAX). Subsequently subjects were asked to continue with the apical four-chamber view (AP4C), since this view normally is considered as important and other apical views can be obtained from this view. From the AP4C instructions were given how to obtain the other apical chamber views. After introducing the views in this order for each subject, they were free to practice.

# *Declarative test and procedural tests*

Subjects have taken two tests: one to measure their declarative knowledge, one to measure their procedural skills. To test declarative knowledge an online multiple-choice test in Nestor (Blackboard) has been used; this declarative test was designed in such a way that minimal/no procedural skills are required. See Appendix 3 – Declarative test for an overview of the test. In order to isolate procedural skills, subjects were asked to perform tasks that required as little declarative knowledge as possible. For that reason subjects were merely asked to make different echoes without interpreting them. The echoes made by subjects are discussed in the practical background. A brief overview of both tests will be discussed.

## Declarative test

This test based on the current test of echocardiogram training, however the test was taken in a more structured way than was implemented in the intensivists training: In Nestor (Blackboard based) a multiple-choice test was taken without a time limit. Afterwards subjects received feedback on their test. The declarative test was taken during planned sessions. The test consisted of 14 videos of structures (moving picture) with an ‘X’ mark to indicate which structure had to be recognized. Further the test had three questions about manipulation of the

probe (to get from one view to another) and three about the functioning of the heart (see Appendix 3 – Declarative test).

## Procedural test

Subjects were asked to make six transthoracic echoes (parasternal long-axis, parasternal short-axis, apical two chamber, apical three chamber, apical four chamber, apical five chamber). There was no time limit to make an echo; subjects were instructed to signal once they were satisfied with the obtained image. Subjects had to cue the researcher to make a screenshot of the final results. In that case the researcher stopped the timer and made the screenshot. In some cases the screenshot did not fully reflect the final results, in those cases a remark was being denoted. All echoes have been assessed by two independent expert raters, of which one was involved in the study. The assessment has been done based on the following categories:

1. One or more compartments are not displayed/not fully displayed
2. All compartments are displayed, however one compartment incomplete
3. All compartments are displayed, however the angle/cross-cut is just off
4. All compartments are displayed with the right angle/cut

In case a remark was made about the screenshot not reflecting the final result during the test, the researcher could change scores.

## Test Scores

The maximum score for the declarative test was 100 (20 answers, 5 points each); for one particular question subjects received points for two of the answer options. As a measure to decrease the difficulty level of the procedural test, ribs/lungs/other tissue were disabled in the simulator. As a consequence subjects got more freedom in making the echoes. This meant that the required rotation to come from the AP2C to the AP4C is somewhere between 45 degrees and 90 degrees. This resulted in one ambiguously formulated question. In case of choosing either 45 degrees or 90 degrees, points were given. The measure of the procedural test was the quality of the submitted echoes, as scored by Dr. R.A. Tio and Dr. I.C.C. van der Horst.

## Feedback

Subjects received feedback on both tests. Directly after completing the declarative test, subjects received an overview of the provided answers. In case of a wrong answers, the correct answers were given without further explanation. During practicing making echoes subjects got real-time visual feedback from the system in a way that is described below in ‘Echocardiogram Simulator’. Subjects could use this direct visual feedback to adjust their echo. After the test, in which subjects made the echoes in random order, the researcher gave feedback on every echo on a few attributes: position, probe angle and rotation. Additionally instructions were given on how to easily obtain certain views. Feedback was given in the same manner to all subjects, but based on the test outcome.

## 2nd measurement

32 out of 42 subjects were invited to take an interim test. The spacing between the first measurement and the second measurement varied between 2 and 3 weeks. Ideally the spacing would have be exactly the same for each subject, however due to availability of both the Skills Center this could not be done. Subjects that had to do both test started with the procedural test. Subjects were not allowed to use the simulator (i.e. practice) prior to the test to ‘recalibrate’. Again feedback was given after finishing the test. The second test was the declarative test. The same test was given for the second measurement, however with a randomized order of question.

## Final measurement

For the final measurements all subjects that attended the initial training day, and if required the interim test(s), were invited again to take both tests. The procedure was similar to the previous measurements; first they had to make the procedural test, followed by the declarative test. In total 35 subjects returned for the final measurements:

| | |
|---|---|
| Group 1 | 7 subjects |
| Group 2 | 9 subjects |
| Group 3 | 9 subjects |
| Group 4 | 10 subjects |

Due to availability of the simulator (not available for several months for maintenance), it was not possible to plan extra timeslots to test the others as well. At the moment the simulator was available again, the retention interval was of such a magnitude, it was not useful anymore.

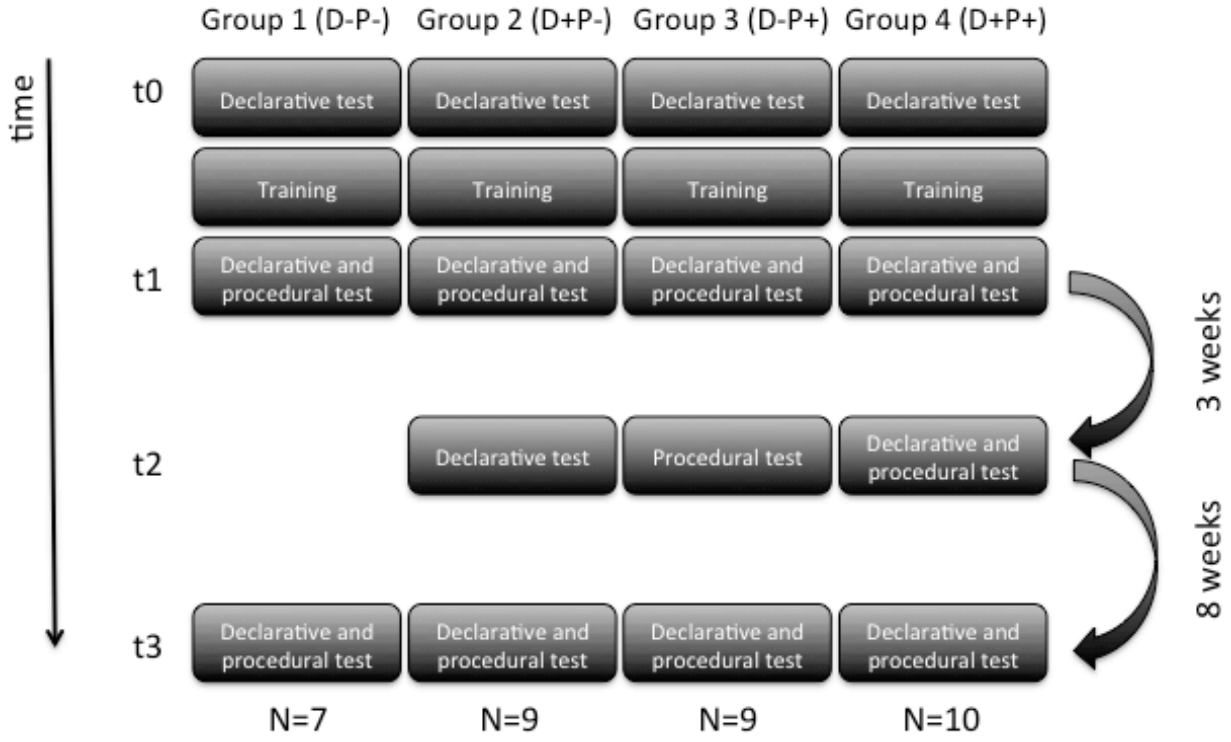


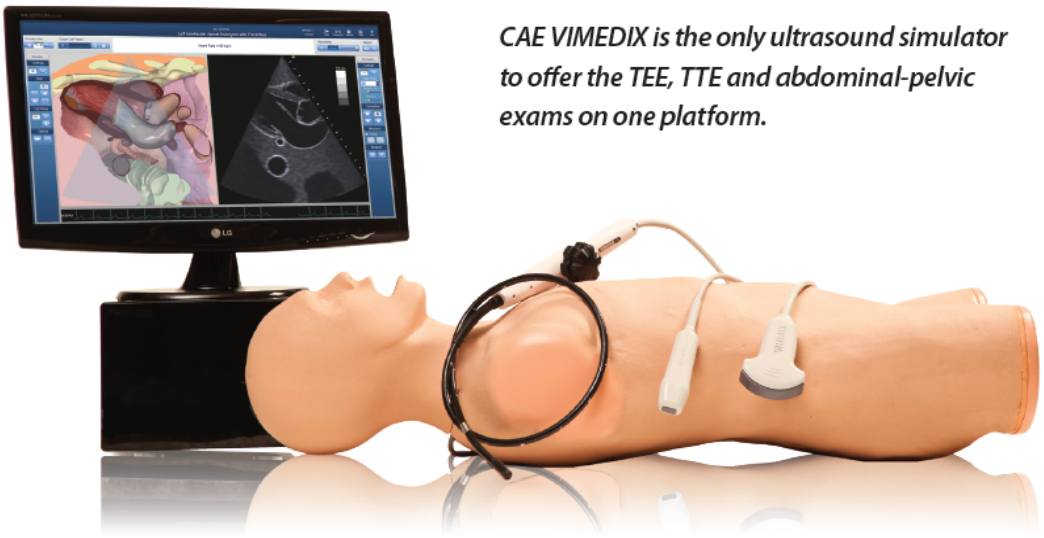

**Figure 11 - Overview experimental design.**

## *Echocardiogram Simulator*

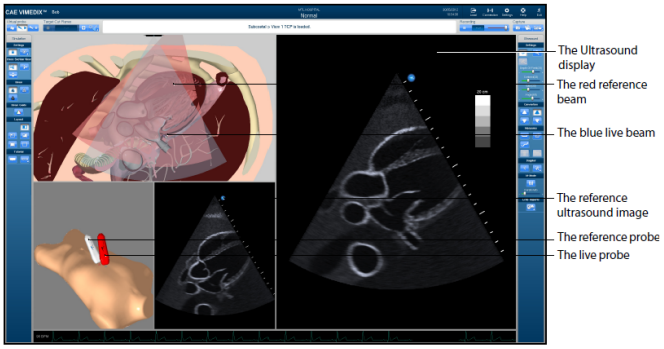
The simulator used in this study is the CAE VIMEDIX™ ultrasound simulator. The simulator consists a lifelike body, two transducers (TTE and TEE) and a computer with monitor. The system can provide users with multiple images simultaneously; a two-dimensional echo image is identical to a normal echo image. Additionally a three-dimension augmented reality model of the human body is given, providing input for better understanding the context. As said, the simulator has a very realistic human body with for instance ribs, providing the option the feel the ribs and search for a certain intercostal space. To decrease the difficulty level of the procedural test, ribs/lungs/other tissue were disabled in the simulator.

While there different types of transducers are available, namely a transthoracic, transesophageal and focused assessment with sonography for trauma (FAST) probes, for this study the transthoracic has been used. The section below shows a summary of the simulator used in this study (CAE Healtcare, 2013).

**CAE VIMEDIX Ultrasound Simulator**
Master Ultrasonography of the Thoracic, Abdominal and Pelvic Cavities



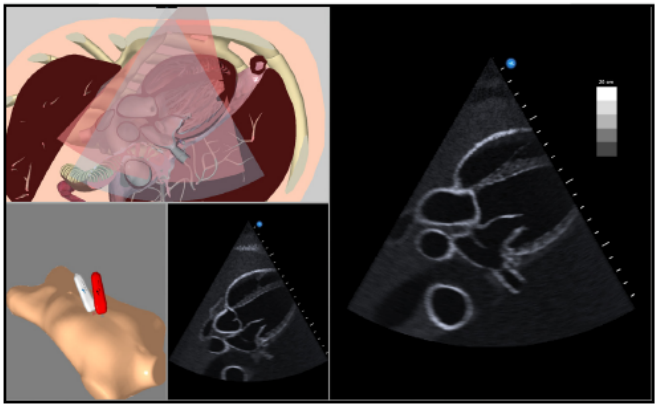

The split screen display was used during practicing. Subjects were shown a simulated live ultrasound image alongside an anatomical representation of the heart. Additionally an indicator of where the probe should be placed on the body was displayed. This view helped students create an understanding for how the system works and how manipulations of the probe results on the screen.

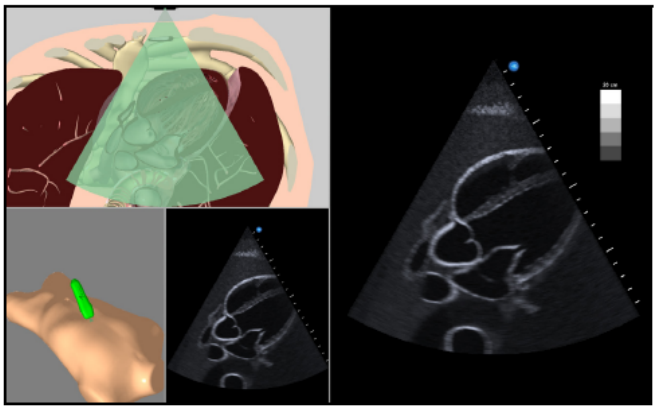



*Target Cut Plane Screen Features*

Live feedback was provided by the system: the cut-plane (upper left image) changed colour based on the placing; red for bad placing and green for correct placing. Additionally the intended placing of the probe got visualized as well (lower left).

In the example on the right an incorrect position is shown, illustrated by the red cut-plane.
A properly achieved target cut plane was shown in the following way (see example on the right): at the lower left, the virtual probe is green, indicating a proper starting position. Furthermore the target cut-plane (upper left) is green.





An Achieved Target Cut Plane

During the test all visual aids were disabled, making sure subjects have to rely on their skills and memory.

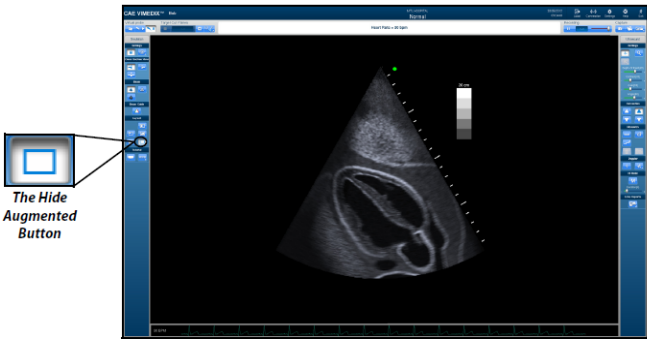



**Figure 12 (A-E) - Screenshots of the CAE VIMEDIX Ultrasound Simulator software. Reprinted from Product & Services: VIMEDIX, CAE Healthcare, Retrieved March 12, 2013, from https://caehealthcare.com/home/product_services/product_details/vimedix.**

# Results

Results for the declarative and procedural test will be discussed separately. In total four tests were taken for the declarative test and three tests were taken for the procedural test:

*Before training:* *Test 0 (Only the declarative test taken before training)*
*After training:* *Test 1*
*Interim test:* *Test 2*
*Final test:* *Test 3*

The groups have been labelled; the labels indicate whether or not groups did the interim test. The label 'D' stands for the declarative test, the label 'P' for the procedural test. The labels are followed by a minus or plus signs indicating if that group did not or did that specific interim test; see Figure 11 for the complete overview.

Before testing on significance the raw data will be inspected. With the current sample size individuals can have a large effect, by inspecting these data the appropriate way of testing on can be selected and it can be checked upfront if any conclusions drawn from the data are right. For testing on significance two separate models will be used. The declarative test will be analysed first by using a variance analysis on the test scores; the procedural test will be tested using a variance analysis on the delta scores.

## *Declarative test*

Subjects have been tested on their declarative knowledge before the training (T0); this was done to determine if there were any differences between subjects in declarative knowledge before participating in the experiment. For the main statistical analysis, the test scores after training (T1) and the final test scores (T3) have been compared. Only group 2 (D+P-) and group 4 (D+P+) had the interim declarative test (T2).

### Visual inspection of the declarative test scores

To understand how respondents scored and to see if there are any obvious differences between respondents, the scores for each respondent individually have been analysed first. The analysis shows no striking or concerning observations. Since the statistical analyses have been done on the group scores, the next step in visually inspecting the data is plotting the average scores per group to identify potential trends (see Figure 13). By doing so some resolution is lost, however it provides better input for understanding potential effects.

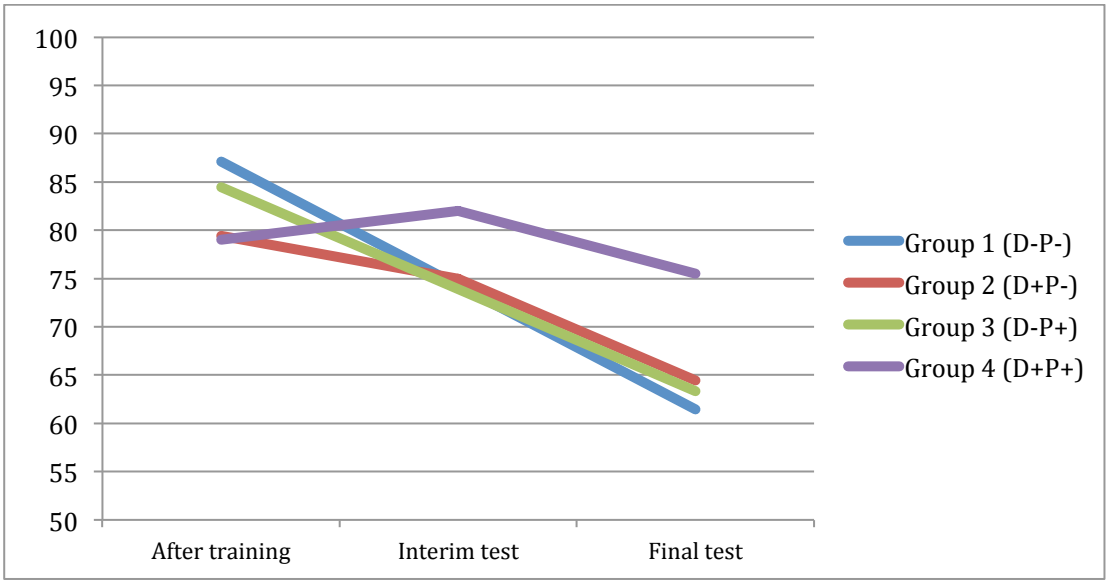
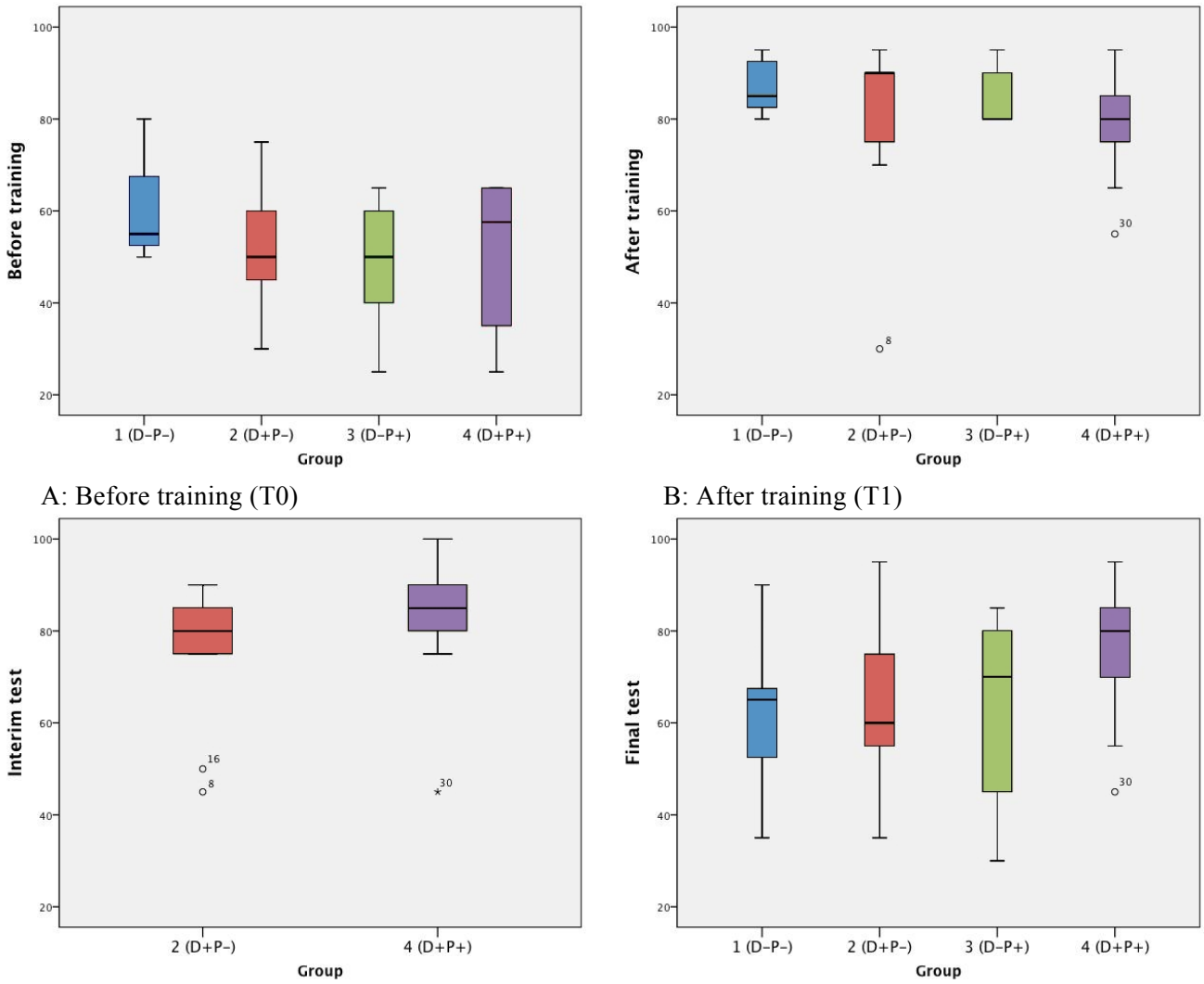



**Figure 13 – Both groups with the interim test (group 2 and group 4) show less decrease in test scores over time.**

There is a difference in test scores after training. Group 1 (D-P-) and group 3 (D-P+) score higher than group 2 (D+P-) and group 4 (D+P+). The decrease in test scores from the test after training to the final test is most obvious for group 1 (D-P-) and group 3 (D-P+). Group 2 (D+P-) shows less decrease over time compared to group 1 (D-P-) and group 3 (D-P+). This could be a first indication that interim testing has a beneficial effect on final test scores. Group 4 (D+P+) outperforms all the other groups, showing nearly the same test scores on the final test (T3) compared to test after training (T1). This provides another indication that interim testing has an effect. The test scores have been plotted split per group for further analysis (see Figure 14 A-D):




A: Before training (T0)

B: After training (T1)

C: Interim test (T3)

D: Final test (T4)

**Figure 14 (A-D) – Comparison of average group scores on all measurements. Boxplots identify multiple outliers and a relatively large spread for group 3 on the final test.**

Boxplots identify outliers in group 2 (D+P-) for the test after training and the final test, and an outlier in group 4 (D+P+) for all three measurements after training. The subject in group 4 scores structurally lower on all tests compared to other subjects, but the scoring pattern is consistent throughout the tests. One of the outliers in group 2 (D+P-) scored lower on all tests, there is no obvious reason for the other outlier in group 2 (D+P-) on the interim declarative test. What furthermore is interesting is the large variation in group 3 (D-P+) on the final test; notes made during the tests do not state any obvious reason. Due to the already low sample size no outliers have been removed from the data.

## Analyses

Based on the visual inspection of the raw data, a concern about differences in test scores before training arose. This concern is not supported by an ANOVA analysis on the start scores of the groups, $F(3,31) = 1.203$; $p = 0.325$. Looking at the raw data (see Table 1), there are signs there is a trend that interim testing has a beneficial effect on final declarative test scores. The combined score of the groups without the interim test scored 62.4 on the final declarative test with a test score of 86.8 after training. The groups with the interim test had a final test score of 70.0 with a score of 79.2 after training.

**Table 1 – Average scores on the test before training, after training and on the final test**

| Group | N | Before training | SD | After training | SD | Final test | SD |
|---|---|---|---|---|---|---|---|
| 1 (D-P-) | 8 | 60.7 | 12.1 | 87.1 | 6.4 | 61.4 | 17.3 |
| 2 (D+P-) | 9 | 53.3 | 13.7 | 79.4 | 20.2 | 64.4 | 18.6 |
| 3 (D-P+) | 10 | 47.2 | 14.8 | 84.4 | 5.8 | 63.3 | 20.0 |
| 4 (D+P+) | 10 | 52.5 | 15.1 | 79.0 | 11.7 | 75.5 | 15.2 |

The next step is to test the hypothesis that interim testing has a beneficial effect on the final test scores. A statistical model will be used with the declarative test scores as the response variable (Y) and with group as the explanatory factor (X). Since multiple response variables are available and the time trend needs to be estimated, a multivariate repeated measurement ANOVA model is used. Due to low group sizes, groups have been combined into two factors. The first factor combines group 1 (D-P-) with group 3 (D-P+), both without the interim declarative test. The second factor is a combination of group 2 (D+P-) and group 4 (D+P+), both with the interim declarative test. Based on the covariance matrices, it can be checked if there is similar variance between groups for the different measurements. Because the Box's M is not significant [Box's M = 18.064; $F(9,7862.770) = 1.764$, $p = .070$], there is no difference in the covariance between groups for the different measurements. Also Levene's Test shows there is no difference in variance between groups for the different measurements: T1, $F(3,31) = 2.262$, $p = .101$; T3, $F(3,31) = 0.541$, $p = .658$. This combined with the Box's M outcome (see above) indicates a legitimate approach.

The model shows that time is of significant influence, $F(1,31) = 32.65$, $p < .0005$. If the testing effect applies on this type of declarative knowledge, there should be a group effect over time. This would indicate that the interim declarative test has a beneficial influence on the long-term retention of declarative knowledge. There indeed is a significant interaction effect for the declarative test over time, $F(1,31) = 1.20$, $p = .019$. A similar effect does not show for the procedural test, $F(1,31) = 1.98$, $p = .169$, indication that groups that did the interim procedural test do not score different than groups that did not. Finally there might be an effect of doing both interim tests, however the variance analysis does not show a significant effect, $F(1,31) = .36$, $p = .551$. What these results show is that an interim declarative test has a positive influence on final declarative test scores.

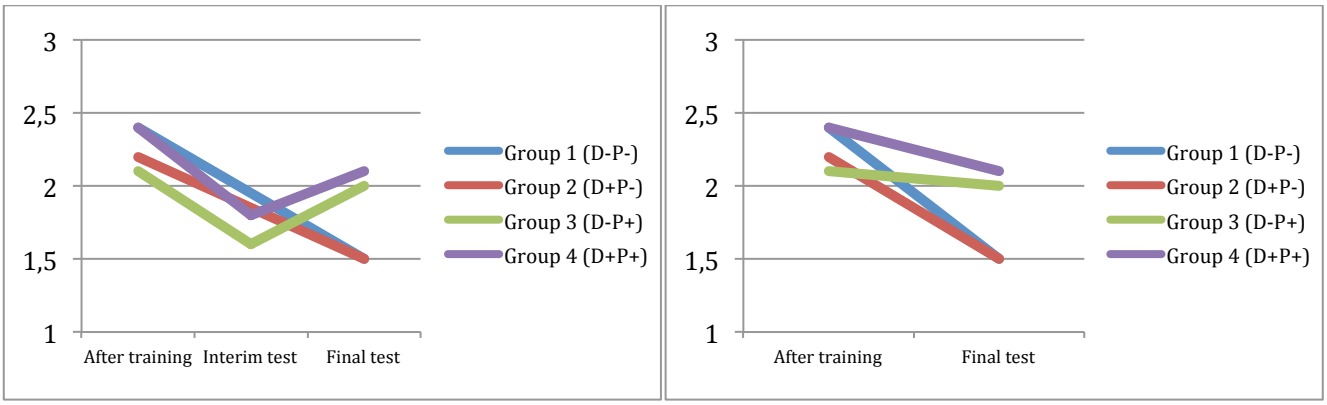
## *Procedural test*

### Merging the separate scores into on scale

Two cardiologists have assessed all echoes. In order to combine their scores into one average score per echo, the reliability between these scores has to be checked making sure echoes are scored in a similar fashion. Once this reliability statistic is at a certain level, the scores of the two different assessors can be combined to one average score per echo. The scores on the test after training and the final test have been checked; both the statistics on the test after training ($\alpha = .860$) and the final test ($\alpha = .812$) are well above 0.7. When this statistic exceeds 0.7, the scores are considered to be highly correlated and scores can be combined into an average score of both individual scores. Furthermore subjects were asked to make six different echoes in random order; the final test score on the procedural test is the average score on these six echoes.

**Visual inspection**

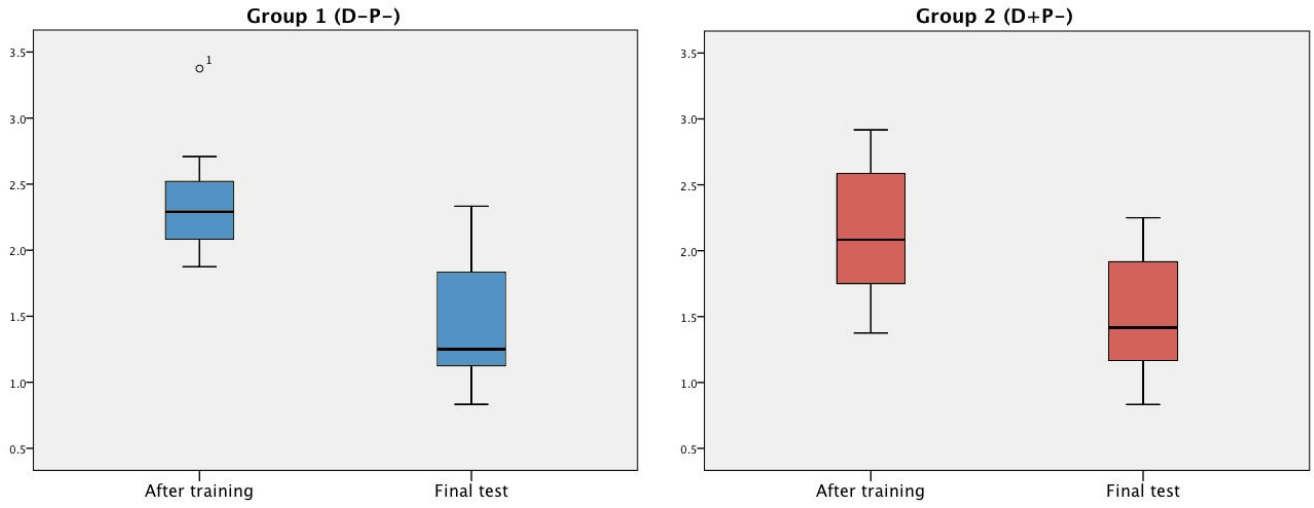
Again the starting point for the analysis of the procedural test scores is a visual inspection. The visual inspection suggests there are two different clusters. The first cluster contains group 1 (D-P-) and group 2 (D+P-) in which most subjects score lower on the final test than on the first test. The general trend in the data shows a decreasing score of about one point. The second cluster contains group 3 (D-P+) and group 4 (D+P+) some subjects show a roughly similar or higher score in comparison to the first test. These visual findings are in line with the hypothesis that the interim procedural test enhances final test performance. As the individual scores may contain noise, the average scores per group per measurement may provide a better insight in score patterns; see Figure 15 A-B for the average scores per group.



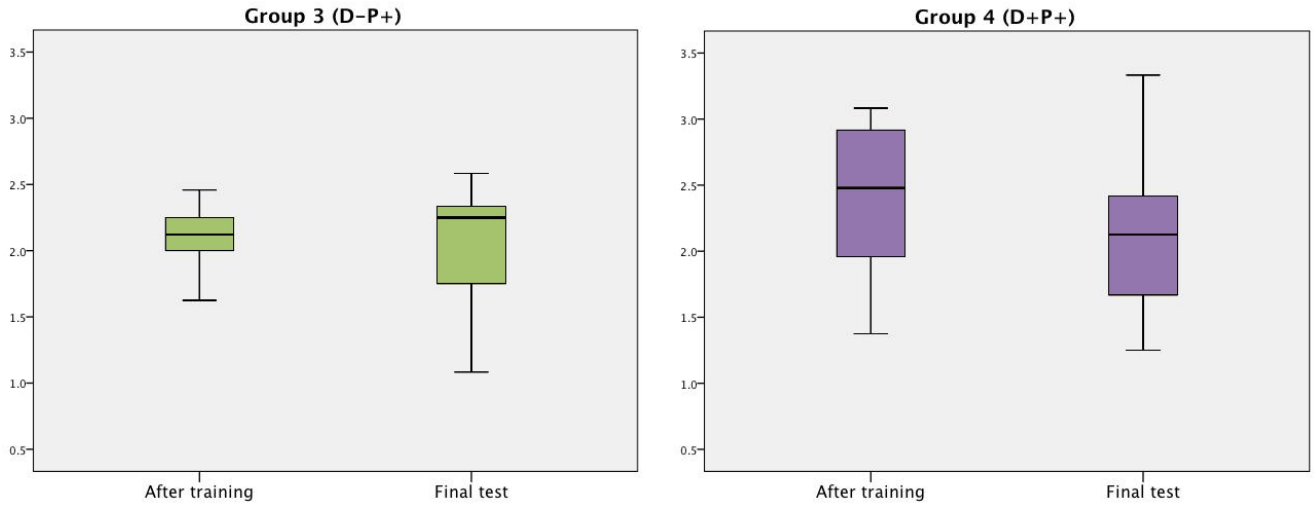

(A) (B)

**Figure 15 (A-B) – Comparison of average group scores per measurement. Group 1 and group 2 show a clear decrease in test performance. The groups with the interim test (group 3 and group 4) show less decrease in test performance over time.**

By looking at the average scores per group per measurement, there is a trend in the data in favour of the hypothesis that the testing effect can be applied on procedural skills. While there is a decrease in scores over time for the group that have not been retested, this decrease in performance is less for the retested groups, almost maintaining their initial level. Interestingly group 3 (D-P+) and group 4 (D+P+) have an interim test score close to the final test score (T3) of group 1 (D-P-) and group 2 (D+P-). The plots in Figure 16 A-D show the scores directly after training (T1) and on the final test (T3) per group:




(A) (B)




(C) (D)

**Figure 16 (A-D) – Comparison of the average group scores on the test after training and the final test. The groups with the interim test (group 3 and group 4) show less decrease in test performance over time.**

Again it shows that group 3 (D-P+) and group 4 (D+P+) perform better on the final test compared to group 1 (D-P-) and group 2 (D+P-). Furthermore the delta scores between the test after training and the final test are lower, indicating less decrease in performance over time. When combining group 1 (D-P-) with group 2 (D+P-) and group 3 (D-P+) with group 4 (D+P+), an even more clear view can be obtained. While the start scores for these combined groups are roughly similar (2.3), the final score for the retested groups is more than 0.5 higher compared to the other groups. Based on the visual inspection it seems that interim testing has a beneficial effect on long-term retention of procedural skills as well.

**Data inspection**

**Table 6 – Average group scores on test 1 and test 3.**

| Group | N | Score on T1 | SD | Score on T3 | SD |
|---|---|---|---|---|---|
| 1 | 8 | 2.4 | 0.5 | 1.5 | 0.5 |
| 2 | 9 | 2.2 | 0.5 | 1.5 | 0.5 |
| 3 | 10 | 2.1 | 0.3 | 2.0 | 0.6 |
| 4 | 10 | 2.4 | 0.6 | 2.1 | 0.6 |

The starting point for the data inspection is checking whether the scores after training (see Table 6) for the groups are the same. A one-way ANOVA has been done on the test scores after training, again using two factors. The first factor is a combination of the groups that did not do the interim procedural test, group 1 (D-P-) and group 2 (D+P-). The first factor combines the group that did the interim procedural test, group 3 (D-P+) and group 4 (D+P+). There is no significant difference in procedural test scores between groups directly after training, $F(1,33) = .82$, $p = .776$. Again it shows that having an interim test indeed has a beneficial effect on final test scores, $F(1,31) = 2.99$, $p = .006$. This gives a strong indication that the testing effect not only applies for declarative knowledge, but for procedural skills as well. Retesting on declarative knowledge has no effect on final procedural test scores, $F(1,31) = .05$, $p = .829$. Furthermore there is no interaction effect between the declarative test and procedural test, $F(1,31) = .44$, $p = .268$).

For both tests it shows that taking an interim test has a beneficial effect on the final test score of the same test type. Additionally it shows that an interim declarative test has no influence on final procedural performance and vice versa.

# Discussion

Most skills have to be learned for a long period, think of riding a bicycle for the rest of your life. In order to test whether the testing effect can be applied to skill acquisition, it makes sense to test over a relatively long period. This study provides a step towards verifying if the testing effect can be generalized to procedural skills by using an interim test at a spaced interval. The entire test set-up was stretched over a period of roughly three months. The study was conducted under controlled circumstances. Subjects were not exposed to the topic of echocardiography, besides encounters due to participation in this study. During the training sessions all subjects received exactly the same video instructions as an introduction to the basic anatomy and function of the heart and to echocardiography. The practical training was done in a similar fashion throughout the study; after the general instruction subjects received the same amount of time for free practice on the simulator. Assigning subjects to test groups over time controlled the effect of an increasingly more effective training. This study had two research questions, namely:

1. Can the testing effect be generalized to procedural skills?
2. Can long-term retention of procedural skills be enhanced when improving declarative knowledge by applying the testing effect?

Before answering the first research question if the testing effect can be generalized to procedural skills, the declarative test results will be discussed. The average scores per group are relatively high for all measurements. Typically people forget much of their knowledge within the first few days after acquisition. The test scores of the groups that did not take the interim test are still at a very reasonable level and for all groups higher than the test scores before training. The most probable explanation is that all subjects follow the Medicine bachelor's programme, are functioning in a medical context daily and were voluntarily participating in this study.

The testing effect explains that the effect of interim testing is more beneficial and goes beyond the effect of restudying. As expected when subjects take an interim test, between the test after training and the final test, subjects in those groups score significantly higher on the final declarative test, $F(1,31) = 1.20$, $p = .019$. Do note that the retested groups have not been compared to interstudy groups that restudied the test materials. Nevertheless the results do show a beneficial effect of the interim test compared to groups that did not take an interim test. Based on that outcome and examples of testing effect studies described in the theoretical background, it is highly likely the effect indeed is the testing effect and goes beyond the effect of restudying. The testing effect has been demonstrated on multiple types of knowledge, but not yet on medical facts. This study demonstrated that that the testing effect can be applied to yet another form of declarative knowledge, showcasing another example of the robustness of this effect.

Coming to an answer to the first research question, it seems interim testing has a beneficial effect on retention of skills as well. After analysing the procedural test scores, it also shows that retesting at that same spaced interval has a beneficial effect on the long-term retention of procedural skills. The groups that took the interim test showed significantly higher test scores in comparison with the groups that did not take the interim test, $F(1,31) = 2.99$, $p = .006$). The test scores of the groups that did the interim test were roughly at the same level of the final test scores of the groups that were not retested. This could indicate that the highest decline in skill performance occurs during the first three weeks after training. The final test scores show that, assuming the found effect goes beyond restudying, the testing effect can be generalized to procedural skills.

So it seems that interim testing indeed has a beneficial effect on both declarative knowledge and procedural skills. And while this is in line with the hypothesis, the main discussion point is that because of the experimental design, conclusions only about the potential benefits of an interim test can be made. It is not

possible to determine if this effect goes beyond the effect of restudying. The magnitude of the effect in comparison with restudying is unknown; hence no final conclusion can be drawn if the effect applies. The results of this study do not show a difference in the effect for either declarative knowledge or procedural skills, showing an interesting area for further research. The performance level was higher for the declarative test, which can be explained by the fact that it generally requires more time to acquire skills. In despite of the fact that no final conclusion can be drawn if the testing effect applies on procedural skills, there are strong indications the testing effect can be generalised to skill acquisition. Based on both the outcomes of this study and the cited literature about the testing effect, it is likely this effect goes beyond spending a similar time on practicing.

Subjects received feedback after both tests, which raises the question whether the found effects are caused purely due to the triggered recollection of the productions needed to make the echoes or if it is due to the additional feedback. Previous studies have shown that feedback can have a positive influence on the magnitude of the testing effect (e.g. Butler & Roediger III, 2008). If feedback increases the testing effect on procedural skills has not yet been studied specifically. As discussed before the testing effect has been demonstrated on declarative knowledge using tests both with and without feedback. It is unknown whether feedback has a similar effect on the acquisition and retention of skills. The hypothesis is that feedback does have a positive influence on the testing effect. Where interim testing without feedback already has a beneficial effect; giving feedback can increase the magnitude of the testing effect.

Looking at the second research question about the potentially beneficial effect of enhanced declarative knowledge on the final performance on the procedural test: (1) The procedural test was created in such a way that little declarative knowledge was required. (2) The declarative test was created in such a way that only little procedural skills were required. Even though declarative knowledge is required to perform the procedural test, one for instance has to recollect what a specific echo looks like, the test were not strongly related. Where both test related sufficiently to answer the second research question? At this point I think both tests should be more related to each other. There were some questions on how to manipulate the probe to switch from one view to another, however the biggest part of the test was not related to the procedural test at all. Based on these data there seems to be no enhanced performance in groups that both took the interim declarative and procedural test. By better connecting both tests, it is more likely a connection between both tests will be found; in follow-up studies tests should be stronger connected.

Finally, not all subjects that were planned for the interim tests were tested; subjects were randomly assigned to other groups to overcome the problem of unequal group sizes. Also not all subjects took the interim test at exactly the same interval; the interval between the first tests and the interim tests was between 2-4 weeks.

# Relevance for HMC

The main practical goal of the current study was defining the optimal way to train students with a simulator in such a way that participants are better prepared for their work while using as little resources as possible to train them. The main theoretical goal was determining whether the testing effect applies to procedural knowledge as well and to see if enhanced long-term retention of procedural skills can be achieved by applying this effect. The impact of the finding that it seems the testing effect can be generalized to procedural skills has an impact on the way current learning models are applied in many situations. From a practical perspective adding an interim test to any training procedure would have a beneficial effect on long-term without demanding additional study time, lowering the effort for students while still performing well. Additionally this study provides valuable insights in how study performance can be enhanced without spending more resources on training.

## *Cognitive models*

One of the aims of Human-Machine Communication (HMC) is providing students with fundamental insights into human cognition (add source: ). This knowledge is then used in applied settings, such as human-computer and human-robot interaction, speech technology and tutoring systems. Especially in the case of tutoring systems, current findings might help to improve the efficiency. The rationale behind the testing effect can be implemented in tutoring systems, enabling alternative learning scenarios in which learning is more efficient I comparison with the current situation in which restudying is used. When computer systems 'know' to apply the testing effect, the learning outcome can be improved by adding tests to the learning model. The outcomes of this study can eventually be used to model the acquisition and retention of different types of knowledge. It has provided valuable insights in how we can create more accurate computer models that can learn in a more natural way. Additionally it will help to better model learning in such a way that it will possible to better predict learning and retention curves.

### Skill Acquisition

Skills are at the base of all human activities. When going to work in the morning, the routine from getting up to taking place behind your desk is full of activities such as for instance preparing breakfast, making coffee, driving your car and many more. The required skills are embedded in our routine in a seamless way, that we often take our skills for granted. Skills vary from opening doors to operating a complex airplane. The more we understand about skill acquisition, basic skills, expert skills and the process from one level the another, the better we can model that knowledge into a model that explain skill acquisition in a more efficient and realistic way. One of the applications would be in the ACT-R architecture. ACT-R has been described in the theoretical background, describing how productions are formed and how skills can be transferred. When further modelling skill acquisition the testing effect could be modelled. By doing so predictions of learning speed and retention could be improved. By implementing current findings, the way machines learn or the way human cognitive performance is modelled and predicted can be done even more realistic. Adding the testing element, and therefore making learning more efficient could influence the neural activation of these productions; i.e. not merely repetitions account for higher activation, active retrieval as required in a test increased neural activation as well. This might give better predictions about different learning situations.

# Future work

An initial step for future work could be to redo this study, using the same experimental set-up with more subjects. With the final number of subjects in this study, there was insufficient power in the statistical test to compare between groups; adding subjects will overcome this problem and more detailed analyses can be done. Furthermore what would be very interesting is to add an additional measurement afterwards. This extra measurement after a few months can help to identify the retention curve. As discussed in the discussion the interim scores on the procedural test are roughly on the same level as the final test scores of the two groups that did not take the interim test. Again it would be very interesting to do additional measurements to gain input for the retention curve.

To answer the question if better declarative knowledge enhances procedural performance, it is important to design tests in such a way that subjects with better declarative knowledge indeed have an advantage for the procedural test. Consequently the declarative test and procedural test should be designed in such a way that there is a stronger relation. Assuming future studies will be done in the same context using an echocardiogram simulator, the following gives an impression of the changes: Rather than merely on how to obtain the different views, the training sessions could focus more on the interpretation of echoes. The current declarative test focused mainly on recognizing structures. This test could be modified in such a way that more questions are asked that require the interpretation of echoes. The procedural test could be designed in such a way that subjects receive a written case and need to obtain the proper images for a diagnosis. By doing so subjects need to use their declarative knowledge to select the appropriate echo. These changes would greatly help in answering the question whether better declarative knowledge enhances procedural performance or not. It would also help if the skill level of subjects would be higher after the training. This would create circumstances in which different treatments will probably elicit higher differences between groups. While all subjects had equal time on the simulator, the time they had was very limited taking into account the difficulty level of making echoes. For future studies it is firmly recommended doubling the time per subject.

The previous point focused mostly on improving the experimental set-up as used in this study. There are several specific suggestions for future studies: To learn more about the effect of feedback it would be highly interesting to conduct a similar study with groups with or without test feedback. Assuming feedback has a beneficial effect on the magnitude of the testing effect, i.e. by giving feedback the testing effect becomes stronger, and by comparing the groups with or without feedback it becomes obvious what the exact influence is. Furthermore it will help in answering the question if the same effect of test feedback applies for the procedural test, and if yes, what the added benefit of feedback is.

To learn more about the magnitude of the influence of interim testing, a similar study would be useful that includes restudy groups. What these groups exactly need to study should be based on the experimental set-up. In the current set-up, as a replacement for the procedural test, subjects could for instance study how to manipulate the probe in general, how to obtain views and how to switch from one views to another.

# Works Cited


Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology, 55* (1), 25-35.

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, 159-177.

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger III, H. L., & McDermott, K. B. (2008). Examining the Testing Effect with Open- and Closed-Book Tests. *Applied Cognitive Psychology, 22*, 861-876.

Anderson, J. R. (1976). *Language, Memory, and Thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1983). *The architecture of cognition.* Cambrudge, MA: Harvard University Press.

Anderson, J. R., Fincham, J. M., & Douglas, S. (1997). The Role of Examples and Rules in the Acquisition of Cognitive Skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 932-945.

Butler, A. C., & Roediger III, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36* (3), 604-616.

Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, 514-527.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory research. *American Psychologist*, 1185-1193.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. B. Resnick, *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 361-392). Hillsdale, NJ: Lawrence Erlbaum Associates.

Berden, H. J., Willems, F. F., Hendrick, J. M., Pijls, N. H., & Knape, J. T. (1993). How frequently should basic cardiopulmonary resuscitation training be repeated to maintain adequate skills?. *BMJ , 306*, 1576-1577.

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In Search of Decay in Verbal Short-Term Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35* (2), 317-333.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co Inc.

Boet, S., Borges, B. C., Naik, V. N., Siu, L. W., Riem, N., Chandra, D., et al. (2011). Complex procedural skills are retained for a minimum of 1 yr after a single high-fidelity simulation training session. *British Journal of Anaesthesia*, 533-539.

Bose, R., Matyal, R., Panzica, P., Karthik, S., Subramaniam, B., Pawlowski, J., et al. (2009). Transesophageal Echocardiography Simulator: A New Learning Tool. *Journal of Cardiothoracic and Vascular Anesthesia, 23* (4), 544-548.

Cull, W. L. (2000). Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall. *Applied Cognitive Psychology, 24*, 215-235.

CAE Healtcare. (2013, March 01). *Product & Services: VIMEDIX.* Retrieved March 12, 2013 from CAE Healtcare: https://caehealthcare.com/home/product_services/product_details/vimedix

Cantù, P., & Penagini, R. (2012). Computer SImulators: The present and near future of digestive endoscopy. *Digestive and Liver Disease, 44*, 106-110.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the Testing and Spacing Effects to Name Learning. *Applied Cognitive Psychology*, 619-635.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulleting & Review, 14*, 474-478.

Crowder, R. G. (1976). *Principles of learning and memory.* Abingdon: Lawrence Erlbaum Associates.

Croley, W. C., & Rothenberg, D. M. (2007). Education of trainees in the intensive care unit. *Critical Care Medicine, 35*, 117-121.

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology.* (H. A. Ruger, & C. E. Bussenius, Trans.) New York: Teachers College.

Daniel, D. B., & Poole, D. A. (2009). Learning for life: An ecological approach to pedagogical research. *Perspectives on Psychological Science, 4*, 91-96.

de Groot, S., Centeno Ricote, F., & de Winter, J. C. (2012). The effect of tire grip on learning driving skills and driving style: A driving simulator study. *Transportation Research Part F*, 3413-426.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and Testing Effects: A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature. In B. H. Ross, *The Psychology of Learning and Motivation Volume 53* (pp. 63-148). Burlington: Academic Press.

Dreyfus, S. E., & Dreyfus, H. L. (1980). *A Five Stage Model of the Mental Activities Involved in Skill Acquisition.* Berkeley: University of California.

Ferman, S., Olshtain, E., Schechtman, E., & Kami, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics, 22* (4), 384-412.

Gates, A. I. (1917). *Recitation as a factor in memorizing.* New York: The Science Press.

Izawa, C. (1992). Test trial contributions to optimization of learning processes: study/test trials interactions. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin, *From Learning Theory to Connectionist Theory* (pp. 1-33). Hillsdale: Erlbaum.

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher, 27*, 10-28.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7* (2), 185-207.

Hillis, G. S., & Bloomfield, P. (2005). Basic transthoracic echocardiography. *BMJ , 330*, 1432-1436.

Kunkler, K. (2006). The role of medical simulation: an overview. *The International Journal of Medical Robotics and Computer Assisted Surgery, 2*, 203-210.

Kang, S. H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38* (8), 1009-1017.

Kirkpatrick, K., & MacKinnon, R. J. (2012). Technology-enhanced learning in anaesthesia and educational theory. *Continuing Education in Anaesthesia, Critical Care & Pain, 12* (5), 263-267.

Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills learning might last 6 months. *Advances in Health Sciences Education, 15* (3), 395-401.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning . *Medical Education, 43*, 21-27.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition.* (J. D. Bransford, A. L. Brown, R. R. Cocking, M. S. Donovan, & J. W. Pellegrino, Eds.) Washington D.C.: National Acadamy Press.

Neelankavil, J., Howard-Quijano, K., Hsieh, T., Ramsingh, D., Scovotti, J. C., Chua, J. H., et al. (2012). Transthoracic Echocardiography Simulation Is an Efficient Method to Train Anesthesiologists in Basic Transthoracic Echocardiography Skills. *Anesthesia & Analgesia , 115* (5), 1042-1051.

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64* (5), 482-488.

Maiss, J., Naegel, A., & Hochberger, J. (2011). The European experience - current use of simulator training in Europe. *Techniques in Gastrointestinal Endoscopy, 13*, 126-131.

Marsick, V. J., & Watkins, K. E. (2001). Informal and Incidental Learning. *New Directions for Adult and Continuing Education* (89), 25-34.

McDonald, R. J., Devan, B. D., & Hong, N. S. (2004). Multiple memory systems: The power of interactions. *Neurobiology of Learning and Memory, 82* (3), 333-346.

McGeoch, J. A. (1943). *The psychology of human learning.* New York: Longmands Green.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in Immediate Serial Recall: Decay, Temporal Distinctiveness, or Interference? . *Psychological Review, 115* (3), 544-576.

Okuda, Y., Bryson, E. O., DeMaria Jr, S., Jacobson, L., Quinones, J., Shen, B., et al. (2009). The Utility of Simulation in Medical Education: What Is the Evidence? *Mount Sinai Journal of Medicine, 76*, 330-343.

Page, R. L. (2000). Brief History of Flight Simulation. *SimTechT 2000 Proceedings* (pp. 1-11). Sydney: The SimTechT 2000 Organizing and Technical Committee.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2006). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14* (2), 187-193.

Perkins, D. N., & Salomon, G. (1992). Transfer of Learning. In *International Encyclopedia of Education, Second Edition.* Oxford, England: Pergamon Press.

Platts, D. G., Humphries, J., Burstow, D. J., Anderson, B., Forshaw, T., & Scalia, G. M. (2012). The Use of Computerised Simulators for Training of Transthoracic and Transoesophageal Echocardiography. The Future of Echocardiographic Training? *Heart, Lung and Circulation, 21*, 267-274.

Portrat, S., Barrouillet, P., & Camos, V. (2008). Time-Related Decay or Interference-Based Forgetting in Working Memory? *Journal of Experimental Psychology: Learning, Memory and Cognition, 34* (6), 1561-1564.

Schacter, D. L., Gilbert, D. T., & Wegner, D. M. (2011). *Psychology.* New York: Worth Publisher.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *The Journal of Neurology, Neurosurgery and Psychiatry, 20* (1), 11-21.

Singley, M. K., & Anderson, J. R. (1985). *The transfer of cognitive skill.* Camebridge, MA: Harvard University Press.

Shakil, O., Mahmood, F., & Matyal, R. (2012). Simulation in Echocardiography: An Ever-Expanding Frontier. *Joarnal of Cardiothoracic and Vascular Anesthesia, 26* (3), 476-485.

Squire, L. R. (1992). Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory. *Journal of Cognitive Neuroscience, 4* (3), 232-243.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82* (3), 171-177.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences* (pp. 13515-13522). National Academy of Sciences.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science, 17* (3), 249-255.

Roediger III, H. L., & Karpicke, J. D. (2006). The Power of Testing Memory - Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science, 1* (3), 181-210.

Taatgen, N. A. (2013). The Nature and Transfer of Cognitive Skills. *Psychological Review , 120* (3), 439-471.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition, 86* (2), 123-155.

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General, 137* (3), 548-565.

The Cardiac Society of Australia and New Zealand. (2009, November 27). *CSANZ Training Guidelines in Adult Echocardiography.* Retrieved November 13, 2012 from CSANZ Training Guidelines in Adult Echocardiography: http://www.csanz.edu.au/Portals/0/Guidelines/Training/Guidelines%20for%20Training%20and%20Competence%20-%20Training%20Guidelines%20in%20Adult%20Echocardiography%20(2009).pdf

Underwood, G., Crundall, D., & Chapman, P. (2011). Driving simulator validation with hazard perception. *Transportation Research Part F*, 435-446.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition* (109), 163-167.

Yan, X., Abdel-Aty, M., Radwan, E., Wang, X., & Chilakapati, P. (2008). Validating a driving simulator using surrogate safety measures. *Accident Analysis and Prevention, 40*, 274-288.

Zaromb, F. M., & Roediger III, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38* (8), 995-1008.

Ziv, A., Small, S. D., & Wolpe, P. R. (2000). Patient safety and simulation-based medical education. *Medical Teacher, 2* (5), 489-495.

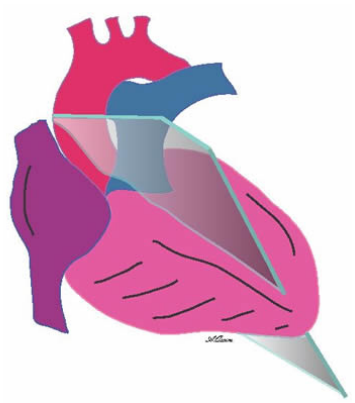
# Appendix 1 - Views and instructions on how to obtain them

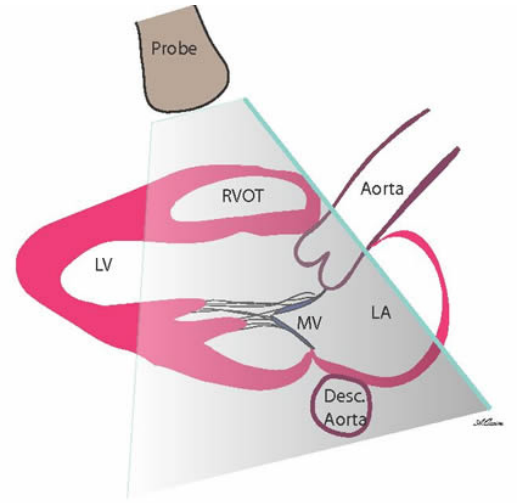
Instructions on the different views and how to obtain them have been shared during the training sessions.

## *Parasternal*

**Long axis:**

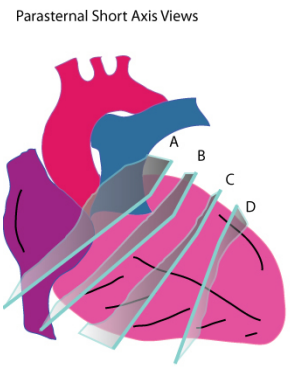
- Transducer position: left sternal edge; 2nd to 4th intercostal space.
- Marker dot direction: points towards right shoulder.

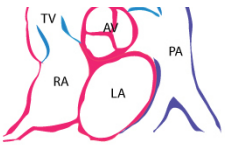


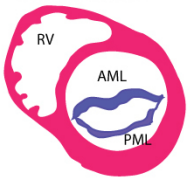


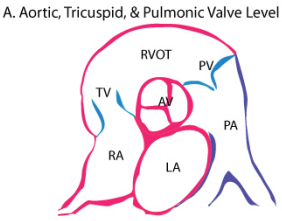

**Short axis:**

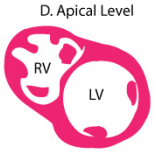
- Transducer position: left sternal edge; 2nd to 4th intercostal spaces
- Marker dot direction: points towards left shoulder (90 degrees clockwise from PLAX view)

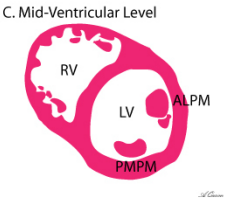


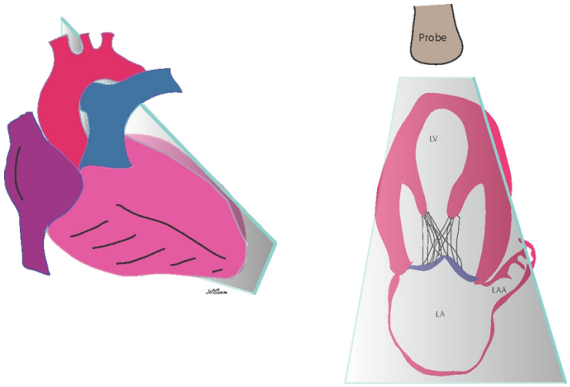



## *Apical*

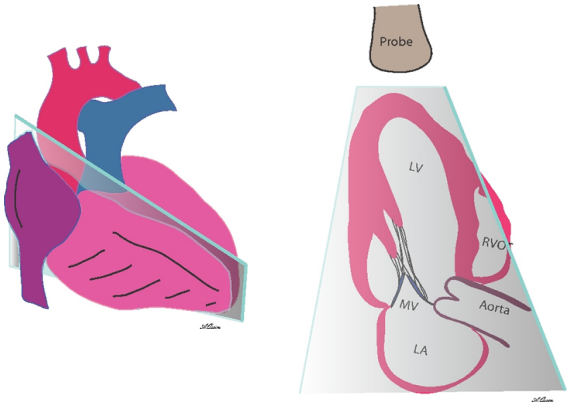
**2 chamber:**

- Transducer position: apex of the heart
- Marker dot direction: points towards left side of the neck (roughly 45 degrees anti-clockwise from A4C view).



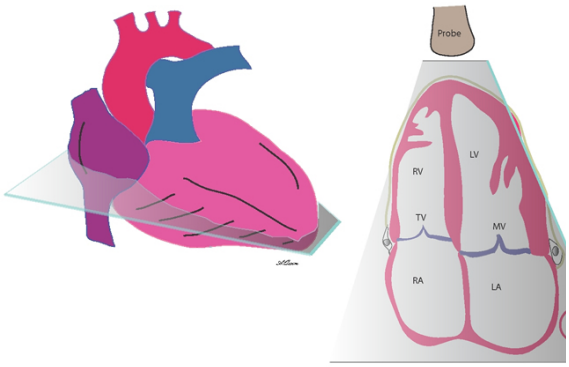

**3 chamber:**

- Transducer position: apex of the heart
- The transducer is placed at the same position as for a A4C view and then turned clockwise by 60°
- Structures visualized: It is similar to a parasternal long axis view seen from the apex and characterized by the presence of the mitral and aortic valves in the same plane.



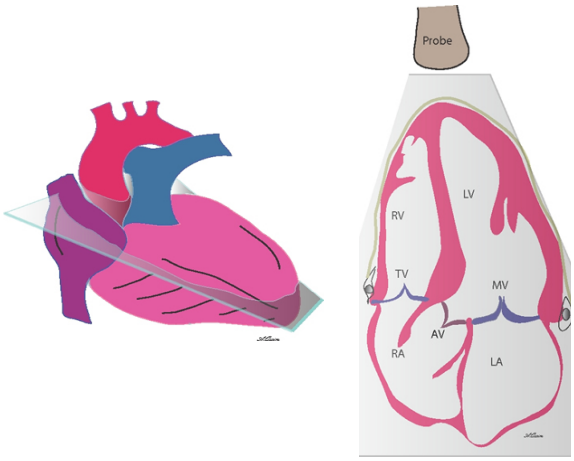
**4 chamber:**

- Transducer position: apex of the heart
- Marker dot direction: points towards left shoulder.




**5 chamber:**

- Transducer position and marker dot direction are same as the A4C view.
- The A5C view is obtained from the A4C by slight anterior angulation of the transducer towards the chest wall.
- The 5th chamber added is the aorta.

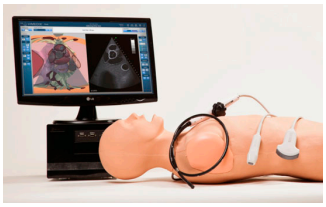


# Appendix 2 – Information letter students

Geachte student,

Zoals in alle academische ziekenhuizen wordt ook in het Universitair Medisch Centrum Groningen onderzoek gedaan. Dit onderzoek kan gericht zijn op het vinden van betere methoden om een ziekte vast te stellen, maar bijvoorbeeld ook op het verbeteren van ons onderwijs. Dergelijk onderzoek is alleen mogelijk met de medewerking van proefpersonen. Daarom vragen wij jouw hulp om aan een onderzoek mee te werken. We willen je in deze brief graag informatie geven over het doel van het onderzoek, de te gebruiken onderzoeks-procedure en de voor- en nadelen ervan. Deelname aan het onderzoek is vrijwillig. Als je deze informatie gelezen hebt en hierover nog vragen hebt, dan kun je contact opnemen met Stephan Mooibroek op s.c.mooibroek@umcg.nl.

## Wetenschappelijk onderzoek



Voor dit onderzoek zal er gebruikt worden gemaakt van de echocardiogram-simulator in het Skills Center. Het onderzoek zal zich richten op het testen van verschillende manieren waarmee het behoud van kennis, in relatie tot de vaardigheid van het afnemen van echo's en kennis van de anatomie en functie van het hart, verbeterd kan worden. Dit zal worden gedaan door vier groepen op verschillende wijze te onderwijzen. Door de resultaten van deze vier groepen onderling te vergelijken kunnen we nagaan welke methode het beste is. Toekenning aan een bepaalde groep wordt door loting bepaald.

Het onderzoek zal binnen een periode van 4 maanden worden uitgevoerd. In die periode is het vereist om één dagdeel aanwezig te zijn voor een training. Deze training begint met een instaptoets, vervolgens komt de theorie aan bod, en daarna oefenen proefpersonen op de simulator. Tot slot worden er twee toetsen afgenomen. Drie van de vier groepen gevraagd om na een periode van ongeveer vier weken na de initiële training terug te komen voor een toets. Tot slot zal iedereen, na een periode van 4 maanden, nogmaals getoetst worden.

## Wat betekent het meedoen voor jou?

Wij bieden de mogelijkheid om op een laagdrempelige manier kennis op te doen over de anatomie en functie van het hart. Daarnaast doe je ervaring op met het maken van een echocardiogram. Dit tezamen biedt een goede ondersteuning voor het begrijpen van de anatomie en fysiologie van het hart. Het enige wat we hiervoor vragen is om een aantal uren (in totaal ongeveer 4 uur) te investeren in deelname. Daar staat tegenover dat je met relatief weinig inspanning en weinig tijd veel nuttige ervaring kan opdoen met echocardiografie. Daarnaast bieden wij je een kleine vergoeding van €10 aan in de vorm van een cadeaubon.

## Vertrouwelijkheid van de gegevens

Alle gegevens die ten behoeve van het onderzoek worden verzameld zullen vertrouwelijk worden behandeld. Jouw naam zal nooit openbaar worden gemaakt. De resultaten van het onderzoek kunnen worden gepubliceerd in wetenschappelijke artikelen, maar ook hierin worden geen namen genoemd. Alle data zal geanonimiseerd opgeslagen worden.

## Vrijwilligheid van deelname

Je bent er uiteraard geheel vrij in om aan dit onderzoek mee te doen. Verder heb je te allen tijde, ook wanneer je schriftelijk hebt verklaard te willen deelnemen, het recht om zonder opgave van redenen af te zien van verdere deelname aan het onderzoek.

## Ondertekening toestemmingsverklaring

Als je besluit mee te werken aan het onderzoek zullen wij je vragen een formulier te ondertekenen. Met deze toestemmingsverklaring ('Informed consent') bevestig je jouw voornemen om aan het onderzoek mee te werken. Je blijft de vrijheid behouden om wegens voor jouw relevante redenen de medewerking te stoppen. De onderzoeker zal het formulier eveneens ondertekenen en bevestigt daarmee dat hij jou heeft geïnformeerd over het onderzoek, deze informatiebrief heeft overhandigd en bereid is om waar mogelijk in te gaan op nog opkomende vragen.

## Nadere informatie

Mocht je na het lezen van de brief, voor of tijdens het onderzoek nog nadere informatie willen ontvangen of komen er nog vragen bij je op dan kunt je altijd contact opnemen met de uitvoerder van het onderzoek, Stephan Mooibroek, per e-mail te bereiken via s.c.mooibroek@umcg.nl.

Met vriendelijke groet,
Stephan Mooibroek

# Appendix 3 – Declarative test

**What structure/area is under the 'X'?**
Fourteen videos with moving 2D heart images have been shown, stopping at a certain point with a cross at the structure/area that had to be named.

**When does the pulmonary valve close?**
**A. When pulmonary artery pressure rises above right ventricular pressure.**
B. When pulmonary artery pressure dives under right ventricular pressure.
C. When pulmonary artery pressure rises above left atrial pressure.
D. When pulmonary artery pressure dives under left atrial pressure.

**When does the aortic valve open?**
A. When left ventricular pressure rises above left atrial pressure.
**B. When left ventricular pressure rises above aortic pressure.**
C. When right ventricular pressure rises above aortic pressure.
D. When right ventricular pressure rises above left atrial pressure.

**What is the main purpose of the heart valves?**
**A. To prevent backflow.**
B. To maintain forward flow of blood throughout the cardiac chambers.
C. To guarantee coronary blood flow.
D. To improve myocardial contractibility.

**How should you manipulate the probe in order to change the 2-chamber view to a 4-chamber view?**
*Note: To decrease the difficulty level of the procedural test, ribs/lungs/other tissue were disabled in the simulator. Due to the extra freedom, the required rotation could be somewhere between 45 degrees and 90 degrees.*
**A. Rotate the probe 90 degrees**
**B. Rotate the probe 45 degrees**
C. Increase the angle
D. Decrease the angle

**How should you manipulate the probe in order to change the short axis view to the long axis view?**
A. Increase the angle of the probe
B. The long-axis view is taken from another position
C. Rotate the probe 90 degrees
**D. Rotate the probe 90 degrees and adjust the angle slightly**

**How should you manipulate the probe in order to change the 4-chamber view to the 5-chamber view?**
*Note: Subjects have been instructed whether to view from inner or outer angle.*
A. Rotate the probe 45 degrees
B. Rotate the probe 90 degrees
C. Increase the angle
**D. Decrease the angle**