Eirik Sørgaard Aksdal
Sebastian Grøvdal Schaanning
(Both students contributed equally to the paper)

Learning habits of medical students at the Norwegian University of Science and Technology, and the application of Serious Games in learning dermatovenereology.

Graduate thesis in Medicine
Supervisor: Jørgen A. Urnes; Brita S. Pukstad
Trondheim, January 2017

Norwegian University of Science and Technology
Faculty of Medicine

NTNU
Norwegian University of
Science and Technology

Abstract

Introduction
The interest in so called “Serious Games” (SGs) in education has increased, with several new RCTs in just the last few years, but there is not yet a clear consensus as to the potential benefit of using Serious Games in education. The objective of our study was to evaluate a Serious Game entitled “Save Your Skin” (SYS). We sought to contribute to the literature on SGs, and their application to higher education, more specifically the teaching of medicine. We also wanted to assess 4th year medical students’ opinions on the learning modalities available to them in terms of usage, degree of motivation, trustworthiness and relevance to exams.

Methods
Students were randomized to receive 100 multiple choice questions (MCQs) in dermatovenereology either as electronic flashcards or in the serious game SYS. Each student’s knowledge in the subject was assessed with 20 MCQs both before and after the intervention. Electronic questionnaires were used to assess students’ reading habits and subjective opinions on the intervention they received.

Results
There was no significant difference between the game group and the flashcard group regarding knowledge acquisition in dermatovenereology. The subjective post-intervention evaluation favored the flashcard setup over the gaming set up. Regardless of group, the students generally reported that they preferred the intervention they received over traditional teaching methods. Reviewing lectures scored consistently high regarding time usage, trustworthiness and relevance for exams on the pre-intervention questionnaire. There were only small differences in degree of motivation when comparing the different learning modalities.

Conclusion
4th year medical students at NTNU generally favor reviewing lectures for learning the curriculum, although it is not known whether this is related to the quality of lectures, or if it follows the focus on lectures in the course structure. Due to the vast heterogeneity of serious games, we cannot on the base of this study dismiss serious games to be used in a learning context. However, the development of serious games is of high economic and academic cost, and it is doubtful whether this is a cost effective way of spending faculty resources. In future research and development of SGs, we recommend identifying success factors of SGs whose efficacy have been proven in high quality studies.

1. Introduction

The term “Serious Games” (SGs) is often defined as *games that do not have entertainment, enjoyment or fun as their primary purpose* (Michael and Chen 2005). The term usually refers to digital games (i.e. games played on a computer, phone etc.), but may also be used to refer to non-digital games. Although it has been argued that digital games fitting this definition can be traced as far back as the 1950s, the current “wave” of development and interest in serious games appears to have begun in 2002 (Djaouti et al. 2011).

To put it simply, the development of SGs for learning can be seen as an attempt to harvest the enormous appeal of video games in a learning context. Several elements of computer games, such as player immersion, dynamic problem-solving and interactivity can be viewed as desirable elements in the learning process as well. It has been pointed out that educational games have often failed to adequately include such elements, succumbing to a “practice makes perfect”-approach, and stripping away most of the perceived beneficial aspects of video games. Games of this kind, packing boring drills and monotonous cramming within a thin video game shell has been pejoratively dubbed a “chocolate covered broccoli” (Green 2014).

Although the interest in SGs has increased, with several new Randomized Controlled Trials (RCTs) the last few years, there is not yet a clear consensus regarding the potential benefit of using SGs in education. A systematic review by Connolly et al. from 2012 explored the literature on SGs in an attempt to establish the effect of SGs on several parameters including *knowledge acquisition/content understanding* and *affective and motivational outcomes*. For knowledge acquisition, there were few RCTs and the evidence they provided about the impact of serious games was mixed. Regarding the motivational aspect of educational games, results were also mixed (Connolly et al. 2012). An update to this review by Boyle et al. was published in 2016, looking at new studies published from 2009-2014. Regarding knowledge acquisition, the update included 7 RCTs, and they tended to report better outcomes for game playing than for the control condition. Regarding motivational properties of serious games, the review mostly included general research into how motivational processing relates to satisfaction with games. (Boyle et al. 2016)

One RCT pertaining to the use of SGs in medical education was discussed in the review by Connolly et al. In this study, Sward et al. showed that fourth year medical students who used a web based game to learn pediatrics over a period of 4 weeks did not perform significantly better on a post-intervention knowledge test than students who used a computerized flashcard approach to the material. However, perceptions about game playing versus self-study as a pedagogical method significantly favored game playing in understanding content, perceived help with learning, and enjoyment of learning. (Sward et al. 2008)

Another example of an RCT pertaining to the use of serious games in medical education was included in the updated review by Boyle et al. (Boyle et al. 2016). Knight et al. used serious game technology in major incident triage training for medical clinicians (Knight et al. 2010). They showed that serious game technologies have the potential to enhance knowledge and skill acquisition. 91 learners were randomly distributed into one of two training groups: 44 participants practiced triage sieve protocol using a card-sort exercise, whilst the remaining 47 participants used a serious game. In a post intervention test, performance was assessed in terms of tagging accuracy (assigning the correct triage tag to the casualty), step accuracy (following correct procedure) and time taken to triage all casualties. Students who used the game scored significantly better in the first two areas than students who trained using a card-sort exercise, while there was no significant difference between the groups in time taken to triage all casualties.

Connolly et al. stated in their review that they had initially planned to conduct a meta-analysis for impact and outcomes of playing games (Connolly et al. 2012). However, because of large heterogeneity in the material, they concluded that a narrative review was most appropriate. Despite the difficulty arising from heterogeneity of data, as described by Connolly et al., another group of researchers undertook such a meta-analysis in 2013. Serious games were assessed by Wouters et al. in terms of learning outcome and degree of motivation in comparison to traditional learning methods. They found serious games to be more effective in terms of learning and retention, but not more motivating than conventional instruction methods (Wouters et al. 2013).

The objective of our study was to evaluate a Serious Game entitled “Save Your Skin” (SYS). We wanted to contribute to the literature on SGs, and their application to the teaching of medicine, specifically dermatovenereology. We were interested to see if we could replicate the results of other studies, and to obtain data which could determine the viability of using SGs at the Faculty of Medicine at the Norwegian University of Science and Technology (NTNU). SYS was originally a SG made for Swiss medical students, and was available through an online learning portal for medical students and doctors called *Dermatology Online with Interactive Technology* (DOIT) (Burg 2016). The founder of DOIT, Günter Burg M.D., was contacted to see if this game could be translated and customized to fit Norwegian medical students. Dr. Burg accepted this request, and technical manager at DOIT, Vahid Djamei, was given the task to do this, according to instructions from Dr. Brita Pukstad, who was responsible for the academic content in the Norwegian version of the game.

The game in question takes place in a virtual hospital, where the player can move between different rooms. In each room, a virtual physician presents the character with a multiple choice question (MCQ) in dermatovenereology. Whenever the student responds to the MCQ, he or she is presented with an explanation, detailing why the selected alternative is wrong or correct. A line of more general feedback is also given, positive or negative, depending on the alternative selected. When the student has answered a question and received the appropriate feedback, he or

she is able to move into another patient room in order to receive a new MCQ. One round of the game concludes when the student has answered 20 MCQs. At the end of each round, a summary is given, detailing the number of correct answers, as well as time elapsed. Depending on these two parameters, the student is given a spot on a ranking list.

A non-game platform was needed as an alternative for the control group. Students in this group were given access to the same multiple choice questions in the form of flashcards, making use of the flashcard platform Anki. Anki is a free and public software that can be downloaded to either PC or Mac. The program is often used to practice solving MCQs. When utilizing Anki for this purpose, a question, typically with three to five answer alternatives, is presented. When the user has made up his or her mind about which alternative is correct, the correct answer can be revealed, and the next question will follow. Each time a question is presented to the student, Anki registers that particular question as answered, so it is easy for the student to keep track of which questions he or she has already answered. In this study, all questions were presented with four answer alternatives.

In addition to our assessment of SYS, we wanted to address the context in which the game was going to be introduced. Students' current learning habits and their perceptions of existing modalities of learning is undoubtedly important in evaluating the need for new modalities such as SGs. At NTNU, the teaching methods for medical students vary between lectures, problem-based learning (PBL), practical skills sessions and clinical practice in the hospital. The university has worked out eight main elements that characterize the medical school at NTNU, thus affecting choice of teaching methods. These elements are: i) Early patient contact, ii) use of PBL, iii) focus on clinical relevance in the understanding of basic science subjects, iv) spiral teaching, v) emphasizing the humane aspects of medicine, vi) focus on environmental medicine, vii) focus on clinical practice, and viii) requirement of writing a thesis on a limited subject (which this article is the result of) (NTNU 2016).

Although not explicitly stated by The Faculty, students are expected to study independently in addition to participating in organized teaching activities. There are few guidelines given from the university, and no readings required. There are also very few assignments required to be completed by the students. As is pointed out by Wass et al., assessment of medical students should include a formative aspect - "students should learn from tests and receive feedback on which to build their knowledge and skills." (Wass et al. 2001) In the current course structure, with formal assessment only in the form of an exam at the end of the year, there is a distinct lack of formative assessment compared to summative assessment. Although not to be considered an assessment per se, a Serious Game is a potential platform through which the students could receive feedback on their performance.

We wanted to establish which modalities the students use for self-study, and to which degree students think these modalities are relevant for exams, how motivating they are and how trustworthy they are believed to be. We identified seven learning modalities that are used to some degree at NTNU: #1 Reviewing lectures. #2 Reading medical textbooks. #3 Using websites that offer official and controlled information, such as *UpToDate.com* and *legehandboka.no* (Norwegian Electronic Handbook for Physicians). #4 Using websites that have limited verification of their information, such as *youtube.com* and *Wikipedia.org*. #5 Solving MCQs from earlier exams by printing those exams and solving them on paper. #6 Solving MCQs through the aforementioned software Anki, where students have manually added MCQs from earlier exams and sorted them by year and field (e.g. dermatovenereology). #7 Using serious games that are made for learning, such as the case app *Prognosis*. We chose the first six learning modalities because we know they are used by medical students at NTNU. The seventh modality, Serious Games, was included because it pertains directly to our study. As the specific flashcard platform Anki was utilized as a control in our RCT, we chose to subdivide MCQ-solving into #5 and #6. (NTNU 2016)

2. Methods

All Norwegian fourth year medical students at NTNU during the spring semester of 2016 were invited to participate in our study. Foreign exchange students were excluded from our study because they did not participate in the dermatovenereology curriculum. The field of dermatovenereology is one of six main subjects taught during this school year. Half of the students have dermatovenereology classes during the fall semester, and the other half during the spring semester. However, all students have their exams at the end of this school year. The students' learning habits were investigated via an electronic checklist questionnaire. More specifically, students were asked to evaluate each of the aforementioned learning modalities with regards to 'time spent', 'motivation', 'trust in information obtained' and 'relevance to exams'. The questions and rating scales are listed in Table 1.

Fig. 1: Trial overview

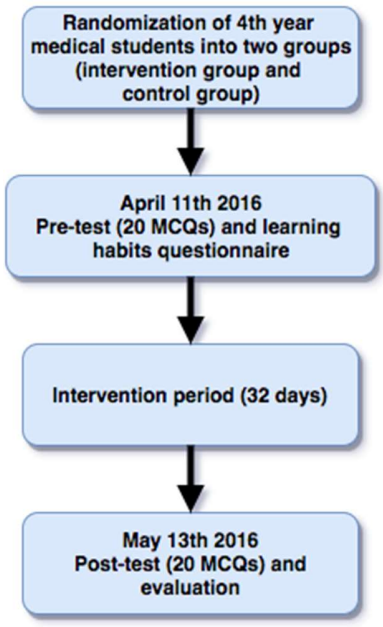


Table 1. Questions used for surveying students on their study habits

Question	Answer alternatives	Num**
How much time do you spend on the following modalities* per week?	Less than 1 hour	0
	1-3 hours	1
	4-6 hours	2
	7-9 hours	3
	More than 9 hours	4
To what degree do you find the use of the following modalities* to be motivating?	Do not use	0
	To a small degree	1
To what degree do you trust information accessed through the following modalities*?	To some degree	2
	To a large degree	3
To what degree do you find the use of the following modalities* to be relevant to exams?	To a very large degree	4

*Lecture notes, Textbooks, MCQs from previous exams accessed directly from The Faculty, MCQs from previous exams in Anki, Quality assured online resources (such as UpToDate), Other online resources (such as YouTube, Wikipedia etc.), Serious Games in the form of apps or computer games (such as Prognosis, Microbe Invader etc.) **Numeric representation of answers used for analysis in SPSS

The evaluation of SYS was conducted as a randomized controlled trial with a pre-test - intervention - post-test setup. Firstly, eligible students were randomized into one of two groups. As half the students were currently in the dermatovenereology semester, and half had already completed this semester, students were stratified by semester before randomization. A pre-test with 20 multiple choice questions in dermatovenereology was then conducted. At the pre-test, each student was given a random, unique numeric ID in range 1-100 to be used at both the pre-test and the post-test. We did not connect names to ID-numbers for purposes of anonymity. After the pre-test we decided by random number generation which group would be designated as the intervention group (referred to hereafter as the game group), and which group that was to be the control group (referred to hereafter as the flashcard group). Students in the game group were given access to a website hosting SYS. For the sake of the intervention, the game itself contained 5 modules of 20 unique MCQs, in total 100. In addition to the game itself and the ranking list, the SYS website contained a collection of reading materials organized by subfields within dermatovenereology. Students were encouraged to make use of the game during a period of 32 days.

For the flashcard group, we electronically distributed a download link to Anki, and the file containing the MCQs to be opened in Anki. The 100 MCQs were equivalent to those used in SYS, and were organized into five 'decks' of 20 MCQs, corresponding to the five modules of SYS. In contrast to how Anki has traditionally been used at NTNU, where only the correct answer has been revealed, students were also shown explanations detailing why each alternative was correct or wrong upon completing a question. This was also in contrast to SYS, in which the student only received an explanation of the one alternative that he or she selected, explaining why this answer was wrong or correct. The students in the flashcard group were encouraged to make use of the flashcards during the same 32 day interval as the game group.

After the intervention period we conducted a post-test using 20 new MCQs in dermatovenereology. The students were asked to use the unique ID that had been assigned to them during the pre-test. At this time, participants were also asked to evaluate the intervention through a questionnaire where they would rate their agreement with statements about the game/flashcards. Statements were designed to assess students' overall opinions of the intervention, including if they found it motivating, relevant to exams, and whether they would recommend it to other students. A detailed summary of statements and rating scales are listed in Table 2. In addition, students were invited to give written feedback on the intervention in a free text box.

Table 2. Statements about the game/flash cards presented at the post-intervention evaluation.

Statements	Answer alternatives (same for all questions)
The [practice questions/game] helped me understand content.	1 = completely disagree 2 3 = neither agree nor disagree 4 5 = completely agree
I enjoyed learning via the [practice questions/ game].	
I would recommend the [practice questions/game] to other students.	
I prefer the [practice questions/ game] to traditional teaching methods.	
The [practice questions were/game was] a good use of my time.	
The [practice questions/game] helped me retain information.	
The [practice questions/game] helped me with application to clinical practice.	
The [practice questions/game] was a good way to prepare for exams.	

Data analysis was conducted in SPSS using imported data from electronic questionnaires. For the purpose of analysis, scale variables in plain text were converted to numeric values. An issue that arose was the fact that participants had the option of selecting 'do not use' with regards to questions about different learning modalities. Because the option 'do not use' doesn't scale numerically with other options, 'do not use' was regarded as a missing value for quantitative statistics. Specifically, quantitative statistics regarding 'motivation for', 'trust in' and 'exam relevance of' different modalities is presented with respect to only those participants who chose another option besides 'do not use'. Thus, the numeric data presented regarding these parameters are interpreted as scores among those students who actually use the specified modalities.

A number of students had not kept their unique ID to input during the post-test. Additionally, some students only participated in the pre-test, and some only participated in the post-test. All of these cases were excluded for analysis of test scores. That is, when calculating mean/median scores of the pre-test and post-test as well as when calculating mean/median difference in score between post-test and pre-test we only included cases which had the same unique ID for both pre-test and post-test. This led to reduced sample sizes, but it was the only way to ensure that we were actually measuring pre/post-test scores for the same person. In addition, students who reported at the post-intervention evaluation that they had not used the intervention they were assigned to, were also excluded from the analysis.

Although instructed to check only one alternative in the digital questionnaires, some students checked more than one alternative to certain questions. Individual answers containing more than one alternative were regarded as missing values for the purpose of analysis. Because of relatively low *n* and relatively large skew in the data, we decided to utilize non-parametric tests instead of T-tests to determine the significance of potential differences between groups. T-tests were conducted during the process for comparison, and generally agreed well with nonparametric tests with regard to significance level.

3. Results

3.1. Compliance

91 students were invited to participate in the study. 74 students attended the pre-test and study habit questionnaire, and 73 students attended the post-test and post-intervention evaluation. 68 of the 73 students (93.2%) reported at the evaluation that they attended the pre-test, meaning that 5 students who attended the post-test did not attend the pre-test. Only 2 of 73 students (2.7%) reported at the evaluation that they did not use their respective intervention, that is, they did not use Anki or SYS at all. These two students were in the flashcard group. The rest, 71 of 73 (97.3%), used Anki or SYS to at least some degree. 56 out of 73 (75.7%) reported that they had completed all 100 MCQs that were included in SYS/Anki. 13 students (17.3%), 6 in the flashcard group, and 7 in the game group, had not completed all 100 MCQs. 4 students (5.4%) were unsure if they had completed all MCQs. All of those unsure were in the game group.

3.2. Study habits

Time spent by students on different study modalities is shown in Table 3. Notably, students spend substantially more time reviewing lectures than any other learning modality that was queried about. Cumulatively, 54/74 (73%) of students report either using 7-9 hours per week or more than 9 hours per week reviewing lectures. We also found that students spend substantially less time using serious games than other modalities, with 70/74 (94.6%) of students reporting using serious games less than 1 hour per week.

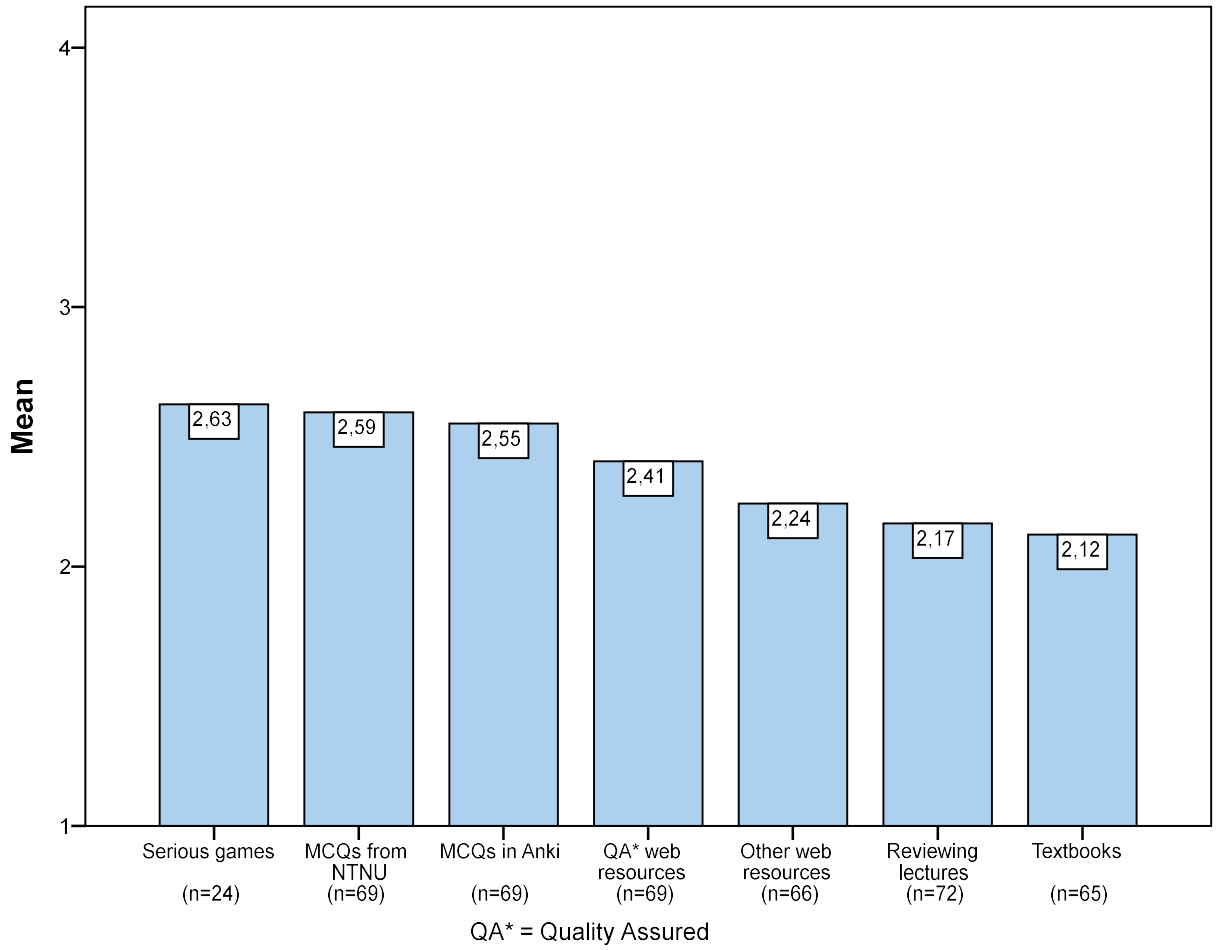
Table 3. Time spent on different learning modalities per week

Modality	Median time spent	Median score*	Mean score*	N
Reviewing lectures	7-9 hours	3	3.09	74
QA* web resources	1-3 hours	1	1.16	74
Other web resources	1-3 hours	1	0.85	73
Textbooks	Less than one hour	0	0.77	74
MCQs in Anki	1-3 hours	1	0.68	74
MCQs from NTNU	Less than one hour	0	0.55	71
Serious Games	Less than one hour	0	0.07	74

*0 = Less than one hour, 1 = 1-3 hours, 2 = 4-6 hours, 3 = 7-9 hours, 4 = more than 9 hours

Figure 2 shows the degree to which students feel motivated by different learning modalities. Numbers presented represent mean scores on a scale of 1-4 (see Table 1). There appeared to be only minor differences between modalities, with mean scores for modalities ranging from 2.12 (textbooks) to 2.63 (games). Students also had the option to select "do not use" (0), and only students who did not choose this option were included in the analysis.

Fig. 2: Motivation for different learning modalities (mean)



The degree to which students trust information gained through different modalities is shown in Figure 3. Numbers presented represent mean scores on a scale of 1-4 (See Table 1). Compared to all other modalities, students place substantially more trust in information accessed through quality assured online resources such as the Norwegian Electronic Handbook for Physicians (NEL) and UpToDate (mean score 3.72). Compared to all other modalities, students place less trust in information accessed through serious games (mean score 2.09), non quality assured online sources such as YouTube and Wikipedia (mean score 2.28) and in previous exam questions collected and organized by students (mean score 2.36). Students also had the option to select "do not use" (0), and only students who did not choose this option were included in the analysis.

Fig. 3: Trust in information accessed through different learning modalities (mean)

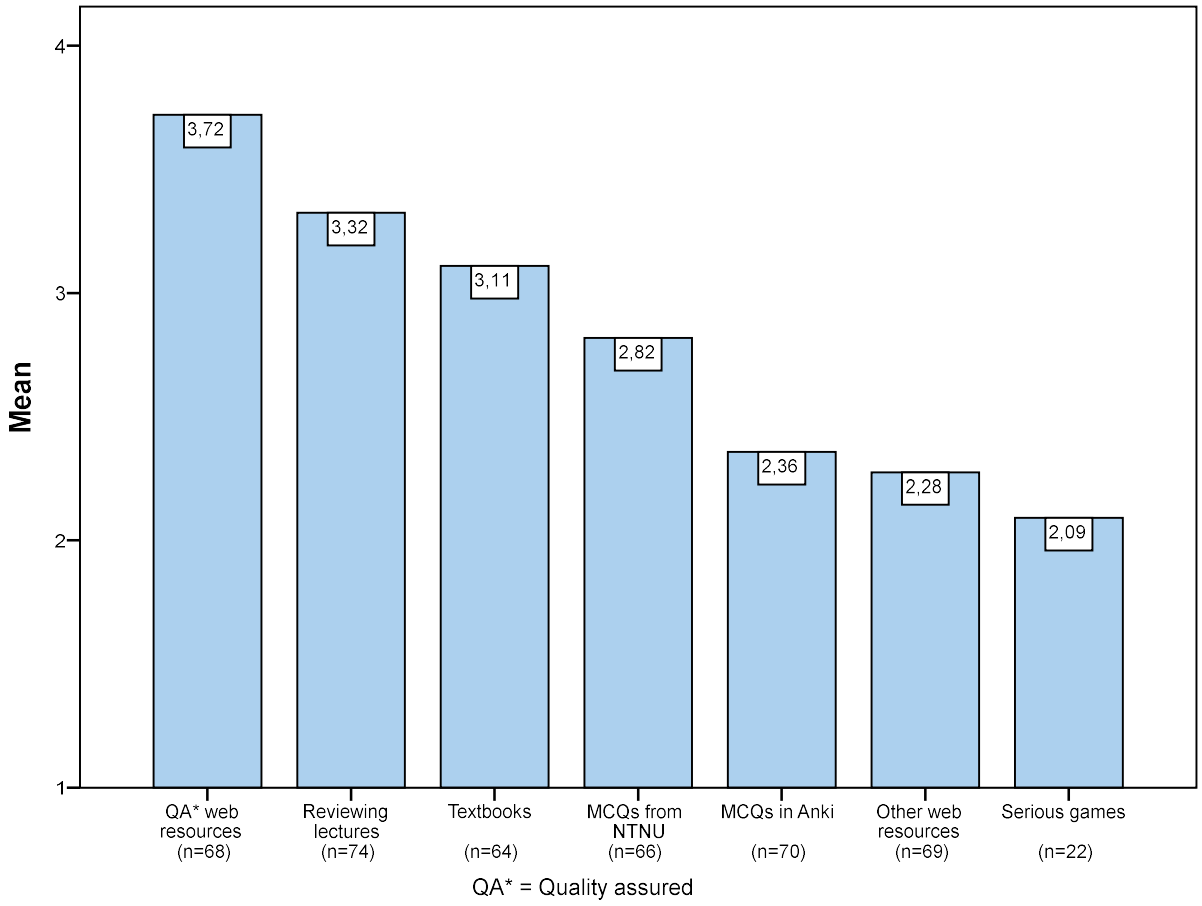
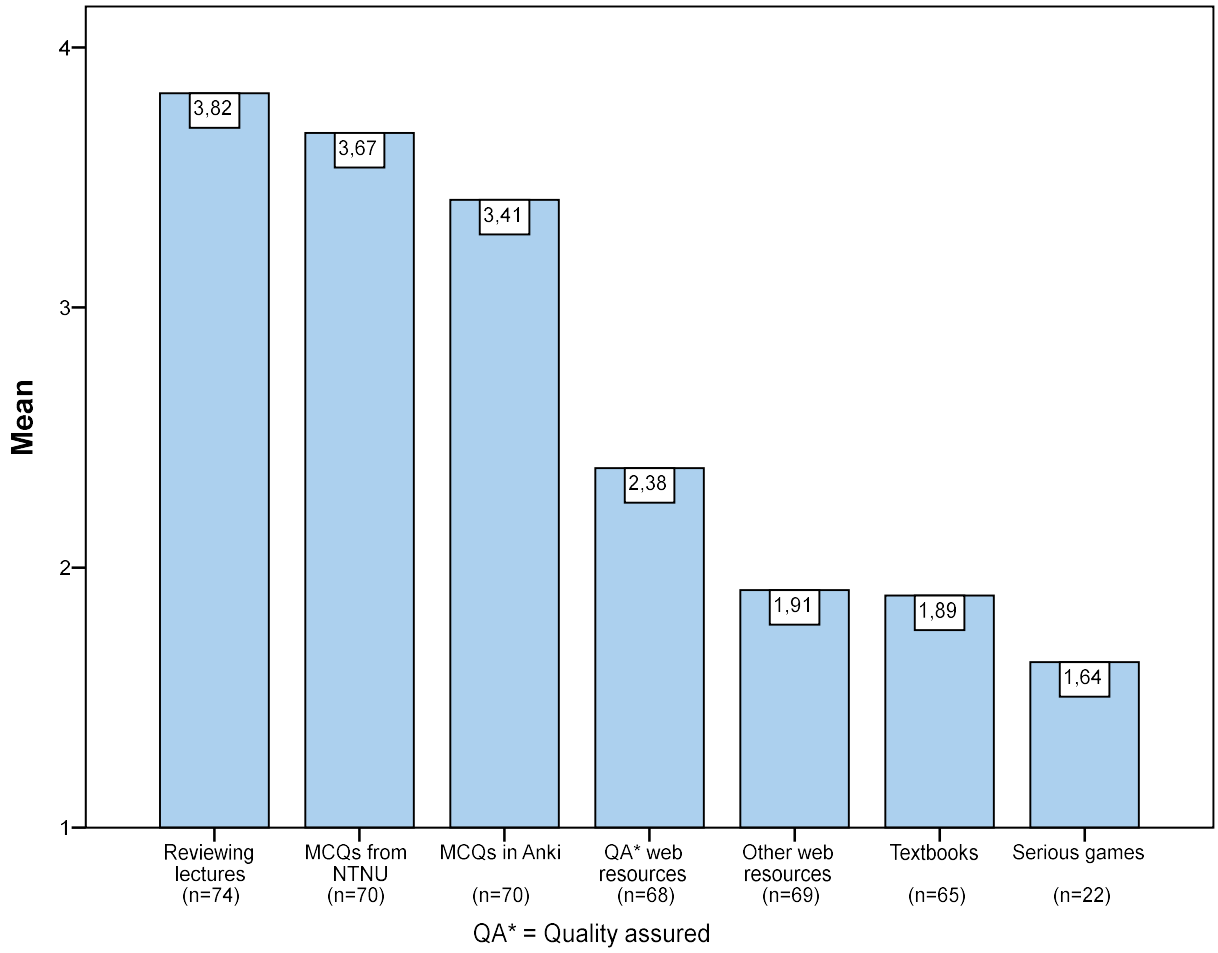


Figure 4 shows to what degree students perceive different learning modalities to be relevant to their exams. Numbers presented represent mean scores on a scale of 1-4 (See Table 1). Students rate reviewing lectures (mean score 3.82), reviewing previous exam questions obtained from the university (mean score 3.67) and reviewing previous exam questions in Anki (mean score 3.42) higher than all other modalities (range of means 1.64-2.38). Students also had the option to select "do not use" (0), and only students who did not choose this option were included in the analysis.

Fig. 4: Perceived relevance to exams for different learning modalities (mean)



3.3. Knowledge aquisition

Both mean and median score at the pre-intervention knowledge test were different between groups (flashcard group mean 12.54, median 12; game group mean 13.42, median 14). However, a Mann Whitney U test for distribution of scores showed no significant difference between groups (p=0.228). The mean score at the post-intervention knowledge test was very similar between the groups (flashcard group mean 18.08, median 18; game group mean 18.27, median 19). A Mann-Whitney test showed no significant difference in distribution of scores between groups (p=0.299). Median difference between post-test and pre-test scores was significantly different from 0 for both groups, signifying improvement from pre-test to post-test for both groups (p<0.001 for both groups). Median improvement in the flashcard group was 6.0, and median improvement in the game group was 5.0. An independent-samples median test revealed the difference in median improvement between groups to be non-significant (p=0.404). A Mann Whitney test supported the non-significance of difference in improvement between groups (p=0.307). A summary of scores stratified by group is shown in Table 4.

Table 4. Pre-score, post-score and absolute difference between post-score and pre-score. Stratified by group.

	Group	Mean	Std. Deviation	Min; Median; Max	N
Pre-test	Game group	13.42	2.788	7; 14; 18	26
	Flashcard group	12.54	3.101	8; 12; 19	26
Post-test	Game group	18.27	1.343	14; 19; 20	26
	Flashcard group	18.08	1.129	16; 18; 20	26
Post-pre absolute difference	Game group	4.85	2.679	0; 5; 10	26
	Flashcard group	5.54	2.789	-1; 6; 9	26

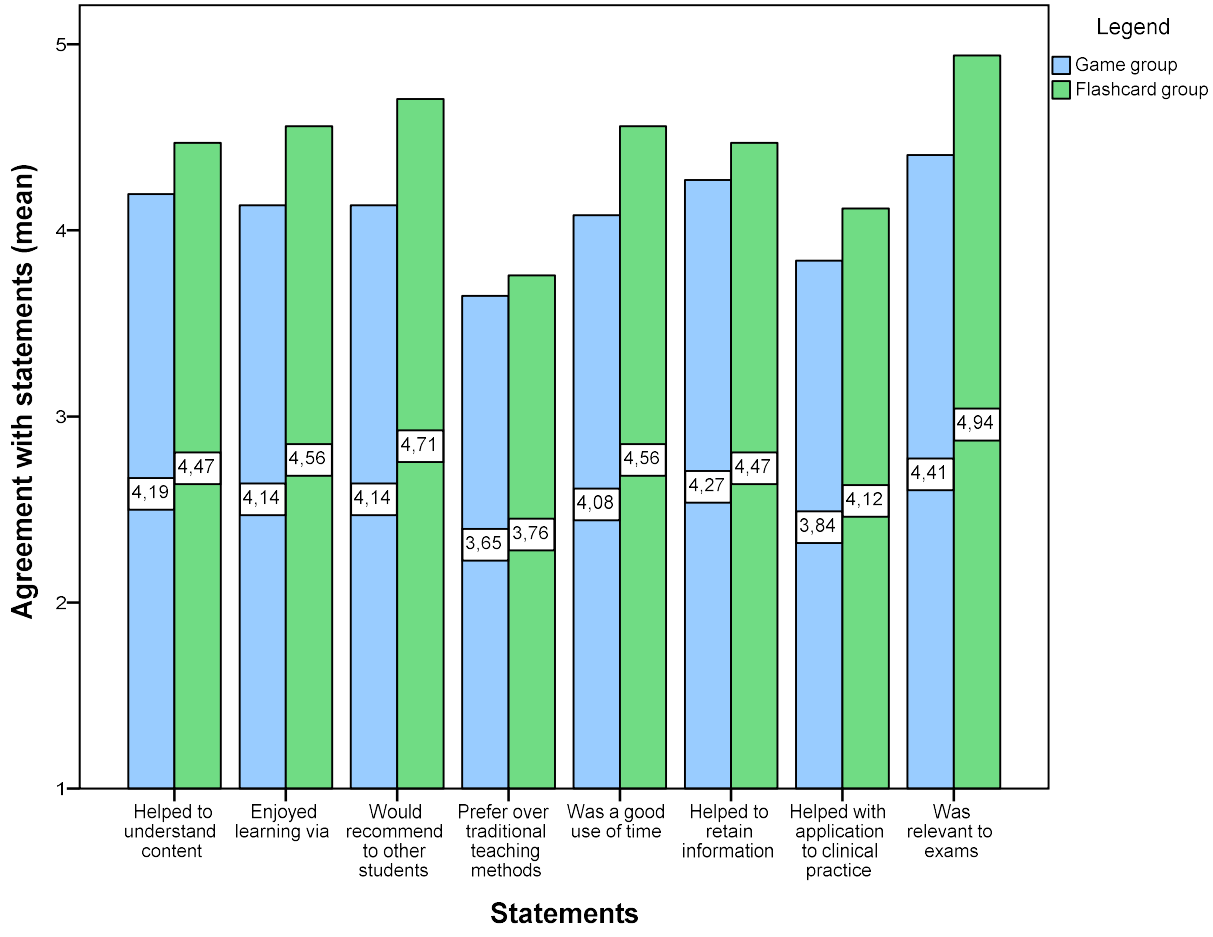
Only cases with the same unique ID for pre-test and post-test were included in the analysis. Those who reported at the post-test that they hadn't used the intervention they were assigned to were excluded from the analysis.

3.4. Post-intervention evaluation

3.4.1. Agreement with statements

When looking at the post-intervention evaluation from students, the trend seemed to favor the flashcard group over the serious game group (Figure 5). A Mann-Whitney U Test for equality of distribution across groups, showed that the flashcards scored significantly higher than the game for ‘relevance to exams’ ($p<0.001$), ‘recommendation to other students’ ($p=0.031$) and ‘good use of time’ ($p=0.022$). Students who reported at the post-intervention evaluation that they had not used the intervention they were assigned to, were excluded from the analysis.

Fig. 5: Agreement with statements about the game/flashcards

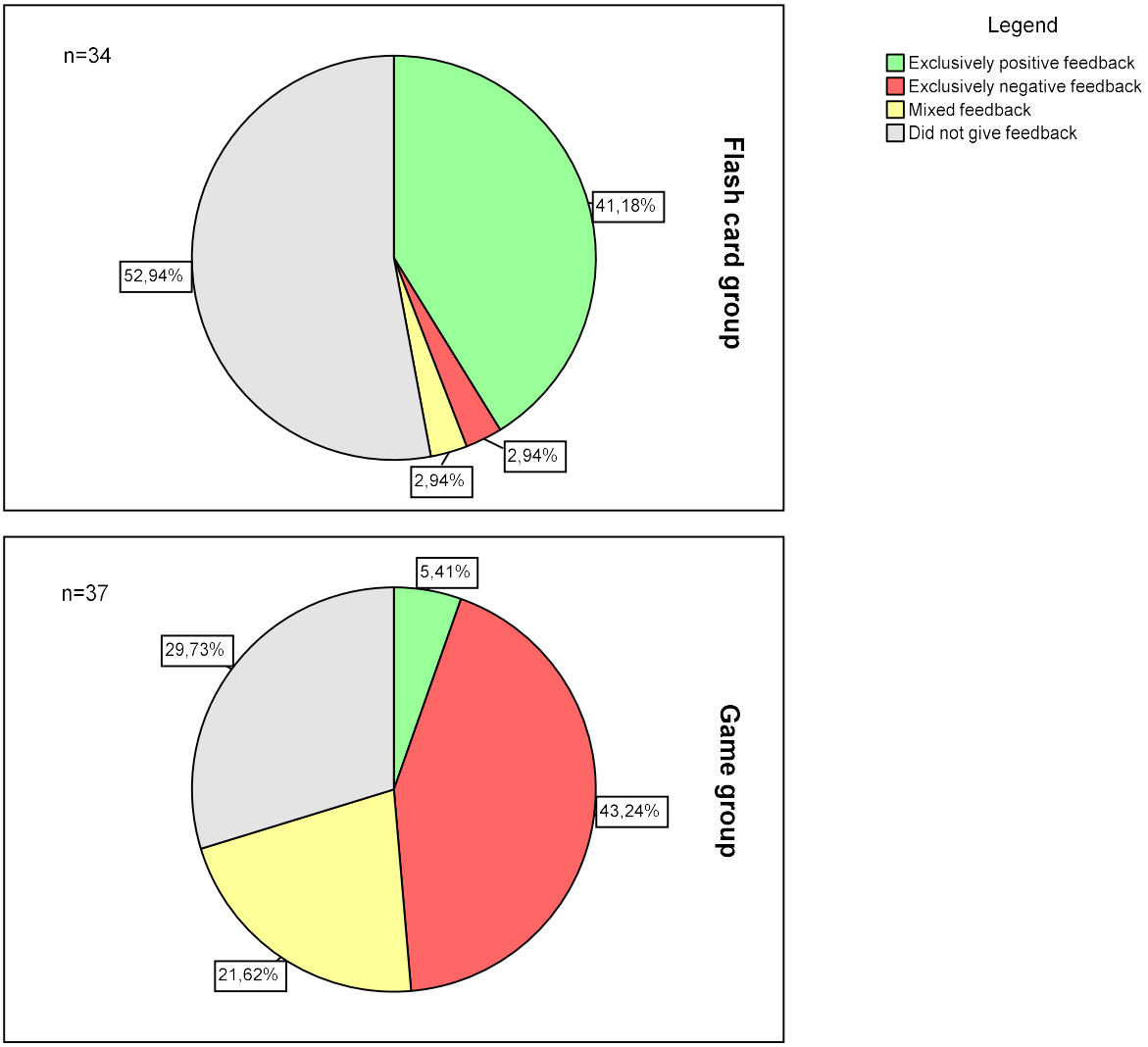


Regardless of group, students tended to prefer both the flashcards and SYS over traditional teaching methods. Out of the 70 students who answered, 41 (58.6%) agreed to some degree with the statement “I prefer the [practice questions/game] over traditional teaching methods”. Only 6 (8.6%) students disagreed in some degree with the same statement. The remaining 23 (32.9%) neither agreed nor disagreed with the statement. Because of rounding, percentages do not add up to exactly 100.0%. The students who reported at the post-intervention evaluation that they had not used the intervention they were assigned to, were excluded from the analysis.

3.4.2. Subjective feedback

42 of 71(59.2%) students who reported making use of their respective intervention gave a free text feedback. 26 of 37 (70.2%) of the students in the game group gave feedback, compared to 16 of 34 (47.1%) in the flashcard group. We grouped the feedback into exclusively positive, mixed positive and negative, and exclusively negative. Note that all types of criticism were regarded as negative feedback. In the flashcard group, of those who gave feedback, 14 of 16 (87.5%) gave an exclusively positive feedback, 1 of 16 (6.3%) gave a mixed feedback, and 1 of 16 (6.3%) gave an exclusively negative feedback. Of those in the game group who gave feedback, 2 of 26 (7.7%) gave an exclusively positive feedback, 8 of 26 (30.8%) gave a mixed feedback, while 16 of 26 (61.5%) gave an exclusively negative feedback. Figure 6 shows the distribution of free text feedback by group, including those who did not give any feedback.

Fig. 6: Distribution of free text subjective feedback on the intervention



When analyzing the free text feedback, there were some recurring themes and issues. Examining the feedback from the flashcard group, 7 of 16 (43.8%) of the free text feedback contained praise towards the quality of questions. 12 of 16 (75.0%) of the free text feedback included reports saying that they liked the feedback they got in Anki. They particularly mentioned the immediate feedback when selecting their answer, and being able to see the explanation of all the alternatives, detailing why they were right or wrong. The small amount of negative and mixed feedback was scattered. Looking at the feedback from the game group, 8 of 25 (32.0%) of the free text feedback included some kind of criticism of the technical performance of the game, saying the game was running slowly or demanding a lot of computing power. Other recurring themes were criticism of not getting the correct answer immediately after choosing an answer, and lack of fully exploiting the opportunity of the game platform.

4. Discussion

4.1. Study habits

It is noteworthy that students report spending substantially more time reviewing lectures than time spent on any other learning modality. Furthermore, reviewing lectures also scored highly on trust in information obtained, and on relevance to exams. The point in the latter sentence may in part explain the former, in that reviewing lectures seems to be perceived as a good all-around approach to learning the material, and so it makes sense that students spend more time on this modality. In addition, all lectures are accessible through the university's intranet, making the academic content from lectures easily and freely available. The finding that students generally rate lectures highly compared to other modalities is interesting. Especially in the context of ongoing changes at the faculty of medicine in Trondheim, where there is currently an initiative to reduce the number of lectures in favor of self-study and other learning modalities (NTNU 2014).

Notice, however, that reviewing lectures did not score particularly well on "motivation", although the differences between modalities were quite small for this metric. The mismatch between motivation and time spent on lectures is interesting to note, in that motivation is likely not the explanation for the widespread use of reviewing lectures. It might be that the apparent superiority of reviewing lectures is related to the perceived importance of lectures in the current course structure. At NTNU, the lecturers are responsible for making MCQs for exams, and it is not unreasonable for students to assume that the content of lectures will, to some degree, predict the content of future exams. It could be that time spent reviewing lectures reflects a feeling of necessity rather than motivation and/or quality of lectures.

If one accepts perceived relevance to exams as one of the most important predictors of time spent, one can question why the time spent reviewing lectures is so much greater than time spent solving MCQs from previous exams, which also scored highly for perceived relevance to exams. It could be that it is merely the volume of lectures which results in substantially more time spent on lectures than on MCQs from previous exams, even though these modalities showed similar perceived relevance to exams.

It is difficult to predict the results of changing the course structure by reducing the amount of lectures. The question remains whether other modalities can replace the removed lectures in an efficient way. It should also be noted that we only assessed students' opinions on self-study learning modalities. We did not investigate students' opinions on group learning activities such as problem-based learning (PBL) (Wood 2003), team based learning (TBL) (Michaelsen and Sweet 2008) and clinical rounds (Spencer 2003). These, and other group learning activities, might be more realistic alternatives for The Faculty when replacing lectures.

4.2. Knowledge aquisition

Previous reviews and one meta-analysis show mixed results regarding difference in objective learning outcomes between a group using a serious game and a control group (Boyle et al. 2016, Connolly et al. 2012, Wouters et al. 2013). We were unable to find such a difference. However, it should be emphasized that there were several limitations and methodological flaws in the RCT-aspect of our study with regards to contributing something meaningful to this particular issue:

1) The sample size turned out to be somewhat smaller than anticipated. 91 students were invited to participate in our study. Of these, 74 participated in the pre-test. Of these, some students didn't participate in the post-test, some had forgotten their numeric ID for use at the post-test, and some reported that they hadn't used their respective intervention. In the end, we were left with 52 cases to be included in the analysis of pre-test and post-test scores, 26 in each group. Although not a negligible amount, the sample size was thus less than optimal. In comparison, in the study by Sward et al., 81 participants completed the pretest, posttest and questionnaire (Sward et al. 2008).

2) In choosing to assign students to groups by simple randomization (with stratification for semester), we assumed the student population to be homogenous, and *n* to be sufficiently large to yield comparable results between groups at the pre-test. As the mean pre-test score differed between groups, it is safe to conclude that one or both of these assumptions did not hold. This led to problems in comparing the quantitative improvement between groups.

3) For both groups the mean post-test result was very close to the maximum possible result. It is therefore difficult to establish whether the equality in post-test means across groups reflects on the quality of the interventions, or if it reflects that the post-test was too easy compared to the knowledge level of the students, leading to a saturation phenomenon. The timing of the post-test was unfortunate, in that it was held immediately following a repetition seminar in dermatovenereology. This was a question of logistics, but in hindsight it is quite likely that this could have affected the post-test scores, contributing to such a saturation phenomenon.

Due to all these factors, caution should be exercised in utilizing our data for strengthening the hypothesis that there are no differences in learning outcomes between students using a SG and students using a control intervention.

4.3. Post-intervention evaluation

In the post-intervention evaluation, the students did not rate the serious game higher than computerized flashcards. This is in contrast to the study by Sward et al., where they showed that the participants favored a serious game over computerized flashcards for several metrics (Sward et al. 2008). The study by Sward et al. is a natural choice to compare our results with, as we used a very similar questionnaire to assess the students' opinions on a game and the flashcards. In addition, the participants in our RCT are comparable to Sward et al.'s participants, as the subjects were medical students in both cases.

The discrepancy could potentially be explained by a small sample size giving insufficient power. However, students actually tended to prefer the control intervention on all our metrics. It is also quite possible that the Serious Game used in our study was inferior in one or more aspects, compared to serious games used in other studies. This is supported by students' written feedback, wherein several students addressed both technical and pedagogical issues with the game used. It might be the case that insufficient care was taken to implement elements thought to promote student involvement within an instructional gaming environment, such as those described by Malone (Malone 1980) and Prensky (Prensky 2001), among others. Trying to adapt the already established game layout of SYS to the academic content left little room for consideration of such factors, and the game may have appeared as a "chocolate covered broccoli" (Green 2014).

It is also possible that the chosen intervention for the control group was of particularly high quality. There is some support for this in that the free text feedback for the control intervention was almost exclusively positive. The fact that many students were already familiar with the format of the control intervention, whereas very few students reported using serious games previously could also have had an impact on the outcome. Furthermore, the Anki version used by students for many years at NTNU has been simpler than the one presented in this study. Earlier, Anki was a for-students-by-students project, based upon MCQs given in earlier exams. No one from the faculty has been involved in adding or controlling the questions, and this lack of content quality control made the earlier version less trustworthy. In addition, the old version did not have explanations included in the answers. In conclusion, one can say that the Anki version used in this study was substantially improved compared to the version the students were familiar with.

The finding that students, regardless of group, tended to prefer the intervention over traditional learning modalities is interesting to note. Given this finding, it might seem paradoxical that neither Anki nor Serious games scored particularly favorable on preference metrics before the intervention. One possible conclusion is that the content presented in both interventions was of such quality that although students normally don't prefer these modalities in learning, the way that they were presented made them favorable to traditional modalities. Although our data do not

support the efficacy of the game over Anki, it could then be argued that utilizing either Anki or the game in teaching is something that students would be in favor of.

4.4. Choice of control group for studying Serious Games

Fundamentally, one could question whether a flashcard-interface such as Anki is an appropriate control intervention. The distinction between a computer based learning platform and a serious game is potentially not so clear. Drawing on Lindley (Lindley 2003) and Prensky (Prensky 2001), Wouters et al. take the following characteristics to define a game: that it is *goal-directed*, *competitive,* conducted within a framework of *agreed rules,* and that it constantly provides *feedback* to enable players to monitor their progress towards the goal (Wouters et al. 2009). Using this definition, it could be argued that Anki, in the form that we used it in our study, approaches the realm of serious games. On the one hand, Anki does not actively utilize feedback. You cannot actively select your answer in the program; you merely choose to continue when you want the answers and explanation to be shown. Accordingly, Anki cannot give you a positive/negative feedback based upon whether the answer you selected was right or wrong. This also means it cannot keep track of your score, so unless the user actively keeps score, there is no obvious competitive aspect. On the other hand, the user does give Anki feedback on the difficulty of the question, which in turn decides the time interval until the question is repeated. Anki also keeps track of cards studied, and the user may set a goal of cards studies per day. In conclusion, although one might not define flashcard-interfaces and similar educational tools as games, this is an aspect that might deserve consideration when choosing a control for a SG.

In general, choosing a control for SGs is not a simple task. One might want to find a control that is something along the lines of “business as usual” or “best current practice”. But neither of these alternatives are easy to define in medical education. On the one hand, one might want to have a control that employs some digital medium, so as to isolate the “game” aspect of the SG. On the other hand, this might not be representative of current practice, and one encounters the problem of having to draw the line between digital learning modalities and digital games. Another aspect is whether the content available to the student through the game is the same as that available through the control. In other words, if one finds a difference between a game group and a control group, is it the “gameiness” or the quantity and quality of content that separates the two? If one tries to solve this problem by making sure the content is as equal as possible between the two interventions, problems may occur. On one side, one risks limiting the platform of the game, and on the other side one risks inflating the control to something which does not reflect the control in the way that it is currently used in education. If the problem is ignored altogether, one will have a hard time in measuring the actual effect of SGs. It should be pointed out, however, that such challenges could potentially be overcome by clever game design. One such example is an RCT by Papastergiou on the impact of using a SG for high scool students learning computer science. The game in question adopted several elements thought to promote student involvement within

an instructional gaming environment, based on the work of Malone (Malone 1980) and Prensky (Prensky 2001). At the same time, great care was taken to ensure that the content presented to the control group, through an educational website, was equal to the content in the game. Students in the game group performed significantly better on knowledge acquisition in a pre-test - intervention - post-test setup ($p = 0.004$), and the game scored significantly better for overall appeal ($p = 0.001$) and educational value ($p = 0.002$) at post-intervention evaluation (Papastergiou 2009).

4.5. Conflicts of interest statement

The primary authors of this study, Schaanning and Aksdal, were also involved in the preparation of both the game intervention and the control intervention in this study. When the first version of the updated game was presented to us, we felt the game was lacking in using the possibilities a game platform can offer. We presented certain suggestions which were implemented in the game. We also helped with some of the technical difficulties that emerged when translating the game to Norwegian. In preparation for the intervention, our role was to insert the questions made by Dr. Pukstad into Anki and SYS. It should be noted that our contribution to the development of the game was quite limited, as the technical framework of the game was already established at the time when we were asked for input. However, we feel that this is important to mention, so the reader can question our impartiality and independence in this study.

5. Conclusion

Our data shows that students are generally very favorable toward reviewing lectures for learning the curriculum. Students spend a lot of time on this, they feel that the information can be trusted, and they believe that it is relevant to their exams. Despite this, it is unclear whether this reflects the quality of lectures, and based on our data, it is difficult to give clear recommendations about the future role of lectures in the medical curriculum.

We have shown that the serious game SYS was not better than flashcards for learning dermatovenereology. The students did not prefer SYS to the flashcards, and in fact tended to prefer the flashcards over SYS. In light of our data, and the fact that development of computer games is significantly more expensive than creating flashcards, the usage of serious games similar in scope and content to SYS does not appear to be beneficial or cost effective in learning dermatovenereology. If The Faculty of Medicine still wants to focus on the gaming platform, we suggest that they should look to already finished games that are of high quality and complimentary to the content The Faculty wants to include in the game. Games should be developed with special consideration of elements thought to stimulate learning in a digital game environment. Such games could also be developed by The Faculty, but this might be substantially more demanding in terms of resources.

In this regard, SYS might not have been the best choice, since it was lacking in technical and pedagogical features. In general, the prospect of creating a game to be used with generic MCQs is very difficult. In order to retain the qualities of video games in an educational setting, it is likely necessary for the content to be tailored to the game or vice versa, to avoid ending up with a "chocolate covered broccoli". For The Faculty, we recommend that resources should be spent on other simpler platforms, and that the focus should be on quality control of the academic content, rather than trying to make learning fun. Given the almost exclusively positive feedback on our flashcard setup and the general ease of constructing such a setup, we would recommend such an approach over serious games in learning dermatovenereology, as an addition to traditional learning modalities.

Regarding the future of SGs, we cannot provide much hard evidence on their efficacy. SYS is probably not a SG worth pursuing for use together with MCQs. However, SYS is not necessarily representative for the platform of SGs used for learning and therefore, SGs as a platform should not be rejected on the basis of this study. A fundamental problem when dealing with the quantitative study of SGs is that different games will necessarily be of vastly varying quality. As more high quality studies concerning SGs are released, it might be constructive for researchers and developers to capitalize on the games whose efficacy can be shown to be superior to traditional teaching methods. By doing so, one might be able to identify factors which characterize successful SGs, and focus on such factors in the development of future SGs.

6. Bibliography

Boyle EA, Hainey T, Connolly TM, Gray G, Earp J, Ott M, Lim T, Ninaus M, Ribeiro C, Pereira J. 2016. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers & Education. 3//;94:178-192.

Dermatology Online With Interactive Technology. Available from http://www.cyberderm.net/en/home.html

Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM. 2012. A systematic literature review of empirical evidence on computer games and serious games. Computers & Education. 9//;59:661-686.

Djaouti D, Alvarez J, Jessel J-P, Rampnoux O. 2011. Origins of Serious Games. In: Serious Games and Edutainment Applications. 1 ed.: Springer-Verlag London. p. XVI, 504.

Green CS. 2014. The Perceptual and Cognitive Effects of Action Video Game Experience. In: Learning by Playing: Video Gaming in Education. Oxford University Press. p. 384.

Knight JF, Carley S, Tregunna B, Jarvis S, Smithies R, de Freitas S, Dunwell I, Mackway-Jones K. 2010. Serious gaming technology in major incident triage training: a pragmatic controlled trial. Resuscitation. Sep;81:1175-1179. Epub 2010/08/25.

Game Taxonomies: A High Level Framework for Game Analysis and Design. Available from http://www.gamasutra.com/features/20031003/lindley_01.shtml

Malone TW. 1980. What makes things fun to learn? heuristics for designing instructional computer games. Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems.

Michael D, Chen SL. 2005. Serious Games: Games That Educate, Train, and Inform: Muska \& Lipman/Premier-Trade. 1051239.

Michaelsen LK, Sweet M. 2008. The essential elements of team-based learning. New Directions for Teaching and Learning.2008:7-27.

En fornyet og fremtidsrettet
legeutdanning ved NTNU. Available from
http://legeforeningen.no/PageFiles/191054/H%C3%B8ringsnotat -Rapport en fornyet og fremtidsrettet legeutdanning ved NTNU.pdf

Studiets oppbygging. Available from http://www.ntnu.no/studier/cmed/oppbygging

Papastergiou M. 2009. Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. Computers & Education. 1//;52:1-12.

Prensky M. 2001. Digital Game-Based Learning New York: McGraw-Hill.

Spencer J. 2003. Learning and teaching in the clinical environment. British Medical Journal; International edition. Mar 15;326:591-594.

Sward KA, Richardson S, Kendrick J, Maloney C. 2008. Use of a Web-based game to teach pediatric content to medical students. Ambul Pediatr. Nov-Dec;8:354-359. Epub 2008/12/17.

Wass V, Van der Vleuten C, Shatzer J, Jones R. 2001. Assessment of clinical competence. Lancet. Mar 24;357:945-949. Epub 2001/04/06.

Wood DF. 2003. Problem based learning. BMJ.326:328-330.

Wouters P, van der Spek ED, Oostendorp Hv. 2009. Current Practices in Serious Game Research: A Review from a Learning Outcomes Perspective.: IGI Global.

Wouters P, van Nimwegen C, van Oostendorp H, van der Spek ED. 2013. A meta-analysis of the cognitive and motivational effects of serious games. J Educ Psychol.105:249-265.