Media Engineering and Technology Faculty
German University in Cairo

A-Learning: Adaptive Educational Games Platform

Bachelor Thesis

Author: Mariz Samir Mounir Awad
Supervisors: Prof.Slim Abdennadher
Eng.Jailan Mohamed Salah

Submission Date: 26 May, 2019

Media Engineering and Technology Faculty
German University in Cairo

A-Learning: Adaptive Educational Games Platform

Bachelor Thesis

Author: Mariz Samir Mounir Awad
Supervisors: Prof.Slim Abdennadher
Eng.Jailan Mohamed Salah

Submission Date: 26 May, 2019

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

---

Mariz Samir Mounir Awad
26 May, 2019

Acknowledgments

First, I would like to offer my sincerest gratitude to Prof. Slim Abdennadher for his guidance and advice during different stages of development of this research work.

I would also like to dearly thank Dr. Nabila Hamdi for taking the time to design and provide us with the question pool from the Pathology course, and offering to supervise and proctor the experiment with the Pharmacy participants.

My utmost gratitude goes to my first-hand supervisor Eng.Jailan Salah El-Din for her positive reinforcement attitudes, and her ungrudging willingness to devote time and effort to help whenever needed.

I am entirely indebted to my mother for her continuous financial and emotional support. I am truly blessed and can never repay you for your troubles.

I would also like to express special thanks to my good friend Omar Khaled for his impressive work in coordinating and gathering enough Pharmacy students for the experiment, as well as all my dear friends and colleagues who helped in managing it.

Abstract

Computer Adaptive Testing (CAT) methodologies have been widely used by test centres for the quick assessment of examinees, in which the test adapts the difficulty according to the ability level of the examinee. A modification approach, termed CAL, for Computer Adaptive Learning is devised to utilize the CAT principles for the purpose of efficiency in learning rather than assessment. The system is then applied to a platform of serious games with multiple choice questions as a proof of concept.
This thesis, hence, investigates the effectiveness of item-based adaptive difficulty achieved through the CAL, and compares it to another adaptivity feature where the game mechanics and UI adapt to the user's emotional state, however, the difficulty increases sequentially in the traditional manner of passing through levels (easy, medium, and hard).
This comparison is carried out in terms of knowledge gain, engagement level and learning efficiency. For this purpose, two versions of the platform are implemented for the two different adaptivity features and subjected to an experiment in order to reach a conclusion. The experiment is comparative in nature and features two groups: one playing the adaptive difficulty version and one playing the emotional adaptivity version.
Results from the experiment confirm that the proposed CAL algorithm achieves a more efficient exposure to questions for the learner, and subsequently improves the learning gain when compared with traditional systems in which the difficulty increases sequentially. However, engagement was found not to differ between the two adaptivity conditions.

Contents

Chapter 1

Introduction

The integration of assistive technological means in educational systems has gained widespread attention in the past few years. In effect, much of the focus has been recently shifted towards the possibilities of having adaptive educational systems tailored to the characteristics, background, abilities, and disabilities of its user. These systems are usually comprised of two complementary models, the student model which tracks and produces information about the user and the instructional model which uses this information to make decisions about the methods of conveying the instructional content to the user. In this Thesis we chose to dedicate our efforts to exploring the effects of adapting the content's difficulty according to the student's inferred mastery level. In this section we will discuss our motivation for this decision and our approach in pursuing the matter.

In the ever traditional educational systems, practice exercises aimed at preparing a student for an exam are usually in the form of mundane paper-and-pencil questions from previous exams. A lot of these examinations rely primarily on the use of multiple choice questions for cognitive evaluation, as they are very convenient for assessing lower order cognition such as the recall of discrete facts, and they can also be designed to accommodate assessment of higher order cognition such as synthesis, creative thinking, and problem-solving [1]. Some of the most notable multiple choice examinations include the ACT and SAT standardized tests for college admission in the United states.

In preparing for such exams, the main problem lies in the fact that during critical times before the exam where the student wants to revise and practice for the subject at hand, they have to answer practice questions in a sequential manner with the inability to identify the questions that most suit their level of mastery of the subject or that quickly address gaps in their knowledge and understanding of a topic, which costs them a lot of valuable time. Therefore, it can be very impactful to design e-learning systems that adapt the difficulty of the questions selected for the student according to their level of ability which can be inferred from their ongoing performance.

A lot of recent research has been devoted to developing e-learning systems that incorporate adaptive difficulty such as in the case of serious games, however, the proposed adaptive difficulty methodology would usually be restricted to the scenario and mechanics

of the game. This raises the need for research into generic serious games that are item-based, so they can be easily refurnished with different pools of questions/items, which then enforces the need for an item-based method for adaptive difficulty.

A well established framework in the fields of Psychometrics and Education, which features an item-based adaptive difficulty mechanism is the computer adaptive testing methodology (CAT). However, the main goal of this method is the quick assessment of examinees, where it can efficiently and effectively estimate an examinee's ability without them having to answer the entirety of the exam's questions by manipulating the order in which the questions are administered. It has been proven that the adaptivity implemented in the CAT's item selection algorithm reduces the test length by up to 50%, ultimately reducing the overall exam time without compromising the test's validity and reliability.

Thereupon, we found that the CAT meets our design criteria in many different ways. It provides accurate and quick identification of the student's mastery level which can be leveraged for real-time adaptation, it offers an already accomplished method of adapting the questions' difficulty, and it is flexible enough to incorporate different styles of questions for different educational content serving different age groups. The one problem was how to adopt it for the purpose of learning rather than assessment.

Hereafter, in an extensive review of literature on the CAT's models and principles, we searched for a model that would be suitable for small scale programs, and that can be easily used by educators. We then found a convenient CAT model which aims to quickly classify the student as a master or non-master, and hereafter, we identified our main aim of implementation to be extending this CAT model to classify the student as an A, B or C student which would in turn serve us in modifying the model in such a way that the goal of the system would be to efficiently advance the student to a higher-order classification (eg. from B student to A student), rather than terminating after finding out the student's category classification the way the CAT should. This redesign achieves the goal of implementing a computer adaptive learning system (CAL) whose principal benefit is efficiency in learning rather than assessment. Furthermore, we investigated and analyzed peripheral problems that might impede our system when transitioning from CAT to CAL.

As a proof of concept, we then applied our CAL system to a pre-existing educational game which incorporates MCQs, and supplied it with a pool of multiple choice questions designed for an educational topic from a university-level course in the field of Pharmacy and Biotechnology.

Furthermore, this study was also concerned with comparing different adaptivity techniques in terms of their effect on the engagement level of the learner. Therefore, we followed a recent study that showed that adapting the game mechanics and UI according to the learner's emotional state had a significant positive effect on the learner's level of involvement, and thus, we implemented this additional adaptivity feature in our system to compare it to our adaptive difficulty feature in terms of the level of engagement it produces. Moreover, we proposed a redesign of the SAM questionnaire used in the system for emotional recognition with the addition of colors and emojis to make the emotional indicators more visually understandable for the user.

As a final step in implementation, we separated the two adaptivity features into two versions of the system, and the final task was to implement an selection algorithm alternative to the CAL algorithm that operates in the traditional manner of sequentially increasing the difficulty (non-adaptive), and which is strong enough for a fair comparison against the adaptive CAL selection algorithm.

Finally, we formulated the research question to be to evaluate and compare the effectiveness of a learning system that adapts its content's difficulty according to the user's performance by using the CAL method and another learning system that adapts the game mechanics and UI according to the user's affective state with conventional non-adaptive difficulty levels, and to weigh up the two systems in terms of 3 factors; (1) learning efficiency, (2) academic gain, and (3) level of involvement. Hence, we conducted a comparative experiment on the two adaptive versions of the system to appropriately answer this research question.

Summing up, the tasks that outline our Thesis are as follows:

- Reviewing literature on the CAT principles and models. **(Chapter 2)**
- Extending the CAT expert system model and modifying it to produce the desired CAL system. **(Chapter 3)**
- Creating an Editor system for the educator to be able to install different pools of MCQs and generate a CAL-based practice system for their students. **(Chapter 3)**
- Changing the mechanics of a generic platform for educational games to incorporate MCQs and improving its front-end. **(Chapter 3)**
- Generating a visual design of the SAM questionnaire for emotional recognition. **(Chapter 3)**
- Applying adaptive game mechanics and UI according to the recognized emotions. **(Chapter 3)**
- Design and structure of the comparative experiment to answer the research question and evaluate the implemented work. **(Chapter 4)**
- Analyzing the results of the experiment. **(Chapter 5)**
- Drawing a conclusion and listing limitations and future work. **(Chapter 6)**

Chapter 2

Literature Review

2.1 Item Response Theory (IRT)

Item Response Theory (IRT) is a psychometrics paradigm invented by Fredrick Lord in the 1950s for the design, analysis and scoring of questionnaires and tests aimed at measuring intangible or arguably difficult-to-quantify traits and abilities. It is used by doctors to give more accuracy to questionnaires that measure physiological abilities, as well as test centres that aim to design tests that measure educational or cognitive achievements. In this paper we mostly focus on IRT in computer adaptive tests (CAT) for educational assessment. For more information on IRT, the reader is reffered to [**?**, **?**]. In this section we will focus on clearly defining some important IRT concepts and terminologies that are needed to get comfortable with for the purpose of this research discussion.

A foundational key to an IRT model is its Item Response Function (IRF) which is an estimate of the likelihood or the probability of different responses to an item by people with different levels of trait to be measured.
For example, in a questionnaire to determine a person‘s fitness level, given that a subject‘s current estimated level of fitness is low, if a question about how difficult they find it to do 50 squats in under 5 minutes is administered to them, it can be estimated that the subject answering “very difficult“ to be the most probable response to this item.

There are three well known models for IRT, each with its own version of IRF depending on the number of parameters they use:

1. The Dichotomous Rasch Model: It is the simplest known model. Dichotomous signifies that there are only two response possibilities to an item either correct or incorrect, or a yes or no, while polytomous models consider multiple answers to an item. A Rasch model means that only one parameter is used in the calculation of the item‘s response function which is the item‘s difficulty denoted by ‘b‘. As seen in

the following figure, b_i denotes the one item parameter which is its difficulty and θ_j denotes the subject's level of trait to be measured. The probability function here, which is the item's IRF, can be considered as an answer to the question what is the probability that a person j answers item i correctly given that his skill level is θ.

$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_j)},$$

2. 2-Parameter Logistic Model (2PL): Intuitively this model incorporates two item parameters to calculate the IRF, item difficulty denoted by 'b' and item discrimination denoted by 'a'.

3. 3-Parameter Logistic Model (3PL): It is the most powerful IRT model as it includes all three significant item parameters in its calculation, item difficulty, discrimination and guessing parameters denoted by 'b', 'a' and 'c' respectively.

A more thorough illustration of these item parameters is needed to get a clear understanding of what calibration is based on in IRT.

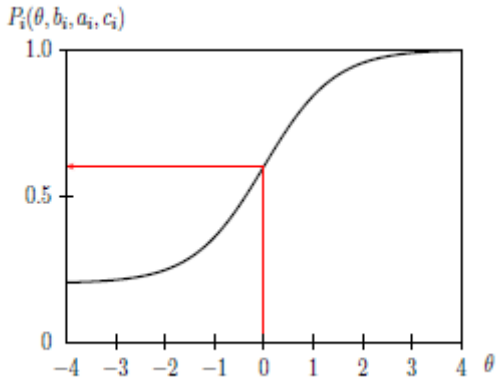


Figure 2.1: Item response function of a 3-PL item

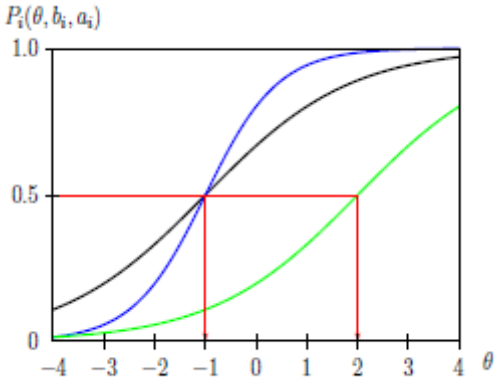


Figure 2.2: Item response functions of three 2-PL item

Figure 2.1 shows a single item whose probability of a correct answer is a function of 3 fixed parameters: difficulty, discrimination and guessing, as well as the varying parameter θ which is the person's skill level. Note that θ varies from -4 (low skill level) to 4 (high skill level).
Figure 2.2 shows 3 items whose probability of a correct answer is a function of difficulty and discrimination parameters only as well the varying skill level θ.

- b - item difficulty: It is θ at the mid-point of the curve, also known as the point of median probability. So in figure 2.1, it is 0 signifying an item of average difficulty. It is worth noting that this is why IRT is very useful at selecting questions in adaptive systems such as CAT (more about this in the following section), as it places the person‘s skill level at the same metric as the item‘s difficulty.

- a - item discrimination: It is the slope of the tangent at the mid-point. A highly discriminating item is very useful at discriminating between very close skill levels. For example, in figure 2.2 there are two items displayed that have the same difficulty -1, however the blue item is more discriminating as a small difference in θ correspond to a big difference in the probability of answering the item correctly.

- c - item pseudo-guessing: It is the lower asymptote of the IRF graph in figure 2.1 or P(-∞) which means that no matter how low θ of the subject is there is a “c“ chance that he could answer this item correctly by “guessing“.

2.2 Computer Adaptive Testing (CAT)

Computerized adaptive testing (CAT) is the redesign of psychological and educational measuring instruments for delivery by interactive computers. [2]. In a CAT, questions are selected whose difficulty level is closest to the particular person‘s ability level. For example, if a person answers a question incorrectly, an easier question is administered and if he answers it correctly, subsequently a more difficult question is selected. According to [2], the main objective of this adaptivity feature is to select for each examinee a set of questions from a pre-calibrated question bank that most effectively and efficiently measure their trait level. This calibration and selection is done through the use of powerful psychometric models such as IRT (2.1).

According to [19, 8, 28, 31], computer based tests have proven to uphold multiple benefits in comparison to conventional paper-and-pencil tests in that they automate the marking process and thus reduce marking workload and shorten the period that students have to wait to get their exam results. One prominent drawback to general computer based tests however is efficiency. The main motivation and reasoning behind replacing simple computer based tests with adaptive ones is that administering multiple advanced

questions to a student of poor skill level does little in assessing his ability as he will just answer most of them wrong, in the same way that giving a large number of easy questions to a proficient student gives little information about his precise skill level as he will mostly answer all of them correctly [23]. Implementing adaptivity in item selection ultimately reduces the exam time and is estimated to reduce the test length by up to 50% without jeopardizing the test's validity and reliability [26, 13].

A lot of examinations including Graduate Management Admission Test, Test of English as a foreign language and Microsoft Certified Professional have adopted the CAT methodology [23] and a lot of research has been devoted to exploring the applicability of CATs to other types of examinations and testing its precision and efficiency compared to normal computer based tests.
However, not enough research has explored the possibility of using CAT principles and theories to design computer programs that target educational training rather than assessment. There is also little effort directed towards maneuvering around sophisticated algorithms and item calibration issues that underlie large scale CAT programs and to redesign them into a simpler framework for the development of small scale adaptive programs.

(Nathan, and David, 2011) [40] lay out a non-comprehensive yet powerful framework for the development of CATs. According to this framework, there are five components that compose a CAT from an architectural point of view:

1. Calibrated item bank
2. Item selection algorithm
3. Starting point
4. Scoring algorithm
5. Termination criterion.

In this section, we will briefly discuss these five components and we will be slightly more comprehensive with some component concepts that interplay with our research direction.

Calibrated Item bank

In an educational context the item bank is simply a calibrated set of questions that the CAT should be programmed to choose from. Calibration means marking each question with a number to place it on a certain parameter scale. For example, marking each question with its level of difficulty on a scale from -4 to 4 in the 1PL IRT model.
There are several critical considerations when designing the item bank, the most important one being which psychometric model will be followed to calibrate this item bank.

The CAT item bank is mostly calibrated using IRT parameters (2.1), usually following the 3PL model. One of IRT‘s most important features being that it places items and examinees on the same scale.
However, papers such as [42] suggest that 500 to 1000 examinees are needed for calibration using the 3 parameter IRT model. There are also multiple IRT calibration softwares such as PARAM [35] and XCalibre, however they all state that you need a minimum of 100 examinees and a large number of items for the system to work. This means that, prior to all the challenges of implementing the actual selection software, a testing phase is needed for the initial estimation of item parameters in which hundreds of participants with varying competencies must take the lengthy and mundane test.
While the large number of examinees required for this initial phase doesn‘t pose a real problem for professional testing agencies and assessment programs, it comes as a real roadblock for small scale instructional programs that want to incorporate CATs [16]. For these kinds of programs, as (Frick, 1992) [16] puts it, “the IRT approach to adaptive testing can be likened to the use of a cannon to kill a mosquito.“ Another calibration model is proposed by (Frick, 1992) [16] that stems from the recognition of CATs as Expert systems. A detailed explanation of this model is handled in 2.3.

Item Selection Algorithm

The item selection algorithm follows from the chosen calibration model. In the 2-PL and 3PL models, the selection algorithm always aims at selecting the item with the most compatible difficulty to the examinee‘s ability estimate θ as well as choosing the most discriminating item for assessment efficiency.
This could pose a small problem where the low-discriminating items are very unlikely to be administered, and these items could have important information. Therefore, there is often a sub-algorithm that implements an item exposure control strategy or a content constraint strategy where a certain percentage of items covering each topic must be administered.

Starting Point

As previously mentioned, one of the selection criteria is compatibility between item difficulty and examinee ability estimate. An initial question when developing CAT engines is what to consider the examinee‘s ability estimate at the very beginning. If previous information is known about the examinee, for example his test score from the previous test administration, then this score could serve as his starting point. If no previous information is available, CAT systems often assume that the examinee is of average ability, hence the first administered item will be of medium difficulty. The latter approach however has a distinct disadvantage in terms of item exposure and test security. If all examinees have the same initial ability estimate, then they will all have the same initial test item and possibly a highly similar initial sequence of items. In such cases where this is a crucial issue, some strategic randomization can be implemented [17].

Scoring Algorithm

On the one hand, the IRT model has shown exceptional precision in point estimation of examinee ability, that's why most CATs utilize IRT for scoring as well as selection. On the other hand, papers such as [34] showed that CATs designed with classical test theory can be quite efficient in the classification of examinees into mastery level categories.
Well known scoring methods used within these models include maximum likelihood and Bayesian methods. Simulation studies are usually conducted later to compare the efficiency of the different scoring algorithms.

Termination Criterion

The CAT terminates when it has achieved some confidence that the current examinee ability estimate is in fact their actual ability. The most common approach to this is the minimum standard error criterion, where the test is designed to stop when the examinee has reached a certain standard error, or equivalently, a certain level of percision. This algorithm is often subject to a couple of practical constraints, typically, the maximum test length and the minimum test length. In the case where an examinee has a latent ability equal to the cut-off score of the test, the test could continue indefinitely without reaching a conclusion, therefore, the maximum test length serves to ensure that the entire item bank is not administered. In general, a test with more items rendered produces more precise scores and vice versa. Therefore, simulation studies are often conducted to make decisions about these termination parameters prior to the publishment of a CAT.

2.3 CAT as Expert Systems

(Frick, 1992) [16] proposed an alternative approach to item bank calibration that combats the problem of the need for such a large number of examinees imposed by the IRT approach to serve for small scale testing projects. However, he emphasizes that the developer following this approach should have an important objective in mind, which is to classify a student into master or non-master with regards to a single instructional objective. Test items should thus be designed to match that objective.
In contrast, large scale tests assess a wide range of instructional objectives (eg. algebra, fractions, geometry..etc), additionally they probably desire to give the student a point grade instead of just a classification and for these tests this model will fail.

In his paper, Frick analyzes and compares CATs to expert systems from the field of artificial intelligence such as MYCIN, a famous early expert system for diagnosing bacterial infections. He presents the following arguments to emphasize his recognition of CATs as one type of an expert system:

- The IRFs in a CAT can be viewed as a set of rules that dictate that if the examinee level is X and item Y is administered, the probability that the student answers the item correctly is predicted to be Z. In this view, a calibrated item bank constitute the knowledge base of a CAT expert system.

- The CAT's scoring algorithm is a form of inference engine that uses this knowledge base and the examinee‘s responses to questions to make deductions about the examinee‘s ability.

- Expert systems aim to make judgments, typically by attempting to choose one option from a number of mutually exclusive and exhaustive decisions. In like manner, a typical goal of a CAT is to estimate an examinee‘s ability level with enough confidence to terminate and make a decision such as master/non-master or a grade classification.

- Expert systems collect information to reach their decisions, likewise a CAT's selection algorithm continuously selects the question that gives the most information about the examinee's ability level.

He follows from this contention to formulate a framework for the development of a CAT as an expert system starting with the construction of the knowledge base, proceeding with the inference engine and concluding with an intelligent item selection algorithm. In this section we will describe his model in the same sequence.

The Knowledge Base

For each item in the item pool, 4 probability rules are developed which are in the same if-then form of the production rules that comprise the knowledge-bases in MYCIN:

- Rule 1: if (examinee is master $\wedge$ item i is selected) $\rightarrow$ probability of correct response = P(Ci|M).

- Rule 2: if (examinee is master $\wedge$ item i is selected) $\rightarrow$ probability of incorrect response = P($\neg$ Ci|M).

- Rule 3: if (examinee is non-master $\wedge$ item i is selected) $\rightarrow$ probability of correct response = P(Ci|N).

- Rule 4: if (examinee is non-master $\wedge$ item i is selected) $\rightarrow$ probability of incorrect response = P($\neg$ Ci|N).

Furthermore, the following instructions are followed to calculate the probability rules for each item in the item pool:

1. Give the test to a representative group of examinees, half masters and half non-masters with an expected wide range of test scores.

2. Choose mastery cut off score (eg.85%).

3. Divide examinees into mastery and non-mastery based on their test scores and the mastery cut-off score.

4. For each item in the test pool, calculate its probability rules as follows:

$$P(Ci \mid M) = (\#T_{im} + 1)/(\#T_{im} + \#W_{im} + 2) \tag{2.1}$$
$$P(\neg Ci \mid M) = 1 - P(Ci \mid M) \tag{2.2}$$
$$P(Ci \mid N) = (\#T_{in} + 1)/(\#T_{in} + \#W_{in} + 2) \tag{2.3}$$
$$P(\neg Ci \mid N) = 1 - P(Ci \mid N) \tag{2.4}$$

Where,
$\#T_{im}$ denotes the number of mastery students who answered that item correctly
$\#W_{im}$ denotes the number of mastery students who answered that item incorrectly
$\#T_{in}$ denotes the number of non-mastery students who answered that item correctly
$\#W_{in}$ denotes the number of non-mastery students who answered that item incorrectly.

Note that this reasoning could be extended to letter grade categorizations.

The Inference Engine

In the CAT expert system, the scoring method is Bayesian and termination follows from a Sequential Probability Ratio Test (SPRT)[14, 37, 41].

After each item administration, a likelihood ratio is computed that compares the current likelihood of the examinee being a master to the current likelihood of the examinee being a non-master. This Bayesian based reasoning utilizes the knowledge base in its calculation of the likelihood ratio as follows:

$$LR = \frac{P_{om} \prod_{i=1}^{n} P(C_i \mid M)^s [1 - P(C_i \mid M)]^f}{P_{on} \prod_{i=1}^{n} P(C_i \mid N)^s [1 - P(C_i \mid N)]^f} \tag{2.5}$$

Where,
P_{om} = prior probability that the examinee is a master
P_{on} = prior probability that the examinee is a non-master
(With no prior information, the alternatives are equally likely-i.e., $P_{om} = P_{on} = 0.50$.)
and
s = 1,f = 0 if item i is answered correctly

s = 0, f = 1 if item i is answered incorrectly
s = 0, f = 0 if item i has not been administered
For a numerical example of this Bayesian reasoning process, refer to [15].
If an extension is desired where classification includes more than 2 mastery categories, more than one LR is computed. For example, in an 'A', 'B', 'C' letter classification, 3 Likelihood ratios are needed, one that compares A to B, one that compares B to C and one that compares A to C.

The SPRT then determines the termination criterion by defining the following 3 decision rules:
Rule S1 If LR $\geq (1-\beta)/\alpha$, then terminate and choose master.
Rule S2 If LR $\leq \beta/(1\text{-}\alpha)$, then terminate and choose non-master.
Rule S3 If $\beta/(1\text{-}\alpha) <$LR $< (1\text{-}\beta)/\alpha$, then select another item, update the LR and apply the three rules again.

$(1\text{-}\beta)/\alpha$ can thus be thought of as the lower bound for the mastery threshold (LBM) and $\beta/(1\text{-}\alpha)$ as the upper bound for the non-mastery threshold (UBN).
α and β are type I and II decision errors.

A lengthy explanation of the SPRT and the decision errors is discussed in 2.4, along with a modified version of the test that classifies the examinee into one of 3 letter categories 'A', 'B' or 'C', instead of the two mastery categories.

Intelligent Item Selection

Frick had previously composed a model where items were selected randomly and without replacement [14]. However, it was criticized by Thomas Plew as it didn't use any information about the items in the selection process [30]. And so, in this paper he joined with Plew to develop an intelligent item selection procedure which was modelled after Weiss and Kingsbury's *"maximum information search and select"* procedure (MISS).

He orderly calculates for each item 3 extra parameters which are derived from the production rules calculated earlier; item discrimination, item/examinee incompatibility index and item utility. Following we explain the idea behind the mathematics of each item parameter and how the parameters interplay in item selection with the aim of testing efficiency.

- **item discrimination:** if the aim of the test is to classify the examinee as a master or non-master, then item discrimination is how discriminating this item is between masters and non-masters. For example, if a large number of master students out of all the master students in the estimation sample answered the item correctly, while a small number of non-master students out of all the non-master students answered

the item correctly then this item is highly discriminating. In contrast, if the two ratios are nearly equivalent irrespective of how small or large their value is, then this item is of low discrimination.
A highly discriminating item means that it is one of the "golden items" that only the master students were capable of answering correctly and which subsequently contributed to their high scores. It's worth noting that this plays the most role in test efficiency, because when an examinee answers highly discriminating items correctly, it can be estimated that he would probably answer other items correctly as well without having to actually answer them, in other words it raises the likelihood that they're a master and the opposite is true for the case of an incorrect answer. Therefore, for an item i, discrimination is calculated as the difference between the probability of a correct answer by a master and the probability of a correct answer by a non-master:

$$D_i = P(C_i \mid M) - P(C_i \mid N) \tag{2.6}$$

- **item/examinee incompatibility index:** Not only do we want to administer items of high discrimination but we also want to administer items whose difficulty matches the examinee's current ability level estimate. The aim of this from an assessment efficiency perspective is that we don't want to ask the examinee questions that are too hard or too easy compared to their ability level, we want to ask questions that the examinee has a fifty/fifty chance of answering correctly, to be able to collect the most information about their actual ability per selection.
Accordingly, we initially calculate for each item its *item difficulty* which is based on the probability that this item is answered correctly in general i.e. by masters and non-masters alike:

$$P(C_i) = (\#r_i + 1)/(\#r_i + \#w_i + 2) \tag{2.7}$$

Where,
$\#r_i$ denotes the number of students who answered item i correctly in the estimation sample.
$\#w_i$ denotes the number of students who answered item i incorrectly in the estimation sample.
And thus, the item difficulty is $1 - P(C_i)$ i.e. probability of an incorrect answer.

Then, we calculate the estimate of the examinee's achievement in a way to be comparable to the items' difficulties 2.1, and so we calculate it as the ratio between the number of items that the examinee answered correctly and the total number of items that the examinee answered so far:

$$E(\theta_j) = (\#r_j + 1)/(\#r_j + \#w_j + 2) \tag{2.8}$$

Where,
$\#r_j$ denotes the number of items that examinee j answered correctly so far.

$\#w_j$ denotes the number of items that examinee j answered incorrectly so far. Finally, we calculate for each item i its incompatibility index as the absolute distance between its difficulty and examinee j's current achievement estimate:

$$I_{ij} = abs\{(1 - P(C_i)) - E(\theta_j)\} \tag{2.9}$$

Note that this means that the achievement estimate is upgraded every time the examinee answers another item (correct or incorrect). And so the items‘ incompatibility indices are also upgraded per answer. Note also that at the *starting point* of the test, when the examinee hasn't yet answered any of the items, their achievement estimate from the above calculation equals 1/2, which means we initially consider the examinee to be of average ability i.e. they answered half of the items correctly. In turn, the initial item to be administered will be of medium difficulty.

- **item utility:** Finally, an item‘s utility is calculated as the ratio between the item's discrimination and incompatibility index. Thereby, utility defines how utilizable the item currently is in terms of discrimination and compatibility, and the amount of resulting information it subsequently promises if selected next for the examinee. The next selected item will then be that of the greatest utility, which means that it will be the remaining one which is *most discriminating* between masters and non-masters and *least incompatible* with the examinee‘s current achievement estimate.

$$U_{ij} = D_i/(I_{ij} + 0.0000001) \tag{2.10}$$

Note that in such manner, the items' utilities will be recalculated along with each update of the incompatibility indices.

In effect, Frick's EXSPRT-I model is comparable to the 2-PL IRT model discussed in 2.1 in that both item difficulty and discrimination are considered in the selection process.

2.4 Multiple-Category Classification in SPRT

In 1947, Abraham Wald devised the sequential probability ratio test (SPRT) as a decision methodology for choosing between two discrete hypotheses when observations are made sequentially. [41] It has since been widely applied in quality control of manufactured goods and in 1983, Kingsbury and Weiss [21], and Reckase [32], among others, have explored the use of SPRT in cognitive tests for classifying examinees into masters and non-masters. The main difference between SPRT and conventional statistical tests is that normal tests make decisions after a certain group of observations have been made, while SPRT applies decision making rules after every observation and terminates when one of the two discrete alternatives can be chosen with a certain level of confidence. The major advantage of this is a reduction in the average sample size needed to reach a decision, or in the case of mastery computer-based tests, a significant reduction in the test length.

While SPRT is usually applied in situations requiring a decision between two simple hypotheses, (Spray, 1993) [38] proposed a modification of SPRT to include situations involving multiple hypotheses. In this section we will discuss the important concepts involved in simple SPRT with 2 discrete hypotheses and steadily transition to Spray's extension of the mathematical model, to classify the examinee into one of three letter categorizations.

Simple SPRT

According to Reckase, the two hypotheses between which we aim to decide are, the null hypothesis where the student is a non-master and in which case his latent ability θ_i is less than the cut-off score δ, and the hypothesis that the student is a master and his latent ability θ_i is greater than or equal to δ. δ is called the passing score or *decision point.*
$H_0 : \theta_i < \delta$. (non-master)
$H_1 : \theta_i >= \delta$. (master)

In order to make resolute decisions about whether the uni-dimensional θ_i is sufficiently low to decide for the null hypothesis H_0 or instead sufficiently high to decide for H_1, we set upper and lower limits to the passing criterion δ, where θ_0 is the lower limit and θ_1 is the upper limit, such that $\theta_0 < \delta < \theta_1$. These upper and lower limits are called *endpoints* and the region in between these endpoints is called the *indifference region.* The length of this region describes the precision the test aims for, as the larger the indifference region, the lower the precision and subsequently the shorter the test length. This, in fact, yields two weaker hypotheses that are used to test the original composite hypotheses:
$w_0 = \{\theta : \theta <= \theta_0\}$. (non-master)
$w_1 : \{\theta : \theta >= \theta_1\}$. (master)

As understood from 2.3, the ongoing comparison, between the likelihood that the examinee is a master and the likelihood that the examinee is a non-master, is accomplished by computing a likelihood ratio using the Bayesian method from equation 2.5. Thereby, this likelihood ratio is regarded as continuously comparing the two hypotheses in question.

$$LR = \frac{Prob(Examinee\ is\ master)}{Prob(Examinee\ is\ non{-}master)}$$

The most critical condition in a SPRT is deciding when we have achieved enough confidence to terminate testing and settle for one of the two hypotheses.

In order to do that, two error probabilities α and β are first defined, where:
Prob(choosing H_1 when H_0 is true) = α.
Prob(choosing H_0 when H_1 is true) = β.
As mentioned in 2.2, these error probabilities are determined a priori and computer simulations can be useful to test varying levels of error.

Wald then defined two likelihood ratio boundaries that are functions of α and β:
The lower boundary = $B >= \beta/(1-\alpha)$.
The upper boundary = $A <= (1-\beta)/\alpha$.
The lower boundary of the likelihood ratio (B) is regarded as the upper bound for the non-mastery threshold (UBN) and the upper bound of the likelihood ratio (A) is regarded as the lower bound for the mastery threshold (LBM)[14].

Finally, 3 decision rules are constructed:
Rule S1 If $LR >= (1-\beta)/\alpha$, then terminate and choose master (H_1)
Rule S2 If $LR <= \beta/(1-\alpha)$, then terminate and choose non-master (H_0)
Rule S3 If $\beta/(1-\alpha) < LR < (1-\beta)/\alpha$, then conduct another observation, update the LR and apply the three rules again.

As previously discussed in 2.2, in practice, this algorithm is sometimes subject to a maximum test length constraint (MTL). When this maximum number of items is reached and no classification under the likelihood-ratio test has occurred, a forced classification can be made according to the following distance rule:

$$MIN\{|log(LR) - log(A)|, |log(LR) - log(B)|\} \tag{2.11}$$

Where,
LR = The likelihood-ratio at the time the classification is made.
A = The pre-determined upper boundary.
B = The pre-determined lower boundary.

Extension to Three Letter Categorizations

In the case where an extension is desired and classification includes more than 2 mastery categories, the number of decisions, hypotheses, likelihood ratios and even item discrimination parameters is incremented. For example, in an 'A', 'B', 'C' letter classification, two cut-off scores or *decision points* ,δ_1 and δ_2, are indicated, which yields 3 possible decisions to choose from:
Decision 1: $\theta_i < \delta_1$. (C student)
Decision 2: $\delta_1 <= \theta_i < \delta_2$. (B student)
Decision 3: $\delta_2 <= \theta_i$. (A student)

As shown in the simple model, in order to be able to make confident decisions, the SPRT tests hypotheses about θ_i defined by the endpoints of the indifference region. In the case of 2 decision points, 3 endpoints are needed to construct 3 indifference regions, θ_1, θ_2 and θ_3, such that:

$$\theta_1 < \delta_1 < \theta_2 < \delta_2 < \theta_3$$

and where the first indifference region $= \theta_2 - \theta_1$, the second indifference region $= \theta_3 - \theta_2$ and the last indifference region $= \theta_3 - \theta_1$.
These endpoints are calculated in an orderly fashion. One endpoint θ_2 is chosen midway between δ_1 and δ_2, and then the second and third endpoints are set to be equidistant from their respective decision points as θ_2 is from δ_1, as per the following calculations:
$MIDIST = (\delta_2 - \delta_1)/2$.
$\theta_1 = \delta_1 - MIDIST$.
$\theta_2 = \delta_1 + MIDIST$.
$\theta_3 = \delta_2 + MIDIST$.

Once the endpoints are established, 3 pair-sets of SPRT hypotheses are formulated:
$H1 : \theta_i = \theta_1 \quad H2 : \theta_i = \theta_2 \quad H3 : \theta_i = \theta_1$
$H1\text{‘} : \theta_i = \theta_2 \quad H2\text{‘} : \theta_i = \theta_3 \quad H3\text{‘} : \theta_i = \theta_3$

It can now be seen that each set of hypotheses compares 2 of the 3 main decisions against each other. For example, hypotheses H1 and H1‘ compare decision 1, where the examinee level is C, against decision 2, where examinee level is B. Therefore, 3 likelihood ratios are calculated to compare the two hypotheses for each pair-set, one that compares B to C, one that compares A to B and one that compares A to C:

$$LR1 = \frac{Prob(Examinee\ level\ is\ B)}{Prob(Examinee\ level\ is\ C)}$$

$$LR2 = \frac{Prob(Examinee\ level\ is\ A)}{Prob(Examinee\ level\ is\ B)}$$

$$LR3 = \frac{Prob(Examinee\ level\ is\ A)}{Prob(Examinee\ level\ is\ C)}$$

While the extension in the likelihood ratio was a mere duplication with simple changes, the extension to the test termination algorithm is a bit more complicated.

- **Establishing the error rates:** In order to formulate the decision rules, desired error rates must primarily be provided. These are used to derive the critical upper and lower limits of the likelihood ratios. Let $P_{h|j}$ designate the probability that θ_h is accepted, when θ_j is true, h = 1,2,3 and j = 1,2,3. Given that, $P_{h|h} = P_{j|j} =$ the power of a single SPRT, this leaves 6 error rates ($P_{1|2}$, $P_{1|3}$, $P_{2|1}$ ..etc), two for each set of hypotheses pair.
  For example, for hypotheses H1 and H1‘:
  $P_{1|2} = Prob(choosing\ H1\ when\ H1\text{‘}\ is\ true)$
  $P_{2|1} = Prob(choosing\ H1\text{‘}\ when\ H1\ is\ true)$
  It makes intuitive sense that the larger the distance is between θ_h and θ_j, the lower the error rate desired (the larger the indifference region, the lower the precision). Therefore, calculations of error rates corresponding to certain endpoints are a function of the indifference region bounded by these endpoints. For the exact formulas, the reader is referred to [38].

- **Establishing likelihood ratio boundaries:** Following the calculation of the error rates, lower and upper boundaries are determined for each likelihood ratio. For a likelihood ratio that is comparing $H_0 : \theta = \theta_h$ versus $H_1 : \theta = \theta_j$, the upper boundary is $power/P_{j|h}$ and the lower boundary is $P_{h|j}/power$, h = 1,2,3; j=1,2,3; h $\neq$ j.
  For example, for LR1 that compares H1 ($\theta_i = \theta_1$) vs H1‘($\theta_i = \theta_2$):
  The upper boundary = $\frac{power}{P_{2|1}}$
  The lower boundary = $\frac{P_{1|1}}{power}$

- **Defining the decision rules:** All 3 sets of hypotheses are tested after each item response and the following decisions are made based on the results:

  - Decision 1 ($\theta_i < \delta_1$) is made when **H1 and H3** are both accepted. (C student)
  - Decision 2 ($\delta_1 <= \theta_i < \delta_2$) is made when **H1‘ and H2** are both accepted. (B student)
  - Decision 3 ($\delta_2 <= \theta_i$) is made when **H2‘ and H3‘** are both accepted. (A student)
  - Otherwise, keep testing.

  The hypotheses are tested by testing their likelihood ratios against the likelihood ratio boundaries. Therefore, the following decision rules are formulated and checked after each observation:

  - *Rule S1:* If ($LR1 <= LB1$) **and** ($LR3 <= LB3$), then terminate and choose C student.
  - *Rule S2:* If ($LR1 >= UB1$) **and** ($LR2 <= LB2$), then terminate and choose B student.
  - *Rule S3:* If ($LR2 >= UB2$) **and** ($LR3 >= UB3$), then terminate and choose A student.
  - *Rule S4:* Else, conduct another observation, update the likelihood ratios and apply the 4 rules again.

  Where,

$$LB1 = The\ lower\ bound\ for\ LR1 = \frac{P_{1|2}}{power}. \tag{2.12}$$

$$UB1 = The\ upper\ bound\ for\ LR1 = \frac{power}{P_{2|1}}. \tag{2.13}$$

$$LB2 = The\ lower\ bound\ for\ LR2 = \frac{P_{2|3}}{power}. \tag{2.14}$$

$$UB2 = The\ upper\ bound\ for\ LR2 = \frac{power}{P_{3|2}}. \tag{2.15}$$

$$LB3 = The\ lower\ bound\ for\ LR3 = \frac{P_{1|3}}{power}. \tag{2.16}$$

$$UB3 = The\ upper\ bound\ for\ LR3 = \frac{power}{P_{3|1}} \quad . \tag{2.17}$$

Note that in item selection algorithms, it is no longer applicable to rank items according to one decision point (eg. according to discrimination between master and non-masters). A reasonable compromise is to rank items according to the decision point that is closest to the current estimate of examinee ability.

2.5 Emotions and Educational Games

An educational game falls under the realm of serious games which are games designed for a primary purpose other than pure entertainment. The study of the advantages of serious games has gradually increased within the research community in the field of E-Learning. A primary key feature in well-designed serious games is the intrinsic pleasure dimension. There is the sensory pleasure for example which lies in the visual richness of the material in terms of colors, patterns and their coherent flow, as well as the acoustic stimuli that strengthen the play experience [3]. Another pleasurable aspect to serious games is providing the learner with a safe environment for curiosity and exploration. Simulated environments provide the player with a safe setting where they can make mistakes without being harshly reprimanded for their false decisions [33, 20]. This reduces the players anxiety towards failures and their real life consequences as well as boosts their self-esteem. A powerful factor in serious games that serves such a purpose over traditional instructional exercises is their capability to give immediate feedback with the aim of encouraging the players exploratory attitudes. Positive feedbacks for example equip the player with the emotional readiness to accept higher goals and explore levels of higher difficulties [5, 22]. Negative feedbacks that focus on the players tactics and not on the players individual identity take the role of notifying the player with the need for a strategy change [4, 25]. These pleasure-related advantages to serious games amongst many others play a key role in inducing positive emotions that enhance the learning experience. Positive emotions were found to capture the learners attention and reduce their distraction with other non-emotional resources [12], as well as empower their memory

in terms of persistent and detailed recall [29]. Despite the many advantages to serious games and e-learning systems, previous studies showed that inadequately implemented e-learning systems can be counter-productive and result in confusion, frustration and reduced learner interest [18, 27].

2.6 Affective Computing

Affective Computing Affective Computing is an interdisciplinary field which incorporates Psychology, Cognition and Computer Science [39]. According to Tao and Tan [39], it aims at assigning human-like empathatic skills to a computer such as observation (noticing signs that indicate an emotion), interpretation (classifying what has been observed into its corresponding emotion) and generation of affective behavior (adapting to these emotions through support, sympathy..etc). In the following section, we will be focusing on the observational aspect of Affective Computing and we will compare the different state-of-the-art assessment methods currently used. We will also give one example of a self-report scale used for emotional assessment.

According to [3], the observation part of Affective Computing is accomplished through either subjective measurements or objective measurements.

Subjective measurements dictate that the subject uses self-report scales or questionnaires to report their emotions and affective state. Advantages of such methodology include that it is unobtrusive and in the case of adaptive systems, it can give the user conscious control over the adaptive mechanics to allow for a form of customization within the system. However, the case of the user self-reporting their affective state is argued to be an inaccurate emotional assessment as its subject to social desirability factors and intentional manipulations. It is also limited by the users awareness and capability of assessing their own emotions. Other disadvantages of subjective measurements in gaming environments for example is that it can be annoying and it would interrupt the flow of the users gaming experience as well as interrupting the unfolding of the emotional process and could thus change it.

Objective measurements on the other hand could be classified into two non-separate categories: physiological correlates monitored through bio-sensors and/or sensory observations of facial expressions, gestures and speech. Advantages of using these measurements are that they offer a higher level of accuracy, in that they are not subject to intentional manipulation or suffer the social desirability factors as in the case of subject measurements. They also dont depend on the subjects capability to communicate their own feelings and they dont interrupt or interfere with the emotional process. In the case of physiological monitoring, physiological data such as changes in skin conductance level, heart rate and body temperature are triggered by the Autonomic Nervous System (ANS) [3] in response to the experienced emotions and are then collected and analyzed through specific sensory devices which translate this data to their respective emotions. However, findings such as [7] suggest that ANS activity indicates more the

arousal level of an emotion (fear as well as anger) rather than detecting and isolating a specific emotion. Thus, physiological correlates arent generally dependable in all systems that deal with affectivity. Moreover, other disadvantages in the case of bio-sensors and other wearables are that they can be quite obtrusive and can limit the subjects freedom or range of motion and their interaction with the system. In the case of assessing facial expressions, gestures and acoustic indications, such various emotional expressions vary between cultures and depend greatly on context as one expression could indicate multiple different emotions, and thus would make it difficult to draw out a single definitive model for classifying the results of these observances and mapping them to their corresponding emotions. Its highly recommended [3] therefore, to use a multimodal approach that incorporates multiple different modalities in objective emotional measurement in order for them to be effective.

Current general challenges in the field of affective computing are such that, emotions are hard to quantify and label, there is an insistent demand for a multi-modal approach and sensor fusion, and that the generation of adaptive affective behavior varies contextually.

2.7 Subjective Measurements

The subjective measurement is a data reporting method at which the user is asked to consciously self-report the required data through given questionnaires [36].

One of the applications of the subjective measurements is in the field of emotion recognition, as there are many existing questionnaires and scales that aim to evaluate the feelings of the user during a certain activity or towards a certain product or given test materials.

In this section we will describe a 3 different examples of questionnaires and scales used to assess the subject's emotional state:

- **Product Emotion Measurement instrument (PrEmo):** PrEmo is a non-verbal self-report instrument that measures 14 emotions that are often related to product design. Seven of these emotions are pleasant and another 7 are unpleasant [9].

  Each of the 14 emotions are expressed by an animation with facial, vocal and bodily expressions to assist the user in identifying the emotion.

  After the participants clicks on the emotion to which they most relate, a three-point scale appears which represents these three states: I do feel the emotion, to some extent I feel the emotion and I do not feel the emotion expressed by this animation. The participant then has to choose one of the three options.

- **Positive and Negative Affect Schedule questionnaire (PANAS):** According to [43], the PANAS consists of a list of 20 adjectives that describe different emotional

states: 10 states of Positive Affect (PA) and 10 states of Negative Affect (NA). Positive Affect denotes activity and pleasure, while Negative Affect denotes stress and fear.

Due to its length, it is unsuitable for frequent evaluations of the subject's emotional state, and is instead used to evaluate long-lasting emotional conditions.

Each of the 20 adjectives is rated by the participant on a 5 point Likert Scale.

- **Self-Assessment Manikin (SAM):** The SAM is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli [6].

  The graphical figures scale of the SAM ranges from a smiling, happy figure to a sad, unhappy one when representing the valence dimension, and ranges from an excited, wide-eyed figure to a relaxed, sleepy figure for the arousal dimension, the dominance dimension represents changes in control with changes in the size of the figures where a large figure indicates more control in the situation. The participant has to choose the figure that describes his current state in each of the three dimensions.

  It has also shown to provide a simple and fast wasy to assess the user's emotion that can depend only on the valence and arousal levels.

2.8 Related Work

In addition to reviewing the CAT principles and models which we decided to use in our implementation of the system, we also reviewed different educational games and programs that embedded an adaptive difficulty feature into their systems. We also reviewed some systems that implemented emotional adaptivity features with the aim of enhancing the learning experience. This review allowed us to pinpoint drawbacks with the current systems in this field and to put effort into making a valuable contribution to what has been reached so far. The most relevant implemented systems are discussed below:

One of the most relevant research studies conducted in adaptive game difficulty was an investigation performed by Jeroen Linssen in the adaptive possibilities in an educational game called "Code Red: Triage"[24]. Code Red: Triage is a serious game created by Van Oostendorp and Van der Spek in 2007, which trained players in performing *triages*. A triage is a procedure performed on an injured human being through which medical personnel can determine in how much need of help that person is, and the triages presented in the game were of varying complexities.

Thus, the main aim of implementation in Linssen's study was to adaptively present the triage cases to the player according to their skill level, and his research question was mainly concerned with whether this adaptivity would improve the learning efficiency of the player in Code Red: Triage, and whether the challenging factor of this adaptivity feature would subsequently improve their engagement level.

A comparative experiment was conducted comparing the game with the adaptive difficulty feature against a non-adaptive version of the game. Results from the experiment confirmed that adapting the game's difficulty let players gain knowledge faster (learning efficiency), however, it proved to be insignificant in terms of improving the engagement level.

Notable shortcomings of the methodology within this research study were first, that the adaptive difficulty featured a "one-way" variation in difficulty, meaning that when the player is performing well, the game is developed to skip triages, however, if the player is performing poorly, there is no way of reducing the difficulty of the game. Second, the adaptive difficulty is to an extent confined to the Code Red: Triage game and can not be easily adopted to suit other educational games. Lastly, the player must always start the game from the beginning regardless of his knowledge prior to playing which restricts the game's efficiency.

Another study concerned with item-based computer adaptive learning explored item selection methods traditionally developed for CATs for their usefulness in optimizing computer based learning [10]. In addition to exploring CATs' selection algorithms, (Eggen, 2012) also develops and compares an alternative selection procedure based on Kullback-Leibner information. He then conducts simulation studies to evaluate and compare the usefulness of the different selection algorithms.

The most prominent strength in this study is that it draws the distinction between optimization in testing and optimization in learning, and thus the proposed selection algorithm is constantly monitoring the student's *growth* in ability, and selecting items that feed this growth rather than items that merely disclose their ability.

The last study we would like to mention in this section [36] is a research into the effectiveness of adapting several game mechanics to a specific set of emotions on the learning experience of the student. The adaptivity feature was implemented in an educational game in which the specific emotion is repetitively inferred using the SAM questionnaire. The mechanics that were changed were the game's timer, scoring method, music, and theme color.

An experiment was then held that compares the game with the adaptivity feature against a non-adaptive version of the game in terms of learning gain and engagement level.

The results of the experiment showed a significant advantage in playing the game with the adaptivity feature over the non-adaptive version in both the learning gain and engagement level.

One impediment in the implemented system was that the participants were not able to instantaneously see which emotions the different SAM choices reflected.

Chapter 3

Methodology

This chapter describes the methodology followed to answer the research question. It starts by discussing the main objective of the study, followed by the details of the reasoning behind the decisions made in the system design. Finally, it describes in detail the games included in the platform, their features and the system implementation.

3.1 Main Objective

In light of the research question, the aim of this study is to compare two adaptivity features within the context of a serious game with regards to their effect on learning, adapting the interface and game mechanics according to the user‘s emotions and adapting the content difficulty according to the user‘s performance.

As previously discussed, practice exercises aimed at preparing a student for an exam are usually in the form of mundane paper-and-pencils questions from previous exams and in many educational systems, this is in the form of multiple choice questions. Therefore, one of the fundamental aims of implementation is to design a platform with engaging serious games that can incorporate multiple choice questions. The platform should also be generic, which means that it can be serviced by educators of multiple subjects to generate games with practice questions from their own syllabi.
However, the main problem lies in the fact that during critical times before final exams where a student wants to revise and practice for a subject‘s exam, they have to answer practice questions in a sequential manner with the inability to identify the questions that most suit their level of mastery of the subject or that quickly address gaps in their knowledge and understanding of a topic, which costs them a lot of valuable time. Thus, the platform was implemented to tackle this issue by implementing an adaptivity engine that cleverly selects the questions that are both compatible with the student‘s mastery level as well as challenging and beneficial.

Additionally, e-learning systems have been previously used to improve the learning motivation and engagement of the student by adapting to the user‘s affective state and impression. However, persistent challenges facing these systems today are in the methods used to observe indications of the user‘s emotional state whether by collecting their subjective measurements or monitoring their physiological correlates. For this reason, a visual design of the SAM assessment technique was created to infer the user‘s affective state and adapt several gaming features and mechanics appropriately to further enhance the learning experience.

Accordingly, the main objective of this study is to compare adapting the gaming interface and mechanics versus adapting the questions‘ difficulty in terms of three factors, learning efficiency, academic gain and engagement. And thus, the main features that were incorporated into our methodology are:

- Serious games platform that incorporates MCQs.
- Generic and reusable.
- Model that infers the user‘s emotions and adapts the user interface and game mechanics accordingly.
- Engine that infers the user‘s mastery level through their ongoing performance and adapts the difficulty of the questions administered accordingly.

And this is what will be further discussed in this chapter.

3.2 From CAT to CAL

As mentioned previously, one of the main features of the platform is an engine that is supposed to infer the user‘s mastery level and administer questions that most suit their ability while still challenging their competence. For this purpose, the Computer Adaptive Testing (CAT) methodology (2.2) was adopted as a foundation for a computer adaptive learning (CAL) system. There is then a feedback loop where this CAL engine feeds the questions to be rendered to the games described in 3.4 and receives the responses to the questions from the games to make calculations about the student's performance.

Choosing the CAT methodology was primarily based on the fact that it already had a mechanism of adapting the questions‘ difficulties according to the examinee‘s performance. If the examinee answers a question wrong, they receive an easier question, if they answer correctly they get a relatively more difficult question, so they don‘t waste their time answering questions that are too hard or too easy for their ability level.

Moreover, a CAT‘s main aim is to efficiently assess the student's ability without them having to answer all the questions in a question pool, which would be beneficial as part

of an adaptive learning system to provide further insight into the user‘s current level of dominion over a subject.

Last but not least, CATs are built on psychometric models which estimate the examinee‘s ability level and the questions‘ difficulties under the same metric. This makes the user‘s monitored ability level estimate feasibly comparable to the current questions in the question pool.

Following this perception, the CAT's different models and principles were investigated as shown in the literature review to design the platform's CAL system.

As mentioned earlier, most CATs are built on IRT (2.1), however, the problem with IRT lies in the large number of students needed for calibration of the item pool 500-1000 which would be burdensome for an educator using our system. (2.2) So instead, Frick‘s model (2.3) is used which views CATs as an Expert system with production rules, inference engine and an informative item selection algorithm. Deciding to use this model was based on the fact that it doesn't need as many calibration students (a minimum of 50). The main difference, however, between Frick‘s model and CATs based on IRT is that IRT provides a point estimation of examinee ability, meaning that it gives a specific number as a final score. On the other hand, Frick‘s model can only classify the student as a master or non-master, with the possibility of an extension to grade classifications. Nevertheless, this suffices for small scale programs.

In light of this rationale, our CAL system consists of two phases which render questions to the user, Phase I, which is a CAT expert system, aims to classify the user (examinee) as an A, B or C student, and Phase II benefits from the user‘s test result in Phase I, along with the items‘ information, pre-attained through calibration, to further elect the questions that most challenge their current ability level.
The CAT in Phase I is an extension of Frick‘s model, with Spray‘s modification of the SPRT (2.4) to serve for the termination criterion.

3.2.1 Phase I: a CAT expert system

In this subsection, we will describe the CAT expert system used to classify the user an examinee into an A, B or C student, by describing its calibration engine and a coherent inference and selection engine. Finally, we will report the approach used to set the initial values of the parameters, which are used in the latter engine, for the starting point of the test.

Calibration Engine

In designing the CAT expert system, Frick first calibrates the questions/items according to two parameters; Item difficulty and item discrimination (2PL model). The calibration

described here constitutes a knowledge-base which is merely an extension of the one detailed in 2.3.

Item difficulty is simply considered to be the probability that the item is answered incorrectly $(1 - P(C_i))$, while item discrimination is not as straightforward.

Since the aim of the test is to classify the user/examinee into A, B or C letter categorizations, there are 3 item discriminations for each item: Item discrimination between A and B students ($D_i\{AB\}$), item discrimination between B and C students ($D_i\{BC\}$) and discrimination between A and C students ($D_i\{AC\}$).

This calibration is derived empirically by giving the complete set of test questions to a representative group of examinees, known as calibration students, with varying competencies (to expect a varying range of test scores).

Afterwards, two mastery cut-offs are chosen (cut-off A and cut-off B) and the calibration students are accordingly divided into those who are A students, those who are B students and those who are C students. Methods for choosing the cut-off scores are described in the section on the Editor (3.3).

Difficulty and discrimination parameters are then calculated for each item through the following production rules:

- $P(C_i) = \frac{\#r_i+1}{\#r_i+\#w_i+2}$
- $P(\neg C_i) = 1 - P(C_i)$ **- - item difficulty**
- $P(C_i \mid A) = \frac{\#T_{iA}+1}{\#T_{iA}+\#W_{iA}+2}$
- $P(\neg C_i \mid A) = 1 - P(C_i \mid A)$
- $P(C_i \mid B) = \frac{\#T_{iB}+1}{\#T_{iB}+\#W_{iB}+2}$
- $P(\neg C_i \mid B) = 1 - P(C_i \mid B)$
- $P(C_i \mid C) = \frac{\#T_{iC}+1}{\#T_{iC}+\#W_{iC}+2}$
- $P(\neg C_i \mid C) = 1 - P(C_i \mid C)$
- $D_i\{BC\} = P(C_i \mid B) - P(C_i \mid C)$ **- - item discrimination between B & C**
- $D_i\{AB\} = P(C_i \mid A) - P(C_i \mid B)$ **- - item discrimination between A & B**
- $D_i\{AC\} = P(C_i \mid A) - P(C_i \mid C)$ **- - item discrimination between A & C**

Note that the notation used in equations [2.1 to 2.4] and 2.7 is only extended, such that $P(C_i \mid A)$ is the probability of a correct answer by an A student, $\#T_{iA}$ is the number of A students in the estimation sample who answered item i correctly, $\#W_{iA}$ is the number of A students who answered item i incorrectly, and the same for $P(C_i \mid B)$ and $P(C_i \mid C)$ designating B and C student correct answer probabilities respectively.

Unified Inference and Selection Engine

After calibration, the items are now ready for the inference and selection engines.

(i) The Scoring method in the CAT Expert System follows a Bayesian based reasoning process where 3 Likelihood Ratios are formulated and updated each time the user answers a question:

$$LR1 = \frac{P_{oB} \prod_{i=1}^{n} P(C_i \mid B)^s [1 - P(C_i \mid B)]^f}{P_{oC} \prod_{i=1}^{n} P(C_i \mid C)^s [1 - P(C_i \mid C)]^f} \tag{3.1}$$

$$LR2 = \frac{P_{oA} \prod_{i=1}^{n} P(C_i \mid A)^s [1 - P(C_i \mid A)]^f}{P_{oB} \prod_{i=1}^{n} P(C_i \mid B)^s [1 - P(C_i \mid B)]^f} \tag{3.2}$$

$$LR3 = \frac{P_{oA} \prod_{i=1}^{n} P(C_i \mid A)^s [1 - P(C_i \mid A)]^f}{P_{oC} \prod_{i=1}^{n} P(C_i \mid C)^s [1 - P(C_i \mid C)]^f} \tag{3.3}$$

Where,

LR1 continuously compares the likelihood that the user is a B student versus the likelihood that the user is a C student.

LR2 continuously compares the likelihood that the user is an A versus the likelihood that the user is a B student.

LR3 continuously compares the likelihood that the user is an A versus the likelihood that the user is a C student.

PoA, PoB and PoC is the prior probability that the user is an A, B and C student respectively.

s=0, f=1 if item i is answered incorrectly.
s=1, f=0 if item i is answered correctly.
s=1, f=1 if item i hasn‘t been answered yet.

n is the number of items in the question pool.

(ii) Test termination follows from Spray‘s extended version of the SPRT described in 2.4. The 4 stopping rules are thus:

- *Rule S1:* If $(LR1 <= LB1)$ **and** $(LR3 <= LB3)$, then terminate and choose C student.
- *Rule S2:* If $(LR1 >= UB1)$ **and** $(LR2 <= LB2)$, then terminate and choose B student.
- *Rule S3:* If $(LR2 >= UB2)$ **and** $(LR3 >= UB3)$, then terminate and choose A student.
- *Rule S4:* Otherwise, select another question, update the likelihood ratios and apply the 4 rules again.

Where,

$$LB1 = The\ lower\ bound\ for\ LR1\ according\ to\ 2.12. \tag{3.4}$$
$$UB1 = The\ upper\ bound\ for\ LR1\ according\ to\ 2.13. \tag{3.5}$$
$$LB2 = The\ lower\ bound\ for\ LR2\ according\ to\ 2.14. \tag{3.6}$$
$$UB2 = The\ upper\ bound\ for\ LR2\ according\ to\ 2.15. \tag{3.7}$$
$$LB3 = The\ lower\ bound\ for\ LR3\ according\ to\ 2.16. \tag{3.8}$$
$$UB3 = The\ upper\ bound\ for\ LR3\ according\ to\ 2.17\ . \tag{3.9}$$

However, it is worth noting that termination in the CAL system only means proceeding to Phase II.

Furthermore, a minimum and maximum test length constraints are inaugurated. In general, a test with more items rendered produces a more precise score and vice versa [40]. Therefore, the two TL constraints ensure that the test *is slow enough* to make precise decisions and *fast enough* to reserve some items and save enough time for phase II, thereby benefiting from the CAT system‘s assessment (reduction of test length).

We will later describe what happens when the maximum test length is reached.

(iii) Selection is based on maximum information search and select (MISS), in which the user‘s achievement estimate $E(\theta_j)$ is saved and continuously re-calculated each time the user makes a correct or incorrect answer, as per equation 2.8. The item incompatibility index (I_{ij}) is also continuously re-calculated using equation 2.9. The previously proposed item selection algorithm in 2.3 then picks the item that is most discriminating and least incompatible with the examinee‘s achievement estimate by choosing the item with the current greatest utility (U_{ij}) after equation 2.10. The desired item selection algorithm, however, is a bit challenging as there is no longer one discrimination parameter for each item but 3.
Therefore, Three item utility parameters are calculated for each item based on the three discrimination parameters along with the incompatibility index:

$$U_{ij}\{BC\} = D_i\{BC\}/(I_{ij} + 0.0000001) \tag{3.10}$$
$$U_{ij}\{AB\} = D_i\{AB\}/(I_{ij} + 0.0000001) \tag{3.11}$$
$$U_{ij}\{AC\} = D_i\{AC\}/(I_{ij} + 0.0000001) \tag{3.12}$$

Then, we choose to rank the items according to one of the 3 utilities depending on the classification the user is currently closest to. For example, if the SPRT is closing in on the user being classified as a B student, then we check, if LR1 is further from its upper bound than LR2 is from its lower bound, then we need to shorten the distance between LR1 and its upper bound, by selecting a question that discriminates best between B and C students, to hasten the classification process.

This is then achieved by choosing the next selected item to be that of the greatest $U_{ij}\{BC\}$.

Accordingly, 3 distance equations, dA, dB, and dC are formulated that designate how far the user currently is from being classified as an A student, B student and C student respectively:

$$dA = (UB2 - LR2) + (UB2 - LR3). \tag{3.13}$$

$$dB = (UB1 - LR1) + (LR2 - LB2). \tag{3.14}$$

$$dC = (LR1 - LB1) + (LR3 - LB3). \tag{3.15}$$

Then, the minimum distance is chosen and the following algorithm is followed:

- If the minimum distance is dA, then the following rules are applied:
  - If $(UB2-LR2) > (UB2-LR3)$, then choose the item of greatest $U_{ij}\{AB\}$.
  - Otherwise, choose the item of greatest $U_{ij}\{AC\}$.
- Else, if the minimum distance is dB, then the following rules are applied:
  - If $(UB1-LR1) > (LR2-LB2)$, then choose the item of greatest $U_{ij}\{BC\}$.
  - Otherwise, choose the item of greatest $U_{ij}\{AB\}$.
- Finally, if the minimum distance is dC, then the following rules are applied:
  - If $(LR1-LB1) > (LR3-LB3)$, then choose the item of greatest $U_{ij}\{BC\}$.
  - Otherwise, choose the item of greatest $U_{ij}\{AC\}$.

Note that after it has been ascertained, which classification the user is currently closest to, the sub-algorithm aims to bring the user closer to that classification. Note also that despite the fact that this Utility selection algorithm is part of the Item Selection Engine, it uses the LRs which are associated with the CAT's Inference Engine. Hence the unification of the two engines in this CAT model.

As previously mentioned, a practical maximum TL constraint is often used for efficiency purposes. When the maximum TL has been reached, a forced classification is imposed according to the following distance rule which is similar to that in equation 2.11:

$$MIN\{dA, dB, dC\} \tag{3.16}$$

so that,
if dA is the minimum, the user is shortly classified as an A student.
if dB is the minimum, the user is shortly classified as a B student.
if dC is the minimum, the user is shortly classified as a C student.

Starting Point

Initially, the CAT must draw out an assumption of the examinee's current ability level estimate and consequent to that, the first item's difficulty will be chosen. A reasonable

compromise is to assign a value to it corresponding to the average score. Therefore, the user's achievement estimate is initially set to $\frac{1}{2}$, as if they answered half the items correctly, thus the first rendered item will be of medium difficulty. Check section 6.2.2 for what happens when a former user takes the test again.

Additionally, the prior probabilities, PoA, PoB, and PoC will be set to $\frac{1}{2}$ so that the likelihood ratios, LR1, LR2, and LR3 all amount to 1, meaning that the user is equally likely to be a B student as they are to be a C student, equally likely to be an A student as they are to be a B student, and equally likely to be an A student as they are to be a C student. In other words, no prior assumption is developed about the user's classification.

3.2.2 Phase II: From CAT to CAL

With the ongoing adaptive difficulty feature at hand, the CAT system works reasonably well on its own in training the user. However, two important questions remain, which are, what to do after the CAT test finishes and how to benefit from its result. Trying to cope with these two questions, an initial approach was followed which consisted of a series of CAT repetitions, however, it generated a number of issues when tested. Thereafter, another approach was investigated, which served as a continuation to the CAT test, and which theoretically promised better results.

Initial intuition and issues

Our initial intuition was to implement test "re-runs" using Frick's CAT expert system as it is, with two mastery categories and only one cut-off score, and to maximize on the difference it makes to change this cut-off score. We set 3 possible cut-off scores (A-Score, B-Score, C-Score), each of which could divide the calibration students in the estimation sample into masters and non-masters, with the 3 cut-off scores set based on a scaling system described in the editor (3.3). Each cut-off score would then be used to create 3 different calibrations of the same question pool -i.e. 3 pools with the same questions, however, the questions are calibrated differently in each. The difference in calibration is only in the item discrimination parameter. Let's consider the situation where there are two cut-off scores, 50%, and 85%. The low cut-off score of 50% would designate any above-average student as a master and the rest will be non-masters, so the items of high discrimination will be items that differentiate between average students and non-average students. On the other hand, if the question pool was calibrated based on the 85% cut-off score, only students who are high-achievers would be considered as masters and the rest will be non-masters, in which case the items of high discrimination will be items that differentiate between high-achieving students, and average or below average students. Therefore, our guess was that a CAT with a question pool calibrated at a higher cut-off score would render discriminating items that are generally more difficult (affiliated with higher category students) than one where the cut-off score was low.

The problem with this approach was that it disrupted the evenness in the population's division into masters and non-masters. This happens when the mastery cut-off score is

diverted to one of the extremes (A or C cut-off scores). For example, if the mastery cut-off score is the A cut-off, a few students would be considered to be masters while a large number of students would be considered to be non-masters. The main problem here is that it does not definitively differentiate between A students and the students of the next lower category (B students) but rather between A students and any other students. This caused the test to terminate very quickly as a few items would determine the student's classification.

Moreover, the short test length issue meant that some items of low discrimination have a zero chance of being rendered, while if the test were to continue indefinitely, low discriminating items would eventually be rendered despite being last in order. This hightlighted our need to extend the model to multiple grade classification.

Fixing the ranking utility approach

Our second approach was to use an extended version of the CAT expert system model to classify the user into one of the A, B, C level categories, and use the test result along with the information about the items to promote the questions that most suitably challenge the user's established ability level and most efficiently address their knowledge gaps. This is achieved by looking at the item utility parameters from a different perspective.

As discussed earlier, during a CAT, after each user's response, the remaining items are sorted according to one of the three item utilities $U_{ij}\{AB\}$, $U_{ij}\{BC\}$, and $U_{ij}\{AC\}$, depending on the classification to which the user is currently closest according to the SPRT. In other words, the sorting utility is different each time. Alternatively, in Phase II the sorting utility is fixed based on the CAT's final classification result to be one of another set of 3 item utilities $\{U_{ij}\{AB\}, U_{ij}\{BC\}, U_{ij}\{A\}\}$ for the remainder of the session. Where $U_{ij}\{A\}$ is a fourth item utility parameter which is calculated according to the following equation:

$$U_{ij}\{A\} = P(\neg C_i \mid A)/I_{ij} \tag{3.17}$$

Henceforth, the following algorithm follows after the user has been identified to be an A, B, or C student by the CAT:
If the user is a C student, fix the ranking utility to be $U_{ij}\{CB\}$.
If the user is a B student, fix the ranking utility to be $U_{ij}\{AB\}$.
If the user is an A student, fix the ranking utility to be $U_{ij}\{A\}$.

In order to understand the idea behind this, consider the following analogy:

On the one hand, from an assessment perspective, ranking items according to their $U_{ij}\{AB\}$ means sorting them according to how useful they are in discriminating between A and B level categories to speed up the classification process.

On the other hand, from a learning perspective, ranking items according to $U_{ij}\{AB\}$ simply promotes the items whose likelihood of an A student responding correctly is higher

than the likelihood of a B student responding correctly, and thus could be viewed as sorting the items according to how beneficial they are in efficiently training a B student to become an A student.

Therefore, the aim of this algorithm is as follows:
If the user is classified as a C student, we rank the items according to $U_{ij}\{CB\}$ to promote rendering the items that C students answer incorrectly while B students answer correctly and thus efficiently transition the user from being a C student to being a B student.
Likewise, if the user is identified as a B student, then we fix the sorting utility to be $U_{ij}\{AB\}$ to continue by rendering items that the B students answer incorrectly but the A students answer correctly.
Finally, if the user is identified as an A student, then we do not have a higher level to transition them to and so we simply target the questions that A students answer incorrectly.

Summing up, after the CAT (Phase I) has terminated, we continue to update the user's achievement estimate every time they make a correct or incorrect answer, as well as, the item incompatibility index for each item to maintain compatibility between the next selected item's difficulty and the user's ongoing achievement.

Furthermore, item selection is fulfilled by using the CAT's result to fix the ranking parameter to be the item utility that would efficiently transition the user to the next higher level.

However, note that we no longer monitor the likelihood ratios for any assessment or termination purposes.

Hence, training could continue until all the questions have been rendered or until the user decides to stop and restart the test.

3.2.3 From CAT to CAL: Limitations, Problems & Solutions

There were a few critical considerations that had to be made when using the CAT model for the purpose of an educational system.

1. **Local Independence:** Probability theory, which is employed in the CAT's Bayesian method for scoring, dictates that the probability of a correct answer to any given question should not change depending on the order in which the questions are administered to produce valid statistical decisions, meaning that no feedback should be given to the user as it would affect their future answers. However, since the purpose of our system is educational development rather than assessment, immediate feedback is implemented where the user is notified when he/she chooses an incorrect answer and is given a chance to try choosing another answer, they are also notified and rewarded when they choose a correct answer. Under these conditions, the assumption of independence of observations is violated which could result in classification errors. The consequences of this compromise remain questionable.

2. **Permanent item removal:** The mathematics involved in the CAT's scoring system also dictate that an item can't be considered twice in the calculation of the likelihood ratios (3.1, 3.2, 3.3). This means that any item rendered to the user will be removed from the pool with no chance of it being re-rendered for reviewing. To compensate for this, an item queuing strategy is employed where any item that the user answers incorrectly is enqueued into a wrong answer queue. This wrong answer queue is then channeled to the car game described in 3.4.3 so that these items are rendered a second time to the user mid-game for reviewing.

3. **Discontinuous change in difficulty:** In the CAT model used the user's achievement estimate is continuously re-calculated according to equation 2.8 and depending on the updated value, the next selected item's difficulty will be determined. However, it was observed that this calculation causes abrupt and discontinuous change in difficulty.

   For example, suppose that the user had just started taking the test so that the number of his/her correct answers ($\#r_j$) is 0 and the number of his/her incorrect answers ($\#r_w$) is 0, thus the achievement estimate would be calculated as $\frac{1}{2}$. This means that we assume that the user is capable of answering half of the items correctly and then the first item to be selected would be of such a difficulty that half of the students in the estimation sample answered it incorrectly -i.e. An item of average difficulty. If the user then answers the selected item correctly, $\#r_j$ would be equal to 1 and the achievement estimate would become $\frac{2}{3}$ which means that we deem the user capable of answering two-thirds of the questions correctly and that the best item to render next is one where 66.7% of the students in the estimation sample answered it incorrectly. Note that, by this means, all items with incorrect answer probabilities (difficulties) between 50% and 66.7% will be skipped, which causes an increase in difficulty discontinuity. This does not pose a problem in a CAT with pure evaluation objectives, on the contrary, this improves the assessment efficiency by always administering questions that the examinee has a fifty/fifty chance of answering correctly. On the other hand, this causes an undesirable effect in our CAL system, where there is an unsteady increase/decrease in difficulty that could prompt feelings of being overwhelmed as a few correct answers would result in a sudden, intense increase in difficulty, or feelings of disappointment or disengagement when a few incorrect answers result in a rapid drop in the difficulty of the questions received.

   Therefore, the user's ongoing achievement in our CAL system is not calculated using equation 2.8 but is alternatively estimated after each item response from the following algorithm:

   - A set X of the remaining items' incorrect answer probabilities is formulated and sorted ascendingly, such that

   $$X = \{x : x = P(\neg C_i),\ 0 <= i <= n,\\ n \textit{ is the number of items remaining in the question pool}\}. \quad (3.18)$$

- If the user is at the starting point of the test, the achievement estimate takes the value of the median incorrect answer probability of the sequence.

$$E(\theta_j) = Med(X) \tag{3.19}$$

- If the user answers a certain selected item k correctly, the achievement estimate takes the value of the next higher incorrect answer probability in the sequence -i.e. The next higher item difficulty.

$$E(\theta_j) = x : x \in X, x > P(\neg C_k) \tag{3.20}$$

  Else,

$$E(\theta_j) = max(X). \tag{3.21}$$

- If the user answers a certain selected item k incorrectly **and** the user hasn't been classified as an A student, the achievement estimate takes the value of the next lower incorrect answer probability in the sequence -i.e. The next lower item difficulty.

$$E(\theta_j) = x : x \in X, x < P(\neg C_k). \tag{3.22}$$

  Else,

$$E(\theta_j) = min(X). \tag{3.23}$$

- If the user answers a certain selected item k correctly and the user has been classified as an A student, the achievement estimate takes the value of the incorrect answer probability less than or equal to current item's incorrect answer probability. -i.e. The next equivalent or lower item difficulty.

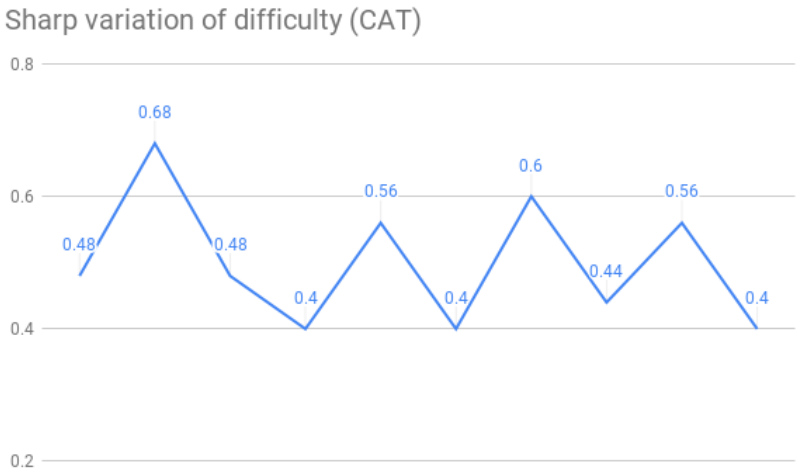
$$E(\theta_j) = x : x \in X, x <= P(\neg C_k). \tag{3.24}$$

  Else,

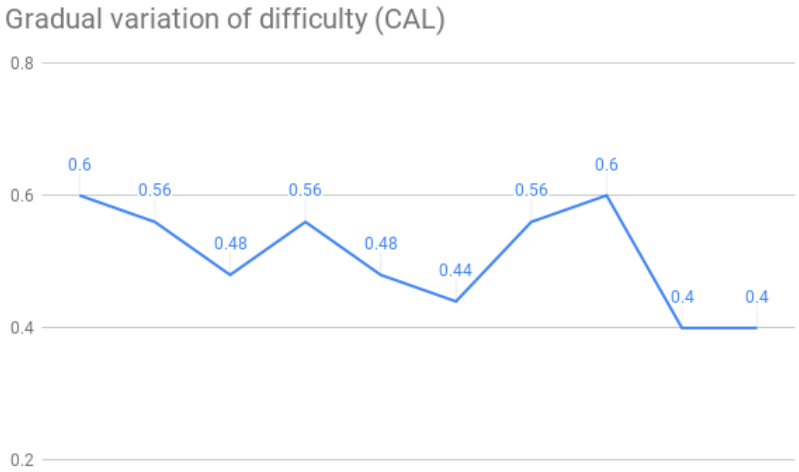
$$E(\theta_j) = min(X). \tag{3.25}$$

Note that in the case where there are multiple items of the same difficulty, equations 3.20 and 3.22 aim to render an item of a higher or lower difficulty relative to the selected item difficulty but not equal to it. Note also that the else part which yields equation 3.21, and both equations 3.23 and 3.25 designates the situation where the selected item was that of the highest difficulty and the situation where the selected item was that of the lowest difficulty, respectively. Finally, note that the aim of equation 3.24 is to slow down the rate of decrease in the difficulty of the questions in the case of a user classified as an A-student, as it is expected that the most difficult questions will be rendered and the user will make more mistakes than the usual.

Accordingly, the learner will experience a slow and steady variation of difficulty.



(a) Sharp Variation of difficulty (CAT).



(b) Gradual Variation of difficulty (CAL).

3.3 Editor

An important feature of this platform is its receptivity to having educators from a variety of subjects refurnish it with multiple choice questions, making it generic and reusable. This is achieved through an editor that allows the educator to install their own set of multiple choice questions and then prompts the user to provide information needed for the calibration process of the CAT expert system (3.2.1).

3.3.1 Multiple Choice Questions (MCQs)

To set up a working game with all the available features, the educator must first type in their set of multiple choice questions. This is done by stating the name they would like

their game to be called and then typing in the number of questions they want in their question pool (minimum 40). Once the user specifies the number of questions, a panel is set with the questions listed to each of which the user must specify the question title, the correct answer and the other 3 incorrect answers. The user then clicks save to move on to the next stage. If the user fails to fill in any of the questions' text boxes, an error message appears prompting the user not to leave any empty text boxes.

After this step, the questions are saved and ready for calibration.

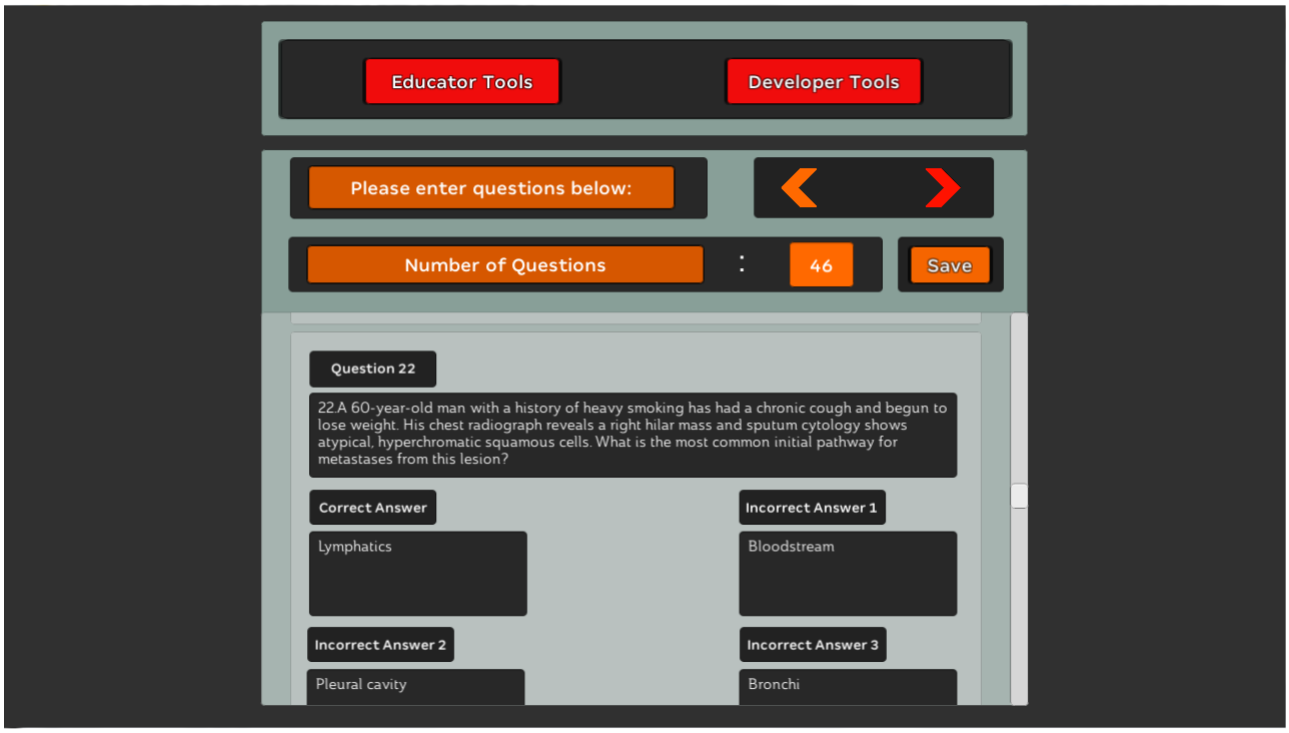


Figure 3.2: Providing the multiple choice questions.

3.3.2 Empirical Calibration

As specified in 3.2.1 and 2.3, the probability rules which are used to calculate the calibration parameters for each question are derived through an empirical process.

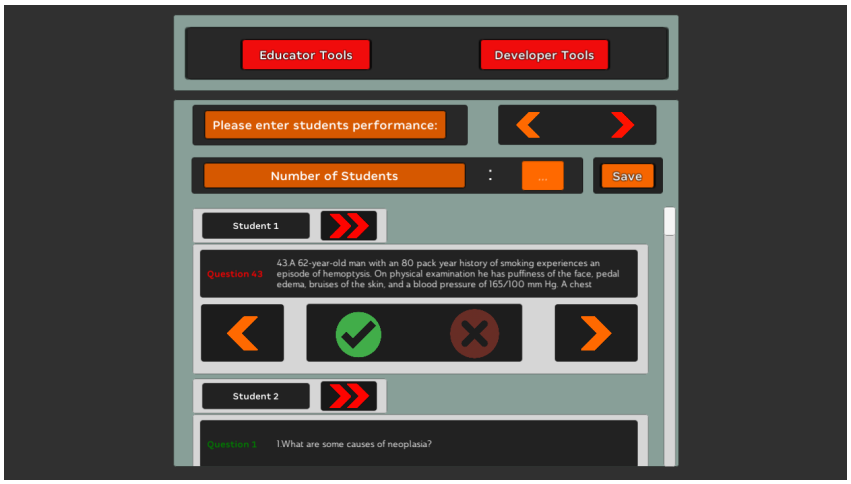
Initially, the educator must give the complete set of questions to a representative group of students as a test. The students selected to take the test must be of varying proficiency of the test's instructional objective. Afterwards, the educator must provide the CAL system with each student's response to each question on the test.

This is done in the editor by first specifying the number of students who took the test.

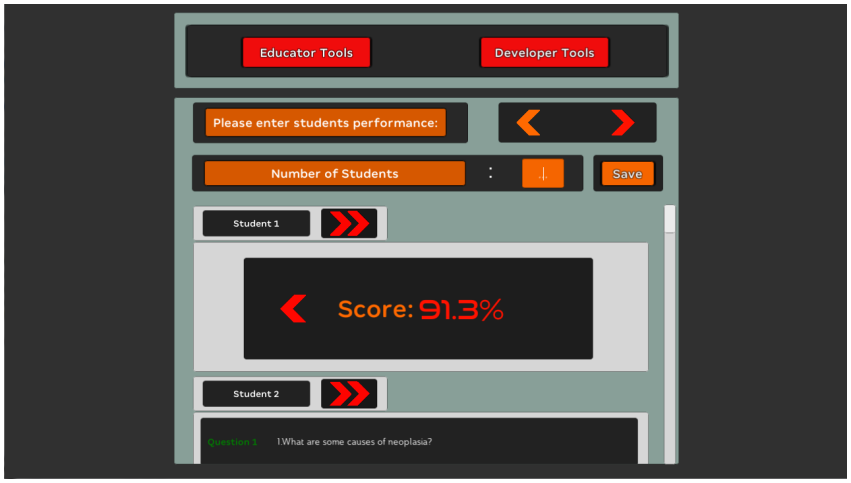
Then, the user must list for each student his/her responses to each of the questions in the question pool, whether their answer was correct or incorrect. This is facilitated

by clicking the green correct-sign button if the student answered the presented question correctly or the red cross button if the student answered the presented question incorrectly, and using the left and right arrow keys the educator could navigate back and forth through the questions as they please. There is also a fast forward button for each student in case the educator wants to skip providing the students answers one by one and proceed to the student's final score, in which case the student's unspecified answers are regarded as incorrect. (Figure 3.3a and 3.3b)

Finally, the user clicks the save button to save the students answers and proceed to the next step. However, all the students' final scores must be displayed in order to move to the next step, to ensure that the user is satisfied with sample's scores.



(a) Providing the calibration students' responses.



(b) Reviewing the calibration students' final scores.

3.3.3 Cut-off Scores

One of the factors that determines how the CAT will work and that should receive the most critical attention is how we decide the test's 2 cut-off scores, the A cut-off

which differentiates the A-level students from the B-level student and the B cut-off which differentiates B-level students from the C-level students. If the test was well designed and the estimation sample students were cleverly selected, ideally one-third of the students would be considered C-level students, one-third B-level students and one-third A-level students. Nevertheless, further insight about the resulting scores of the estimation sample is needed to carefully choose the cut-off scores.

To assist the educators in choosing the cut-off scores, the following distribution scales are produced that distribute the students in the estimation sample into the 3 level categories according to their given scores. The educator can then click on one of the distribution scales to view the A and B cut-off scores that would be need to achieve that specific distribution. These scales are:

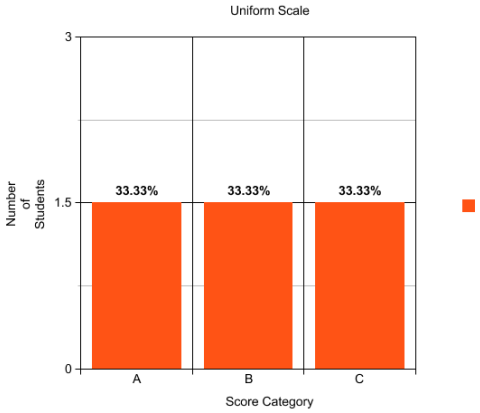
- The uniform scale, where the number of students is uniformly distributed among the 3 level categories. (Figure 3.12a)

- The average scale, where most of the students are categorized as B-level (average) students. (Figure 3.12b)

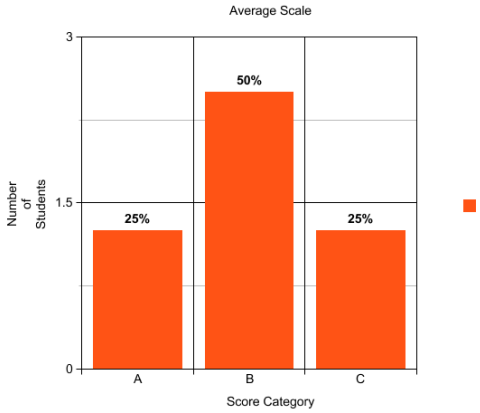
- The excellence scale, in which there are less A students than B students and less B students than C students -i.e. The higher the level, the fewer the students that qualify for it. (Figure 3.12c)

- The extremity scale, in which there are more students around the two extreme level-categories A and C than there are in the middle level-category B. (Figure 3.12d)

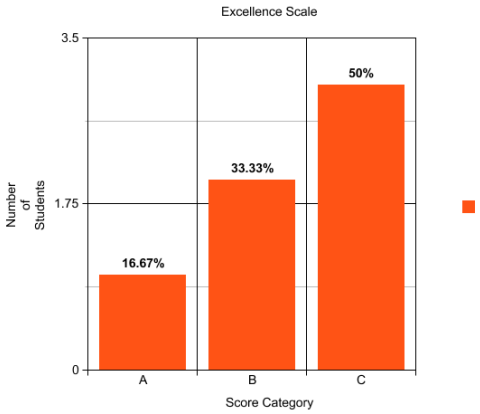
- The underdog scale, in which there are more A students than there B students, and more B student than C students -i.e. There is a larger opportunity for students to qualify for high level categories.(Figure 3.12f)



(a) Uniform Scale.



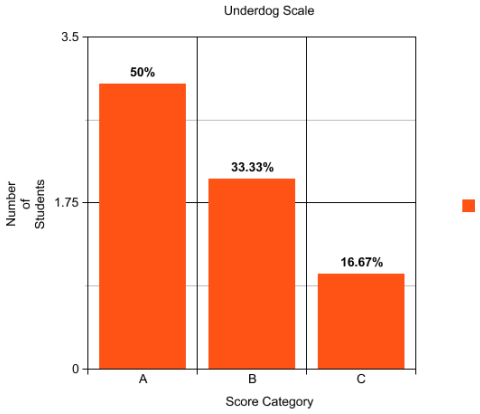
(b) Average Scale.



(c) Excellence Scale.

Scale.png

(d) Extremity Scale.



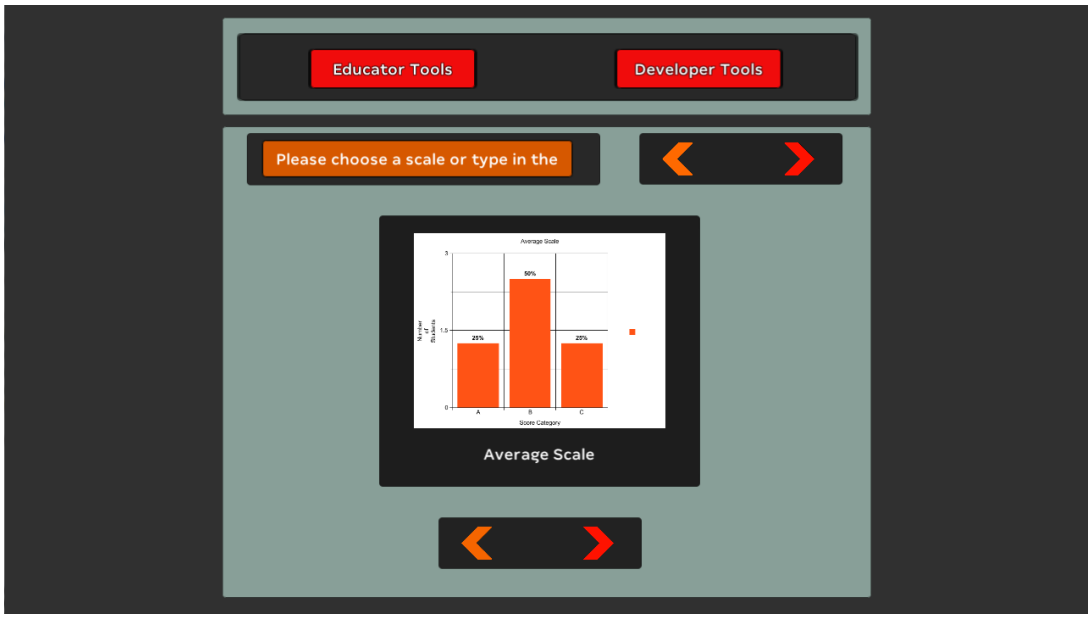
(e) Underdog Scale.

When the educator clicks on the scale according to which they choose to distribute the students in the estimation sample, the A and B cut-off scores are shown that would produce the chosen students' distribution as shown in Figure 3.5b. If the educator is satisfied with the resulting A and B scores, then they can click Save to finalize the 3 calibration stages and publish their fully adaptive MCQ game.

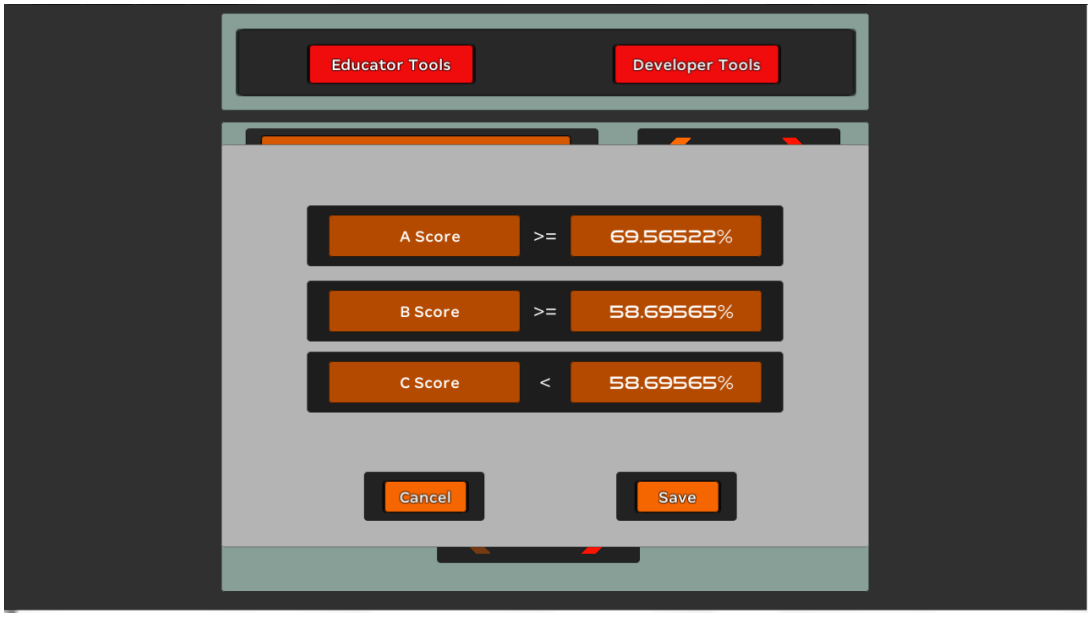
On the other hand, if the educator would like to hard-code the cut-off scores them-

selves, they are welcome to choose that option as shown in figure 3.5c . This is also assisted by providing the educator with a short report of the statistical mean, median, minimum and maximum of the population.

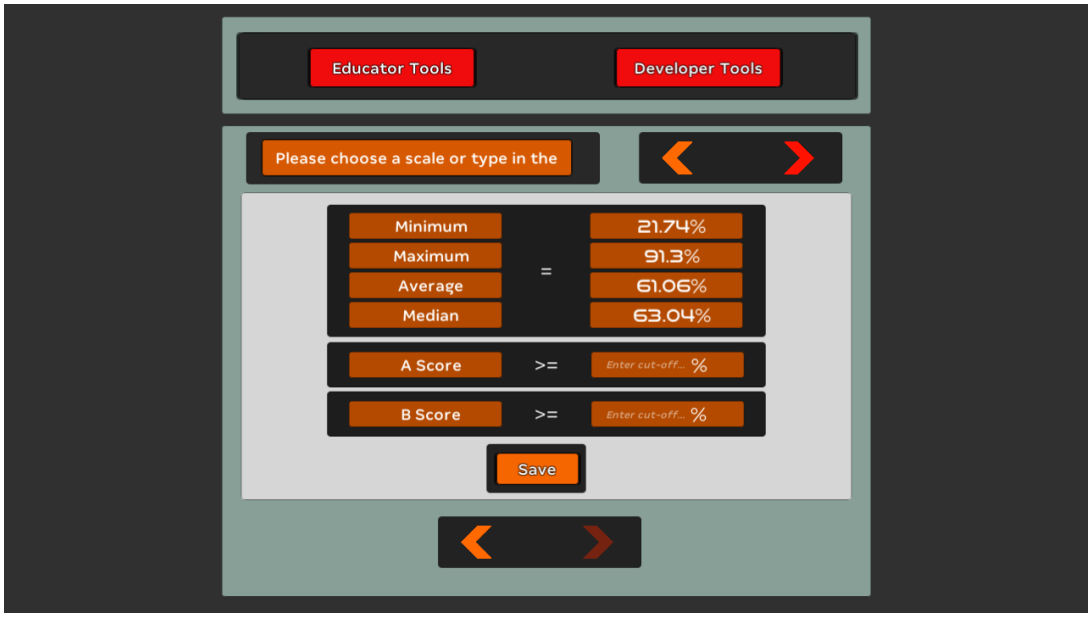
Note that in the CAT system, the A cut-off score is usually chosen to mark students who have fully mastered the educational objective and therefore do not need any training, but in the CAL system, the A cut-off score should be lower as it should be used to mark students who are of a very high level in training but still need training.



(a) Example of choosing the Average Scale.



(b) Resulting scores shown after choosing a Scale.



(c) Hard-coding the cut-off scores.

3.4 Games

As mentioned previously, the fundamental objective of the platform is to render MCQ practice exercises from a pre-designed question pool to train the learner in a specific topic. We already had the back-end for a platform that allows the educator to generate a collection of serious games and customize them to meet the needs of their subject. We decided to use 3 of the serious games in the platform and change their mechanics to be suitable for rendering multiple choice questions. We also worked on building the front-end for each of the games to give the player an attractive UI and improve the gaming experience. In this section we will describe the mechanics and features of these 3 games.

3.4.1 Pipes

The objective of this game is to place as many pipe parts as possible before the time ends to create a long pipe. What the player tries to avoid is the pipe eating itself or going outside the grid borders.

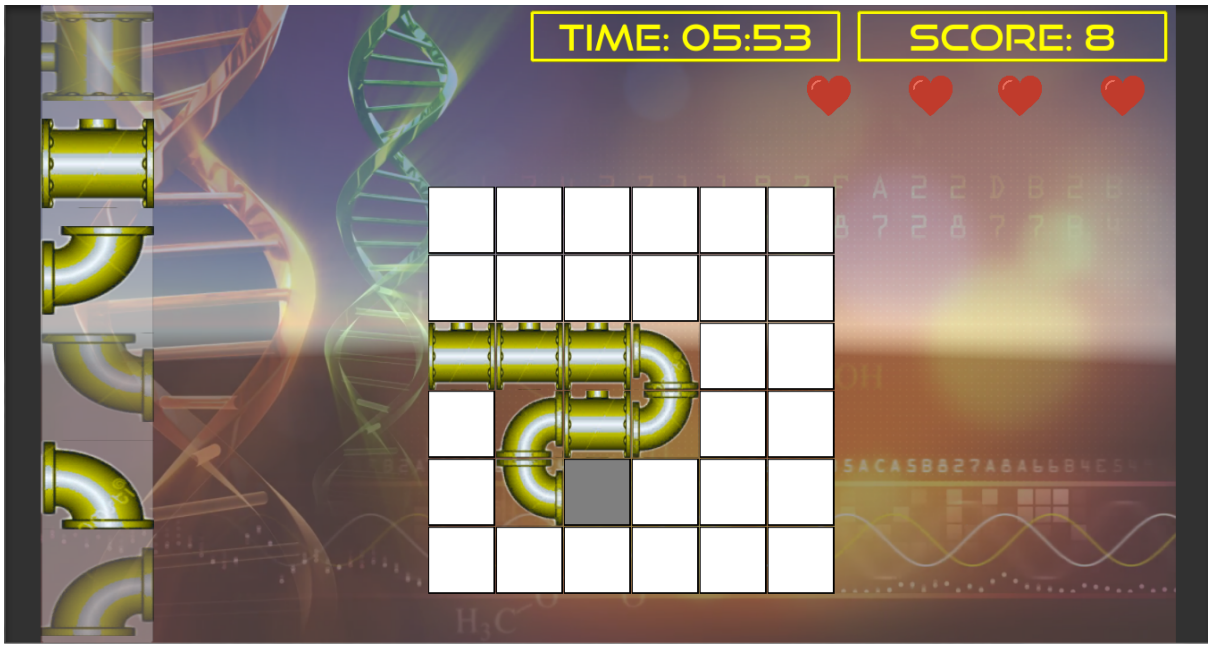
Each time, there is an empty tile in the grid and the player is prompted to fill it with a pipe part of his/her choice from the left menu, they will then be given an MCQ with 4 possible answers.

If the player chooses the correct answer, the pipe part he/she selected will be placed in the empty tile. If the player answers incorrectly, their incorrect answer will be excluded and they are given another chance to choose the correct answer. If the player answers incorrectly again, they are shown the question with the correct answer in green and prompted to use a Save-me to get another question to save themselves or a random pipe part will be placed instead of the pipe part they initially selected. However, the player has a limited number of 4 'Save-me's.

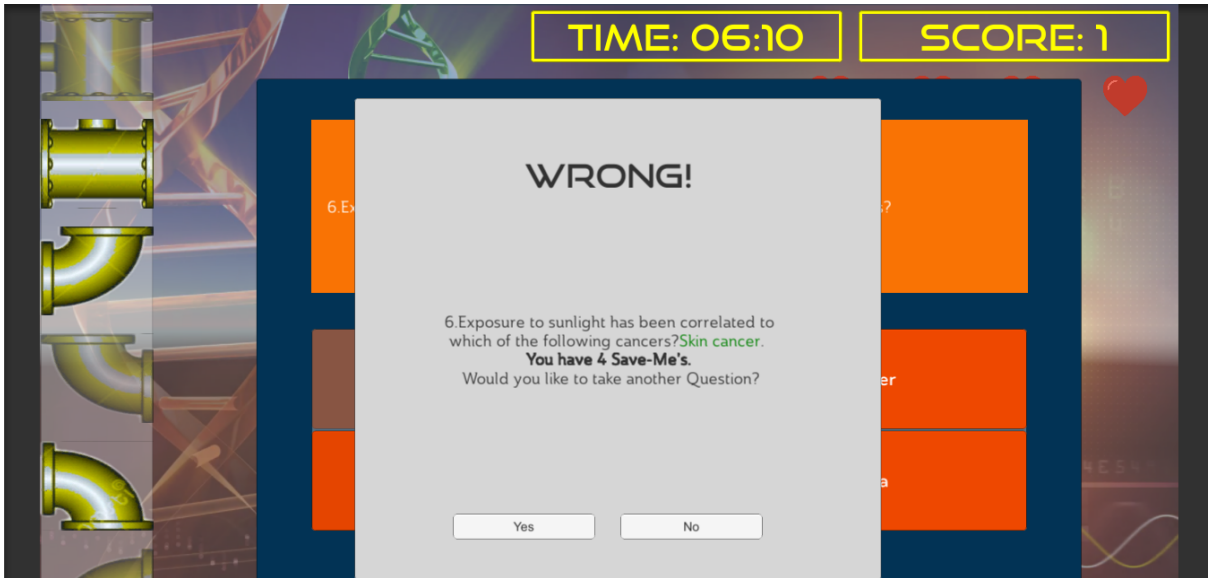
The aim of this is not to immediately penalize the player for an incorrect answer and to give them more time to think about the question and why their initial answer was incorrect, and to reconsider their tactics for answering the question correctly.

The player also gets points for each question that they answer correctly on the first try with an immediate score deduction for each question they answer wrong.

During this game, the CAT is most likely going to terminate and the user will be classified as an A, B or C student and would start receiving questions according to this classification. Hence, once the classification is determined, the user is notified that they are now "Exploring (A/B/C) Level questions" so as to equip them with the emotional readiness to accept levels of higher difficulties. Also, the choice of wording used in this feedback is so that we don't focus on the learner's individual identity but rather to reinforce their exploratory attitudes as advised by the experts (2.5).


TIME: 05:53
SCORE: 8


TIME: 06:50
SCORE: 0
23.A previously healthy 42-year-old man has a skin nodule on his right hand that has become larger and darker with more irregular outlines over the past 3 months. Three nontender enlarged lymph nodes are palpable in the right axilla. Which of the following risk factors is most important for development of this neoplasm?
Chronic exposure to ultraviolet radiation
Cigarette smoking
Allergy to latex gloves
Inheritance of a faulty RB gene


TIME: 06:10
SCORE: 1
WRONG!
6.Exposure to sunlight has been correlated to which of the following cancers?Skin cancer.
You have 4 Save-Me's.
Would you like to take another Question?
Yes
No

3.4.2 Locked Doors

The main objective of this game is to unlock as many doors as necessary to reach and unlock the golden goal door at the bottom right as shown in Figure 3.9.

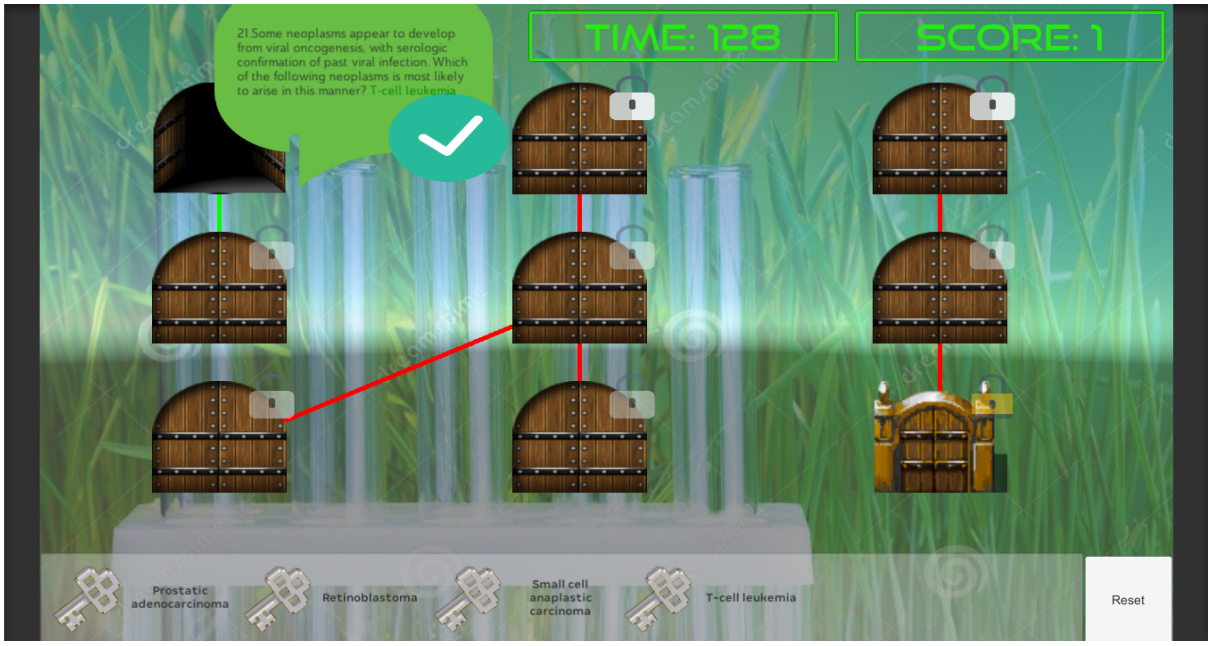
Each door is assigned a MCQ that pops up when the player hovers of the door. When the door is clicked, 4 possible keys appear in the key panel at the bottom, the player can then drag one of the keys to the door to unlock it. The 4 keys correspond to the 4 answer choices for the question.

To get a clear understanding of this game, it can be viewed as reducible to a puzzle with MCQs embedded in it. There is an apparent hierarchy in the way the doors are placed, such that some doors are *parents* to other doors below it and connected to it by red lines. The player simply must solve and unlock the parent door to be able to attempt to solve any of its children. Furthermore, there is a less obvious dependency set-up between the doors. Some doors offer rewards when unlocked which can be seen when the player clicks on the door's rewards drawer as shown in Figure 3.8. These rewards serve as keys to other doors somewhere in the puzzle. Therefore, when the player wants to solve-and-unlock a door whose keys are not yet "found", he must search through the puzzle to find the door holding its keys as rewards and solve it first, in addition to solving any parent door above it in the hierarchy. The doors whose keys are not yet *found* can be seen as closed doors with a dimmed lock icon attached to them, and once the player finds their keys (solves the door that holds the keys), the lock icon is made bright indicating that the player can solve them now.

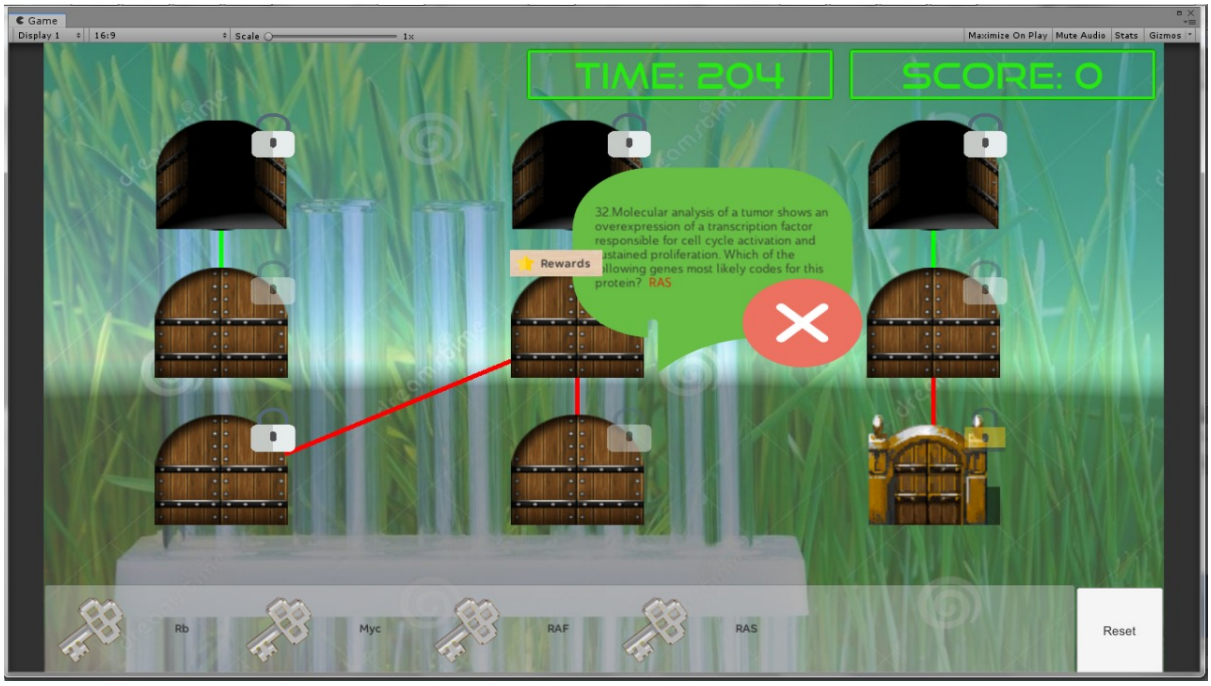
Accordingly, for each door that the player can solve:
If the player drags an incorrect answer key to the door, an incorrect answer bubble pops up giving feedback that the answer was wrong (Figure 3.7b). If the player drags the correct answer key to the door (Figure 3.7a), a correct answer bubble pops up, the door is opened, the path to its children doors is unlocked as the line between them turns green, and the other doors whose keys this door was holding in its reward drawer are freed and their attached lock icons undimmed.

The player then solves through the puzzle until they reach the golden goal door. Once they solve-and-unlock this door, the player wins the game.

It is worth noting that the CAL system operates differently for this game, as the game requires that 9 questions from the question pool be selected all at once to be assigned to the doors before the game can start, whereas, the CAL system is designed to render a question and receive its response before it can select another question. The CAL engine is therefore interrupted and a batch process is executed where the 9 currently most utilizable questions are all selected and assigned to the doors, and the CAL inference engine uses the responses to the 9 questions collectively in the calculation of the likelihood ratios, and the next question's difficulty will be based only on the last question the player answers in all of the 9 questions. Note that the next question means the question rendered after the whole game ends. Due to this undesirable interruption in the regular CAL operation, this game was omitted when conducting the experiment.



(a) Correct Answer



(b) Incorrect Answer

Figure 3.7: MCQ Puzzle.

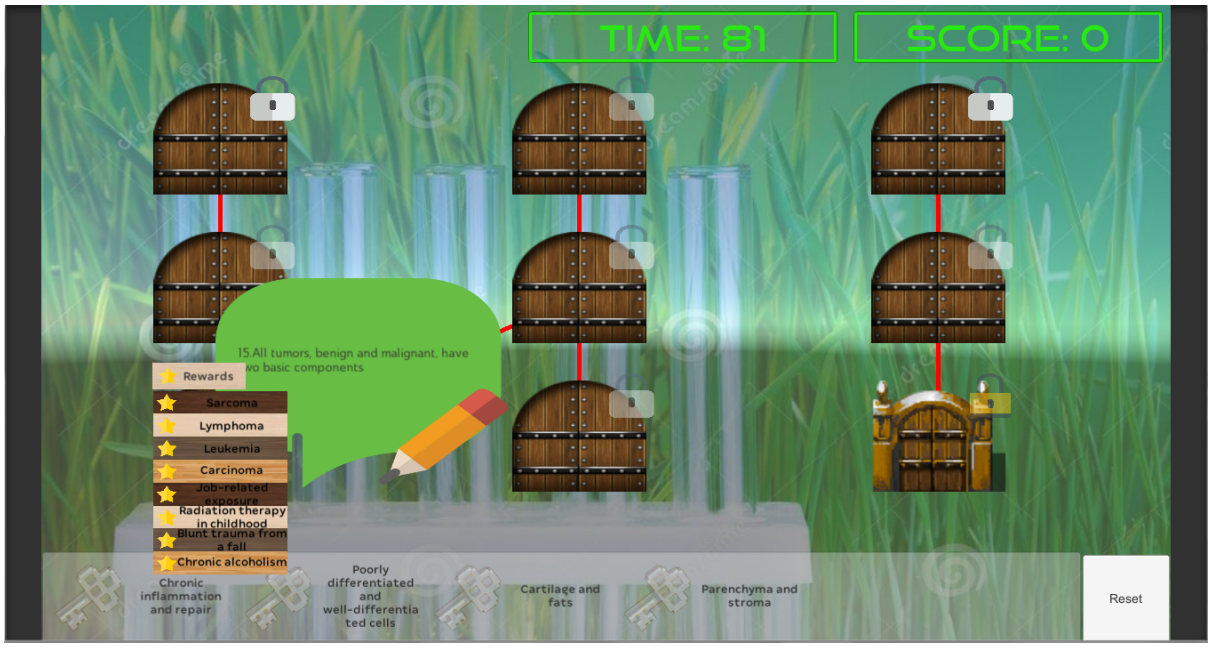


Figure 3.8: Rewards Drawer.

Figure 3.9: Goal Door.

3.4.3 Car

In this game the player gets the chance to review the questions that he/she answered incorrectly in any of the previous two games. The reason for choosing this game for that

purpose is that it requires the player to multitask (answer the questions while dodging the obstacles) so they don't have as much time to read the questions and are compelled to think more quickly. This game is also considered to be more fun and rewarding and could therefore be used as a break from slow cognitive work.

The objective of the game is to run the maximum amount of distance before the time ends, the distance at the end translates to score.

The player uses the arrow keys to move left and right. There are coins lying around that they can pick up. When they pick up a coin they are presented with an MCQ with 4 possible answers arranged up, down, left, and right (as in the illustration). The player uses the standard W, A, S, and D keys used for gaming movement to choose up, left, down, and right answers respectively.

If the player answers the question wrong, the wrong answer is colored red and the correct answer flashes green. If they answer the question correctly, only the correct answer will be colored green.

Additionally, when the player makes a correct answer, the car gains speed so that they are able to cover a larger distance to get a higher score. When the player makes an incorrect answer the car loses speed.

There are also obstacles along the way that the player must avoid. If the car runs into an obstacle it loses speed, however, it will not be stopped.

3.5 Emotional Adaptivity

The last feature to be discussed in our platform is the model for emotional recognition and affective adaptation. For this purpose, we compared the pros and cons of using subjective

measurements versus objective measurements as discussed in the literature review, and we decided to design a visual model for the self assessment manikin discussed in (2.7) for observing and interpreting the user's emotions. Moreover, we followed the adaptivity techniques suggested by [36] for implementing affective behavior that has shown to boost the user's engagement and learning gain.

3.5.1 Emotion Recognition: Visual SAM design

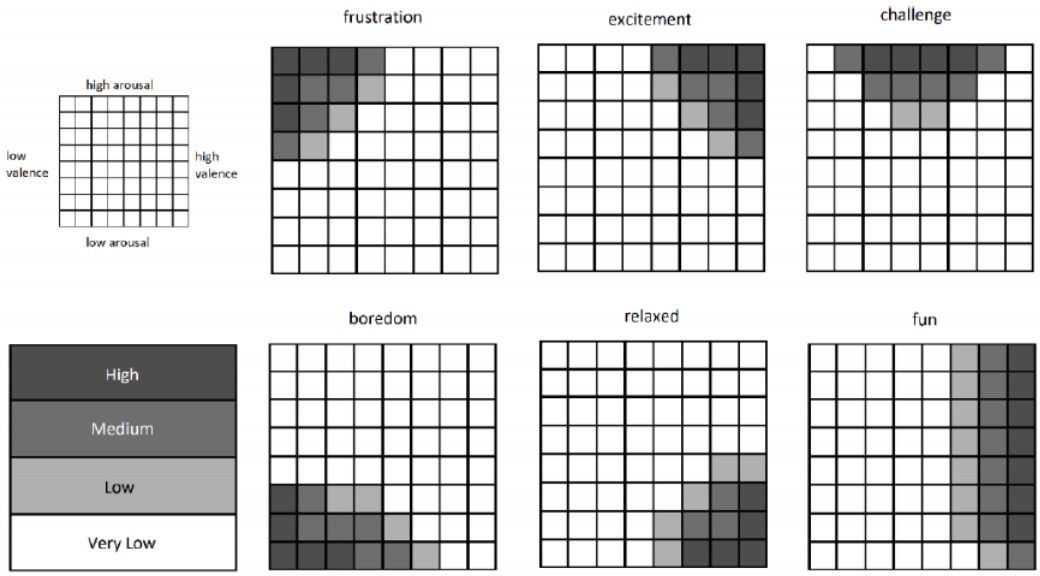
In our system, we need to detect the user's current emotion each time they finish one of the above games so we can adapt the interface and mechanics of the next game accordingly.

In comparing the different methods used for emotional measurement as discussed in 2.6, we found that objective measurements would not be suitable for our system since bio-sensors and other wearables, which are used to collect physiological data, would be obtrusive and would limit the user's freedom and range of motion when playing the game. Moreover, findings such as [7] showed that emotional measurement through physiological correlates indicate more the arousal level of an emotion (fear as well as anger) rather than detecting and isolating the targeted learning emotions.

Therefore, alternatively we reviewed a number of subjective measurement questionnaires that are designed to evaluate the affective state of the user (2.7). After reviewing the different questionnaires, we decided that the Self-Assessment Manikin (SAM) questionnaire would be the most suitable for frequent assessments during game-play as it is non-verbal, quick and easy to understand.

According to [6], the Self-Assessment Manikin is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli. It can also provide a simple and fast way to assess the emotional state depending on the valence and arousal levels only [36].

As discussed in 2.7, in a typical SAM the participant has to choose one of the five figures representing the valence scale and one of the five figures representing the arousal scale. The reported valence and arousal levels are then mapped to a specific emotion according to a scale designed by [11], for assessing emotions along the AV space (Figure 3.10).



space.PNG

Figure 3.10: Assessing emotions along the AV space.

Accordingly, a 5x5 grid was designed which maps each valence-arousal combination into one of the following emotions: Excitement, Fun, Relaxation, Challenge, Boredom, and Frustration.

However, a general limitation to the SAM model is that some users experience difficulty in interpreting what valence and arousal mean. For this purpose, we used the HSV color model to visualize the AV space for the user where the hue and value (brightness) are used to represent the valence and arousal dimensions respectively. This color model was chosen because it is defined in a way that is similar to how humans perceive color. Hue degrees are used to represent the valence levels, where green signifies positive valence and red signifies negative valence, and value (brightness of color) corresponds to the level of arousal of the different valence levels as shown in figure 3.11 .

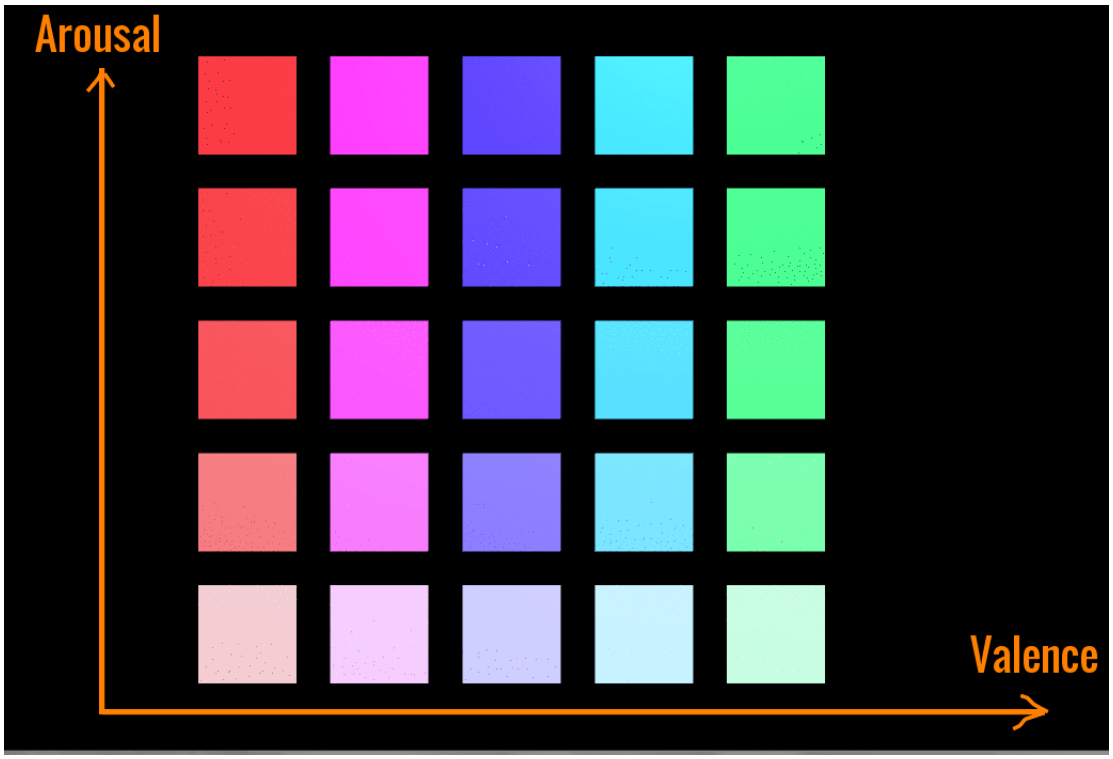
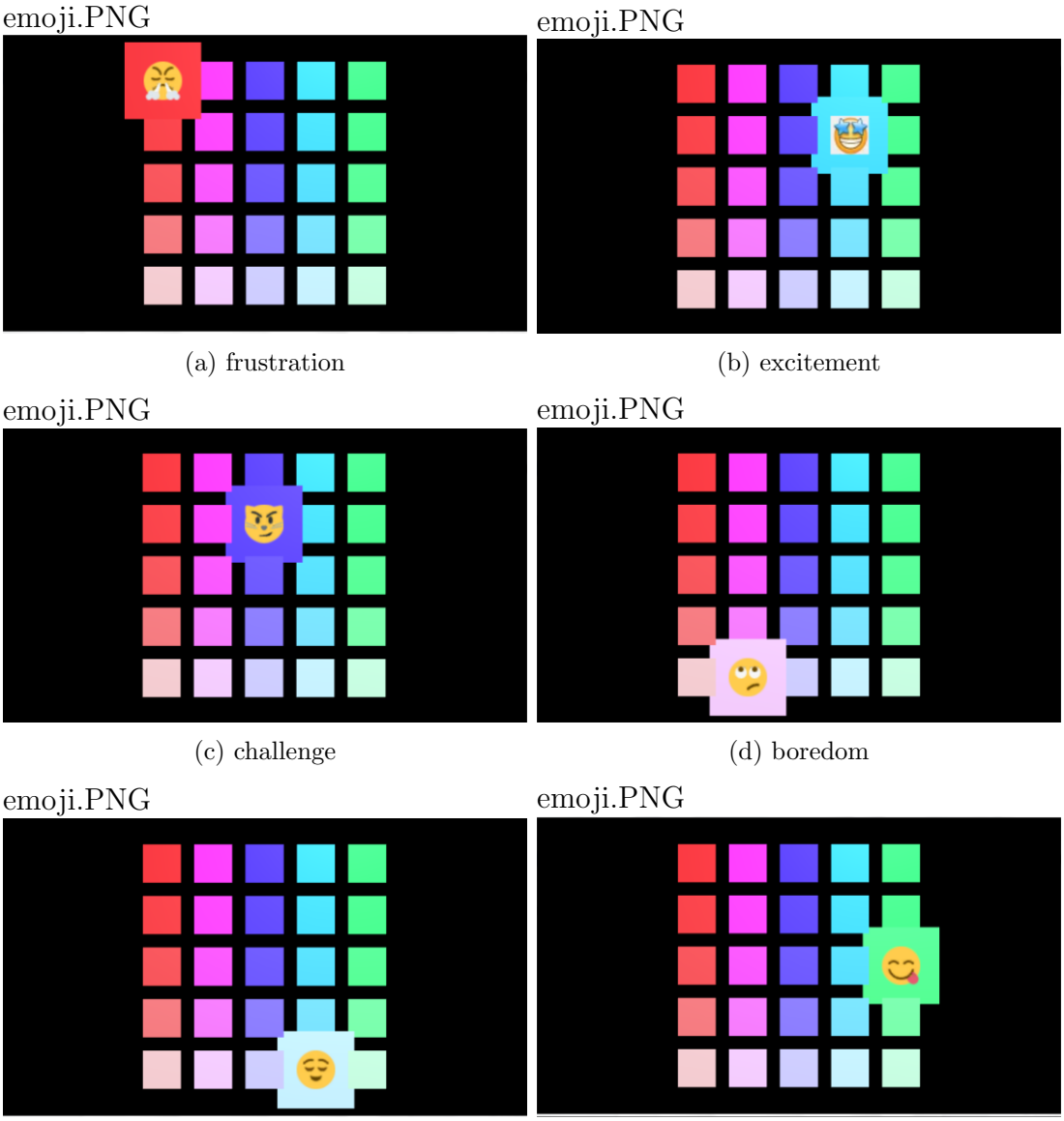


Figure 3.11: Valence Arousal grid using the HSV color model.

Finally, emoticons are allocated to each tile in the 5x5 grid to visualize the emotion that this tile expresses. The decision to add emoticons was based on the fact that people frequently use emoticons in their everyday lives and are thus acquainted with which emotions they indicate. This ultimately reduces the need for importunate explanations of the model as well as the probability of errors in the user self-assessing their emotions.

Hence, when the user hovers over a tile in the valence-arousal grid, the emoticon of the corresponding emotion pops up and they can click on it if this is the emotion they're feeling. Then, the adaptivity techniques are applied according to that recognized emotion and the changes appear directly in the next game (if the recognized emotion requires any changes) as will be discussed in the next section.

Figure 3.12: Expressing emotions along the AV space using emoticons.



(a) frustration (b) excitement

(c) challenge (d) boredom

(e) relaxed (f) fun

3.5.2 Adaptivity Techniques

For the purpose of the experiment, we followed the design proposed by [36] which has previously shown to increase the engagement level of the student and used it to implement the affective adaptivity features which would be compared to the adaptive difficulty feature we designed in 3.2.

According to [36], the main aim of the affective adaptations is to avoid feelings of boredom, frustration and unchallenged, which correspond to 3 of the 6 emotions detected by the model in 3.5.1. Hence, the system changes its metrics only if one of these emotions

is detected, otherwise, the system continues without any changes as it is desirable to keep the student in one of the other 3 states (Excitement, fun or challenged). Moreover, it has been decided to limit the adapting metrics to one of the following four: Timer, Scoring method, Music and Theme color.

Accordingly, the following changes happen if one of the undesirable emotions is detected in any of the games:

- **Timer:** For each game there is the option to choose between a maximum time limit (eg. 7 minutes) or a minimum time limit (eg. 4 minutes). If the user reports feeling bored or relaxed the minimum time limit is chosen to increase the challenge level. On the other hand, if the user reports feeling frustrated, then the maximum time limit is chosen to reduce feelings of agitation. Note that if the timer in the game is a count-up timer, the time limit is the time under which the user must finish the game to get a bonus, otherwise, they lose the bonus. On the other hand, if the timer is a count-down timer, the time limit is simply the time after which the game will end.

- **Scoring method:** In each game, there is a minimum and maximum deduction penalty per incorrect answer. If the user reports feeling bored or relaxed, then the maximum deduction penalty is enforced. If they report feelings of frustration, the minimum deduction penalty is applied.

- **Music:** For each game, there is a default theme music playing in the background. This music varies between the default music track, a more active one if the user feels bored and a relaxing one if they feel frustrated.

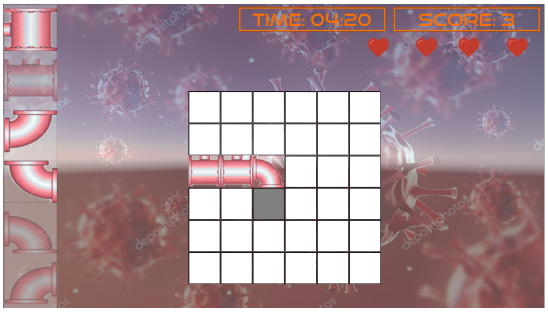
- **Theme Color:** For each game, there are several features that change color according to the user's emotional state. For example, in the pipes game the color of the pipe parts, the background wallpaper and the MCQ panel varies between red in case the user feels bored, green in case the user feels frustrated and orange in case the user feels relaxed. The aim of this is to use high arousal colors (red and orange) to avoid negative low arousal emotions (boredom and relaxation), and in contrast, low arousal colors (green) to avoid negative emotions of high arousal (frustration). Example of the different color schemas is shown for each game in figure 3.13.

The mapping of these changes to the detected emotions is summarized in the following table:

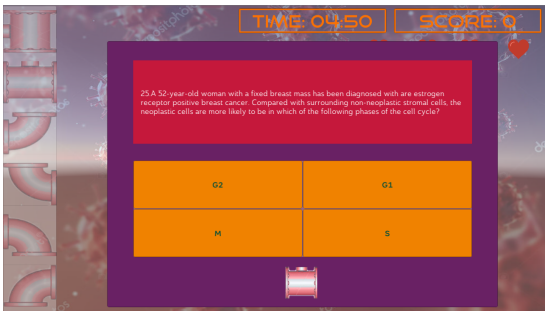
Emotion	**Timer**	**Scoring Method**	**Music**	**Theme Color**
Boredom	Min. Time Limit	Max. Deduction Penalty	Active Music	Red
Frustration	Max. Time Limit	Min. Deduction Penalty	Relaxing Music	Green
Relaxed	Min. Time Limit	Max. Deduction Penalty	Default Game Music	Orange

Table 3.1: Mapping metrics changes to the recognized emotion.

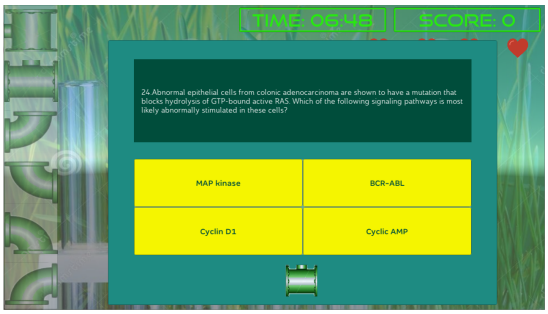
Figure 3.13: Color scheme variations.



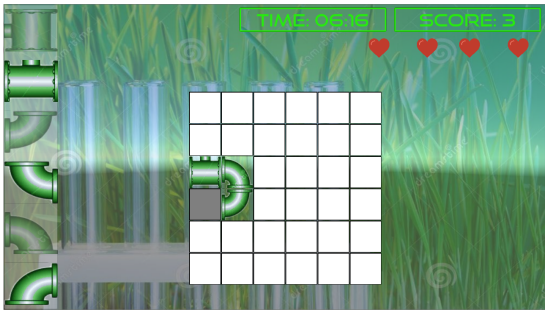
(a) Pipes: Red Theme Color (a)



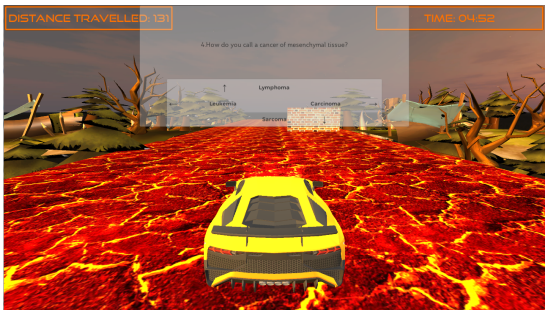
(b) Pipes: Red Theme Color (b)



(c) Pipes: Green Theme Color (a)



(d) Pipes: Green Theme Color (b)



(e) Car: Red Theme Color

(f) Car: Green Theme Color (b)

Chapter 4

Testing and Experimental Design

4.1 Question Pool

To test the system, we asked an educator from the field of Pharmacy to provide us with an MCQ question pool that tests an instructional topic from their Pathology course. We chose the Pathology course because the students' final exam was in the form of a computerized MCQ exam. We further instructed the educator to design and divide the question pool into 3 levels, the first level consists of 13 very easy questions, the second level consists of 13 questions of medium difficulty, and the third level consists of 20 very difficult questions.

4.2 Calibration

4.2.1 Procedure

After designing the question pool, we invited some of the students that were currently taking the Pathology course to solve the full question pool in the form of a paper-and-pencil exam (46 MCQs). The students were selected such that one third of the students had a high GPA, one third were of an average GPA and one third were below average. In addition, some of the students selected had attended a recent quiz and so were expected to have studied the instructional topic, while some of the students hadn't yet attended any quizzes for the topic. This selection criteria was set in the hopes of diversifying the competencies in the estimation sample.

The students, who attended, were first told that the aim of this experimental exam was to collect data to identify difficult questions within the pool with the future goal of implementing a training system with adaptive difficulty.

For each question, there were the standard 4 answer choices as well as an extra check box with the phrase "I don't know". The students were accordingly instructed not to

guess the answer to any of the questions and to tick the "I don't know" choice box in case they did not know the answer.

Finally, the exam papers were collected, corrected and scored, and the A and B cut-off scores were chosen to divide the students into A, B and C students.

4.2.2 Results and Insights

The students who attended the calibration exam were a total of 23 students with varying test score results. Statistics for the test score results were then calculated as shown in the following table:

Average	**Minimum**	**Maximum**	**Median**	**1^{st} Quartile**	**2^{nd} Quartile**	**3^{rd} Quartile**
61.05860113	21.73913043	91.30434783	63.04347826	45.65217391	63.04347826	75

Table 4.1: Test Score Results Statistics

Thereby, the A cut-off score was established to be 70% and the B cut-off score was established to be 56.5%. The A cut-off was chosen to be between the 2^{nd} and 3^{rd} quartiles to mark students who are of a very high level in training but still need training as discussed in 3.3.3. The B cut-off was likewise chosen to be between the 1^{st} and 2^{nd} quartiles. As a result, 7 (30.4%) of the students in the estimation sample were categorized as C-Level students, 9 (39.1%) of the students in the estimation sample were categorized as B-Level students and the remaining 7 students (30.4%) were categorized as A-Level students -i.e. even distribution among the 3 grade-categories.

The questions and student responses were then installed into a spread-sheet in Microsoft Office Excel and the probability rules and difficulty, discrimination, and utility parameters for each question were calculated as described in 3.2.1.

This calibration yielded a lot of additional information about the items in the question pool. For example, one of the questions that the expert (educator) had put in the easy level questions showed that only 44% of the students answered it correctly, which means that this was actually one of the difficult questions for the students. Another question that the expert had classified into the hard level questions showed that 88% of the students answered it correctly, which means that this was actually an easy question for the students. This proves that this calibration system was indeed successful in complementing the expert's knowledge about the questions' difficulties.

However, some of the questions' discrimination parameters were of a negative value, meaning that for example more C students answered a specific question correctly than B students. This indicates possible limitations which distorted the calibration results.

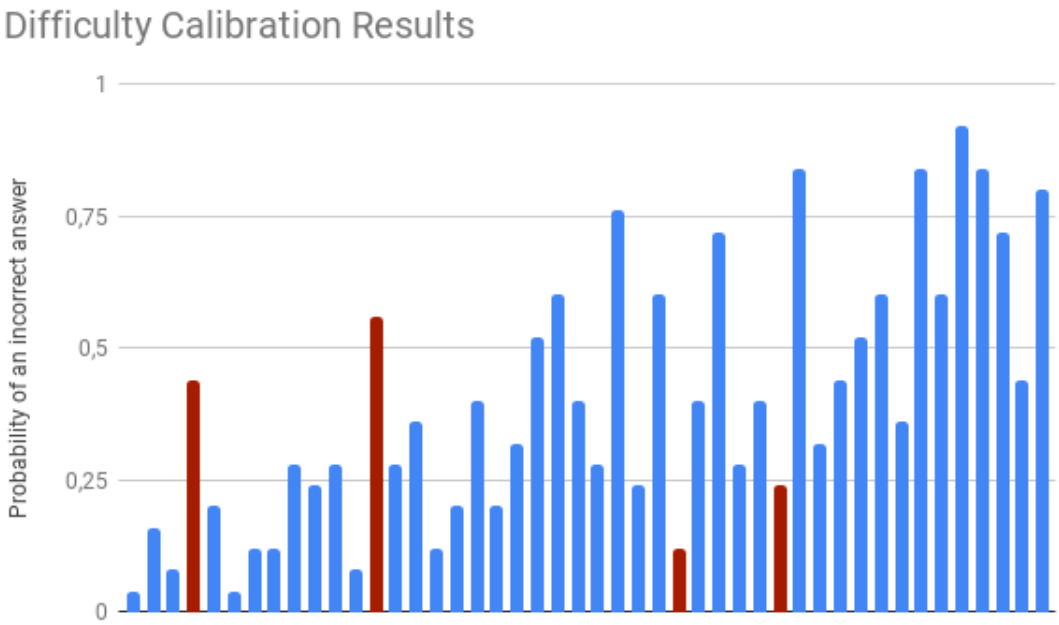


Figure 4.1: Difficulty Calibration Results (Red bars = unexpected results)

4.2.3 Further Testing and Decisions

Moreover, simulations of the CAT test described in 3.2 were made for each student in the estimation sample by using their responses in the paper-and-pencil exam. These simulation studies were used to test the CAT system and to make decisions about the values of the termination parameters such as the error probabilities (2.4) and the maximum and minimum test length constraints.

When running simulations of the CAT test on the calibration students responses, we found that the test takes from 4 to 16 items to be able to classify the student into A, B, or C category. This test length variation is expected. For example, an examinee who consistently answers questions correctly will be classified as an A student within a small number of questions, likewise, an examinee who consistently answers questions wrongly will be classified as a C student within a small number of questions. However, for an examinee to be classified as a B student (an "in-between" category), more test items will be needed as their answers will vary between correct and incorrect.

Moreover, when we ran the test simulation on a student who achieved an overall score of 69.6% (approximately equal to the A cut-off score), we found that the CAT fails to categorize him/her as an A or B student and the entire item bank could be administered with the test remaining inconclusive. For such cases where the student is an "edge student" (on the verge of being classified into any one of two categories), a maximum test length constraint has to be imposed to ensure that the question pool is not depleted before proceeding to phase II (3.2.2).

Therefore, a minimum and maximum test length constraints were set as previously discussed in (ii). The minimum TL is set to 7 items to ensure that the test is *slow enough* to make precise decisions, and the maximum TL is set to 10 items, after which the CAT resorts to a forced classification in an effort to reserve some items and save enough time for phase II.

4.3 Two Versions

The experiment is comparative in nature, therefore we created two separate versions of the system described in the methodology chapter, one with the adaptive difficulty feature, and one with the emotional adaptivity feature.

The first version uses the calibrated question pool and operates the CAL system, however, the game mechanics and user interface discussed in 3.5 were static and the SAM questionnaire does not appear after every game.
The second version, on the other hand, operates a different selection algorithm (which we will describe shortly), asks the user to report their emotions after every game using the SAM questionnaire, and adapts the game mechanics and UI according to the recognized emotion.

For the second version, we created another strong randomized selection algorithm to be able to compare it to the CAL system. It selects a total of 21 questions (less than half the question pool) from the 3 difficulty levels designed by the educator. It renders 7 questions randomly selected from the easy difficulty level, followed by 7 questions randomly selected from the medium difficulty level, and then 7 questions randomly selected from the high difficulty level.

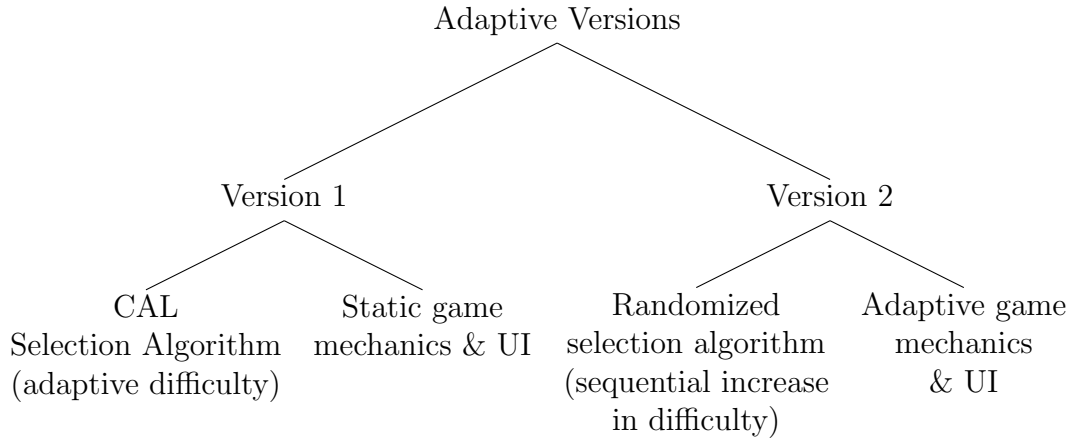
In this way the randomized selection algorithm has no bias towards any one level, operates in the conventional manner of increasing the difficulty sequentially (non-adaptive), and within every level all the questions have an equal chance of being rendered. The important differences are that it lacks the CAL system's intelligence attained through the calibration, it doesn't skip questions the way the CAL system does if the user is performing well, and it is indifferent to the users skill level so it can render questions that are too easy or too difficult compared to the user's ability.

Given these points, we made the first version also render 21 questions, however, depending on the CAL selection algorithms. In this way, the two systems render an equal amount of questions from the same pool, which makes it easier to compare their efficiencies.

Finally, with all things considered, we construed both systems to be as alike as possible where the adaptivity features are not affected. That is, only 2 of the 3 games are played in the same way (the locked-doors is omitted to avoid the batching process within the experiment 3.4.2), the questions are administered in the same way, the feedback system is unchanged, the incorrectly answered questions are gathered and channelled to the car game for reviewing, and the user repetitively plays the 2 games back to back until they finish answering the 21 questions, after which, the session ends abruptly.

4.4 Test Planning

The main objective of the experiment is to compare the two versions of the system described in the following block diagram:



4.4.1 Testing parameters

To compare these two versions, we first define certain parameters against which the systems are tested and their results compared.

The CAL selection algorithm in *version 1* is compared to the randomized selection algorithm in *version 2* in terms of two parameters:

- **Learning gain:** The amount of knowledge gained after answering 21 questions.
- **Exposure efficiency:** The amount of beneficial questions to which the user is exposed within the 21 rendered questions.

The learning gain describes what information the student actually remembers from playing the game and how much this information improved his overall academic level. On the other hand, exposure efficiency describes the selection algorithm's efficiency in exposing the user to questions they hadn't already known, with disregard to whether or not the user actually remembers these questions.

Finally, the CAL's adaptive difficulty feature in *version 1* is then compared against the emotional adaptivity feature in *version 2* in terms of a third parameter:

- **Engagement:** How engaged and immersed the user felt during gameplay.

An extensive explanation of these parameters, the ideas behind them, and how they're calculated is explained in the following section.

4.4.2 Hypotheses

With the performed work in mind, more definitive hypotheses can be fleshed out:

- **Hypothesis 1 (H1):** There is no significant difference between the two versions of the system in terms of the learning gain.
- **Hypothesis 2 (H2):** There is no significant difference between the two selection algorithms in the two versions in terms of exposure efficiency.
- **Hypothesis 3 (H3):** There is no significant difference between the two adaptivity features in terms of the level of engagement the user experiences.

4.5 Test conduction

The participants in the experiment were evenly divided into two groups, one group played the adaptive difficulty version of the game and the other group played the adaptive mechanics and UI version of the game, and the two groups were subject to the same 3 tests that aim to compare the two versions. In this section we will describe these tests and the reasons behind them.

4.5.1 Learning Gain

The idea behind this test is to assess the academic level of the participants in the given Pathology topic before and after using the adaptive difficulty version of the game and the adaptive gaming mechanics and UI version of the game.

The test consists of a pre-test and a post-test, which are identical and are comprised of the entire question pool of 46 MCQ questions. The pre-test is taken before playing the game and the post-test is taken after playing the game, and the decision to make them identical was so as to ensure that they were at the same level of difficulty and structure.

The raw score for each test is calculated for each student as a percentage of the total number of questions which the student answered correctly.

By comparing the results of the pre and post tests we can assess the learning gain of the students, and conclude whether each version has affected the student's learning achievement level or not.

Afterwards, the learning gain of each group is calculated separately as the difference between the students' scores in the post-test and their scores in the pre-test. Then, the learning gain of both groups is compared to see which group achieved better academic results.

This test was used to prove or disprove the first hypothesis (H1).

4.5.2 Exposure Efficiency

The aim of this test is to define a calculative measure to compare the efficiency of the two selection algorithms of the two systems.

The question pool in the pre and post tests is the same pool from which the two systems would select the questions to be administered to the user. The main difference is that in the pre and post test the user is instructed to answer the entire question pool (46 questions), while in the game they're only exposed to 21 questions (less than half the question pool).

The two selection algorithms are then evaluated according to how many questions they were able to successfully target and render out of all the questions that the student answered incorrectly in their pre-test within the game's "21 Questions Period".

This is then achieved by marking the questions that the student answered incorrectly on the pre-test. Then, we log all the questions that are rendered to the user during the game. Finally, we match and mark the questions rendered to the user which they had answered incorrectly in the pre-test, and calculate the exposure efficiency according to the following equation:

$$Exposure\ Efficiency = \frac{w_e}{w} \tag{4.1}$$

Where,
w: is the number of questions the student answered incorrectly on the pre-test.
w_e: is the number of questions the student answered incorrectly on the pre-test and was then exposed to in the game.

Afterwards, the exposure efficiency of both groups is compared to see which group experienced better exposure.

This test was used to prove or disprove the second hypothesis (H2).

4.5.3 Engagement

Many studies nowadays are stressing on the fact that using the motivating aspects, in general, can be used to facilitate learning for students. Therefore, one of the main goals of this study is to compare the effect of adapting the difficulty of the educational content according to the user's performance versus the effect of adapting the game mechanics and UI according to the user's emotions on the student's engagement during his learning experience.

Thus, a test that aims at measuring the engagement level of the students being exposed to the two different adaptive versions of the same game was held through using a Likert Scale survey that measures their overall experience. The Likert Scale is a five-point scale that is used to rate the agreement (or disagreement) of the participant with a particular statement, or question.

This approach is widely used in subjective research topics (especially usability related studies) [11].

The Likert Scale survey used in our study is a subset of a standardized questionnaire that is generally used to test the engagement level or the overall state that the user experiences (Engagement paper). It is composed of 9 items (listed in Appendix A) that measure the overall engagement level in any activity through measuring the control and the enjoyment levels.

Each participant was asked to rate their agreement with each statement in the questionnaire on a scale from 1 to 5, where 1 means "Strongly agree" and 5 means "Strongly disagree", the mean of their 9 ratings was then calculated to quantify their engagement level.

Finally, the engagement level of both groups was compared to see if one adaptive version was more engaging than the other.

This test was used to prove or disprove the third hypothesis (H3).

4.6 Participants

The participant group consisted of 50 students that were currently enrolled in the Pathology course. The age range was from 20 to 22. There was 1 student who indicated that she had no familiarity with playing computer games, while the rest stated they had no problem with playing a computer game. The participants were randomly and evenly divided in each of the two conditions so that 25 persons participated in each of them.

4.7 Procedure

Having taken place at a computer lab at the University, it was ensured that all the PCs for the experiment were identical and were configured to the same resolution.

The experiment was scheduled on 3 different days, with a week between the first and second session, and the second and third session occurring on two consecutive days.

The timing of the experiment was chosen so that it took place within 2 weeks before the students' Pathology final examination in the hope that some students would have studied a little for their examination. This would reflect in some diversification in the experiments' results as students of different abilities would profit differently from the system's efficiency factor.

During any of the 3 sessions, the participants would be randomly and evenly divided between the two versions of the system. Hereafter, they took part in determining their pre-game level of proficiency in the instructional topic by answering the paper-and-pencil pre-test. Then, they were directed to read the instructions of the game, in which they were

informed about the goal of the whole game as well as the specific instructions for playing each of the 2 games (Pipes and Car). Nothing was revealed to them about the condition they took part in, and the students playing the emotional adaptivity version of the game were given extra instructions about the definitions of the valence and arousal dimensions. When they felt they were ready for the task, we had them enter their student ID in the game to later be able to identify their respective logged data, after which, they could commence the game. Playing the game from start to finish took each participant at most 30 minutes and during this time, data was gathered about the questions being rendered to the student and saved in a text file along with their registered student ID. Directly after the participants finished playing the game, they were again given their paper-and-pencil test which they answered before playing, and they were asked to correct any answers they now knew were wrong with a different colored pen to account for the post-test. The reason for not making them re-answer the exam from scratch was because the students did not show keenness towards re-answering the lengthy exam as the questions were bulky and quite comprehensive. Afterwards, they were asked to fill out the engagement questionnaire. Finally, the participants were thanked for their cooperation.

Chapter 5

Results

5.1 Learning Gain Test Results

This section describes the findings behind holding the learning gain tests (the paper based pre and post tests) that both groups have been exposed to. The design of this test is mentioned in Section 4.5.1.

5.1.1 CAL Group

The results of the tests of the group of students who have played the CAL version of the game revealed that playing this version improved the students' scores in the Pathology MCQs of the pos-test comparing to the pre-test. The students had an average pre-test score of 52.26% (n=25, M=52.26, SD=10.2965), and an average post-test score of 76.4% (n=25, M=76.4, SD=9.609). This has achieved a significant learning gain by an average of 24.14% (n=25, M =24.14, SD= 6.8851). The results of the tests of this group are present in the following table:

CAL Group Results		
	Mean	**Std. Deviation**
Pre-test	52.260	10.2965
Post-test	76.400	9.9609
Gain	24.1400	6.8851

Table 5.1: CAL Group: Learning Gain Test Results

5.1.2 Emotional Adaptivity Group

The results of the tests of the group of students who have played the version of the game, in which the game mechanics and UI adapt to the user's emotions, revealed that

playing this version improved the students' scores in the Pathology MCQs of the pos-test comparing to the pre-test. The students had an average pre-test score of 52.4% (n=25, M=52.436, SD=13.4323), and an average post-test score of 70% (n=25, M=69.9736, SD=11.66764). This has achieved a significant learning gain by an average of 17.5% (n=25, M =17.53760, SD= 7.48787). The results of the tests of this group are present in the following table:

Emotional Adaptivity Group Results		
	Mean	**Std. Deviation**
Pre-test	52.436	13.4323
Post-test	69.9736	11.66764
Gain	17.53760	7.48787

Table 5.2: Emotional Adaptivity Group: Learning Gain Test Results

5.1.3 Independant t-Test Results

The analysis test type that was held for the learning gain test is the well known independent t-test. The aim of this test is to compare between the two versions tested in the experiment in terms of the amount of knowledge gained by the students.

Therefore, an independent t-test was run on the data as well as 95 percent confidence intervals (CI) for the mean difference.

By applying the treatment, it was found that after the two interventions, the learning gain resulting from the group that used the CAL version of the game (M = 24.14, SD= 6.885129) was significantly higher than the gain of the other group that used the Emotional adaptivity version of the game (M= 17.5376, SD=7.487869) (t(25) = 3.245323, p = 0.002141) with a mean difference of 6.602400 (95% CI, 2.511894 to 10.692906).

This rejects (H1) stating that there is no difference between the two versions of the system on the knowledge gain of the students.

Learning Gain Test of Both Groups			
	Mean	**Std. Deviation**	**Std. Error Mean**
CAL Group	24.140000	6.885129	1.377026
Emotional Adaptivity Group	17.537600	7.487869	1.497574

Table 5.3: Independent t-test results of the learning gain

Independent Sample Test Results (Learning Gain)						
	t	**Signif. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% confidence interval**	
					lower	higher
Difference	3.245323	0.002141	6.602400	2.034435	2.511894	10.692906

Table 5.4: Independent t-test results of the learning gain

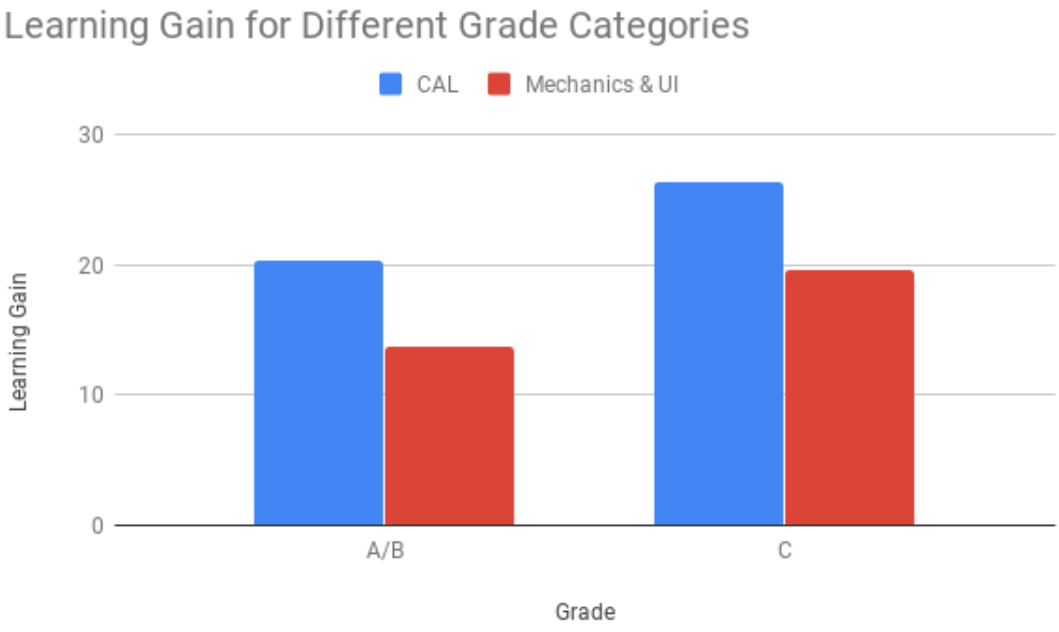


Figure 5.1: Learning Gain for Different Grade Categories

5.2 Exposure Efficiency Test Results

This section describes the findings behind holding the exposure efficiency tests that both groups have been exposed to. The design of this test is mentioned in Section 4.5.2.

5.2.1 Independant t-Test Results

The analysis test type that was held for the exposure efficiency test is the well known independent t-test. The aim of this test is to compare between the two selection algorithms implemented in the two versions in terms of how efficient they are in targeting the questions that the student did not know.

Therefore, an independent t-test was run on the data as well as 95 percent confidence intervals (CI) for the mean difference.

By applying the treatment, it was found that after the two interventions, the exposure efficiency resulting from the group that used the CAL version of the game (M = 0.493899, SD= 0.091873) was significantly higher than the exposure efficiency of the other group that used the Emotional adaptivity version of the game (M= 0.410766, SD=0.065600) ($t(25) = 3.682069$, $p = 0.000586$) with a mean difference of 0.083133 (95% CI, 0.037737 to 0.128529).

This is an indication that the efficiency definition from section 4.4.1 would probably not yield different results in terms of the efficiency of the CAL selection algorithm versus that of the randomized selection algorithm implemented in the emotional adaptivity version.
This in turn rejects (H2) stating that there is no difference between the two selection algorithms in the two versions in terms of exposure efficiency.

Exposure Efficiency Test of Both Groups			
	Mean	**Std. Deviation**	**Std. Error Mean**
CAL Group	0.493899	0.091873	0.018375
Emotional Adaptivity Group	0.410766	0.065600	0.013120

Table 5.5: Independent t-test results of the exposure efficiency

Independent Sample Test Results (Exposure Efficiency)						
	t	**Signif. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% confidence interval**	
					lower	higher
Difference	3.682069	0.000586	0.083133	0.022578	0.037737	0.013120

Table 5.6: Independent t-test results of the exposure efficiency

5.2.2 Independant t-Test Results: A&B Students

An important point to consider when analyzing the results is that the CAL system's adaptive selection algorithm is mostly beneficial for A and B students as they are the students who have prior information about the subject and thus would benefit more from a system that quickly addresses their knowledge gaps. On the other hand, C students would benefit from any kind of training (efficient or inefficient). Therefore, we set apart the students from each of the sample groups who achieved a score greater than 56.6% (the B cut-off score), and an independent t-test was run on their data alone which yielded the following results:

Exposure Efficiency Test of Both Groups (A and B students)			
	Mean	**Std. Deviation**	**Std. Error Mean**
CAL Group	0.553920	0.093514	0.031171
Emotional Adaptivity Group	0.401486	0.081175	0.027058

Table 5.7: Independent t-test results of the exposure efficiency (A and B Students)

Independent Sample Test Results (Exposure Efficiency A-B Students)						
	t	**Signif. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% confidence interval**	
					lower	higher
Difference	3.692934	0.001972	0.152434	0.041277	0.064930	0.239937

Table 5.8: Independent t-test results of the exposure efficiency (A and B Students)

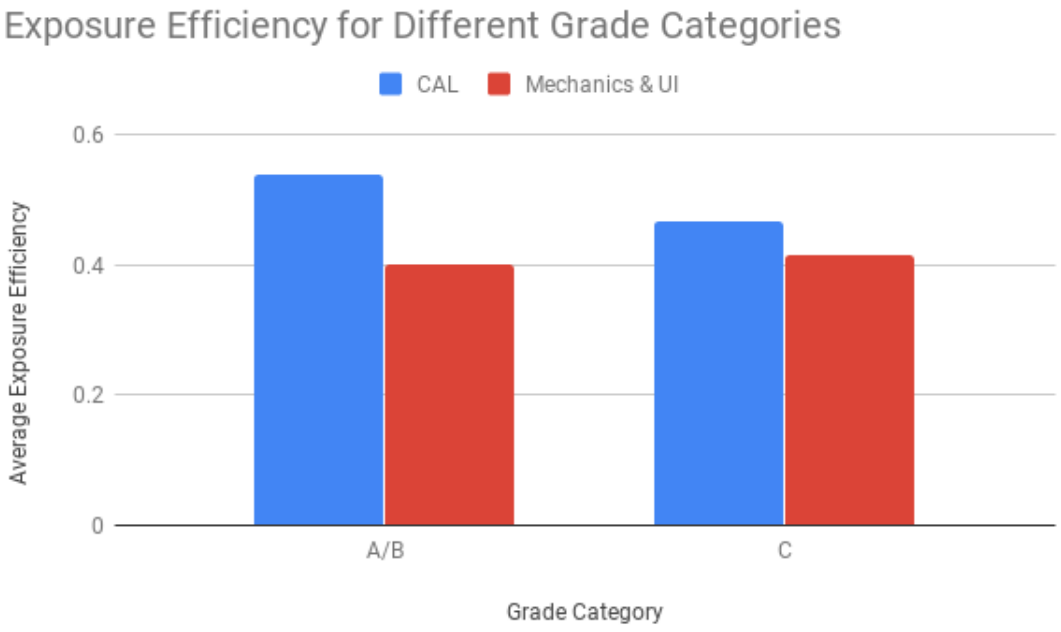


Figure 5.2: Exposure Efficiency for Different Grade Categories

5.2.3 Independant t-Test Results: C Students

Another independent t-Test was run on the data for the students who were classified as C students from their pre-test scores, however, it showed that the exposure efficiency resulting from the group that used the CAL version of the game (M = 0.460138, SD= 0.073918) was not significantly higher than the exposure efficiency of the other group that used the Emotional adaptivity version of the game (M= 0.415986, SD=0.057361) (t(16) = 1.887562, p = 0.068793) with a mean difference of 0.044152, which proves the point that the CAL selection algorithm mostly benefits A and B students, while C students can benefit from any type of training.

However, it is also worth noting that there was a slight difference in the mean exposure efficiency of both groups (0.044152) with the exposure efficiency of the CAL group being marginally higher than the emotional adaptivity group. This is likely due to the intelligence advantage of the CAL system which was achieved through calibration.

5.3 Engagement Level Test Results

Concerning the engagement level measurement held at the end of the experiment, this test was held to evaluate the overall involvement state of the students being exposed to both the CAL adaptive difficulty version and the emotional adaptivity version. It compares between the results of the two adaptivity features in order to determine which has achieved a better engagement level for the participants (Section 4.5.3).

The results of the independent t-test between the two groups revealed that the engagement level ratings that represent how absorbed the students were while using the version with the adaptive game mechanics and UI as the adaptivity feature (M = 2.1667, SD= 0.4507) were not significantly higher than the engagement level ratings of the other group that used the version with the adaptive difficulty as the adaptivity feature (M= 2.0505, SD=0.5652) (t(22) = 0.75368, p = 0.4552) with a mean difference of 0.116 (95% CI, -0.1949 to 0.4272).

This proves (H3) stating that there is no significant difference between the two adaptivity features in terms of the level of engagement experienced by the user.

Engagement Level Test of Both Groups			
	Mean	**Std. Deviation**	**Std. Error Mean**
Emotional Adaptivity Group	2.166667	0.450684	0.096086
CAL Group	2.050505	0.565236	0.120509

Table 5.9: Independent t-test results of the engagement level

Independent Sample Test Results (Engagement Level)						
	t	**Signif. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% confidence interval** lower	higher
Difference	0.753679	0.455247	0.116162	0.154126	-0.194878	0.427201

Table 5.10: Independent t-test results of the engagement level

5.4 Discussion

By gathering and analyzing the test results of the experiment. The results show that there exists statistically significant differences in the learning process of the students during the course of playing an educational game when two different adaptivity features are applied: (1) adapting the difficulty of the questions according to the student's performance using the CAL methodology, and (2) adapting the game mechanics and UI according to the student's emotional state.

Regarding the learning gain, the independent t-test results show the existence of a significant difference between the two adaptivity groups (p = 0.002). The participants

who played the game with the adaptive difficulty feature showed a better improvement in their test results than those who played the game with the adaptive game mechanics and UI.
Accordingly, (H1) is rejected which claims that there is no difference between the two versions of the system on the learning gain of the students.

Regarding the exposure efficiency, the independent t-test results show that there is a significant difference between the two selection algorithms implemented in the two version of the system: (1) the CAL selection algorithm, and (2) the randomized selection algorithm (p 0.001). The participants who played the game with the CAL selection algorithm were better exposed to questions they had not already known than those who played the game with the other selection algorithm, within the same designated 21 questions time period. Accordingly, (H2) is rejected which states that there is no difference in the efficiency of the two selection algorithms.

Further independent t-tests on the exposure efficiency of the two selection algorithms showed that the CAL system's efficiency is mostly effective with A and B students as they are the ones who benefit most from skipping questions they had already known (or having them ordered last), while C students would make a lot of mistakes and would thus benefit from any manner of exposure (Figure 5.2).
Nevertheless, the significant difference between the two systems' learning gain results for C students shown in figure 5.1 proved that exposure does not directly correlate with learning, and this, therefore, reflects the benefit of the CAL's compatibility considerations when dealing with C students.

Regarding the engagement level, the independent t-test results showed no significant difference between the two different adaptivity features ($p = 0.455$), which accepts (H3) stating that there is no considerable difference between the two adaptivity features in terms of the level of engagement the user experiences. An important limitation in the experiment that might explain the mentioned results is that the students playing the emotionally adaptive version of the game usually reported they were feeling "excited", and according to the system's implementation "excitement" is a desirable learning emotion and therefore no noticeable changes were made in the game. Moreover, another problem was that the session was considerably short with not many switches between the game scenes, so the user did not get a chance to experience all the different changes that correspond to the different emotional reports. As a result of all of this, the emotional adaptivity feature was latent for the most part of the experiment.

It can be concluded that using the proposed CAL selection algorithm can be beneficial in item-based e-learning systems to optimize the learning process. It offers effective means for improving learning efficiency and rendering items that are compatible with the student's level of mastery of a certain instructional topic.

Chapter 6

Conclusion

The literature review showed that there exists a diversity of methods for adapting the difficulty in an e-learning system. There is also different models for the CAT methodology which achieve item-based adaptive difficulty. CATs based on IRT provide point-estimation of examinee ability, however, they need a large number of students for the calibration phase. Meanwhile, CATs based on expert system models need less students for calibration but can only provide a classification for the examinee according to a discrete number of categories.

We have used a CAT expert system model and extended it to initially classify the learner as an A-Student, B-Student, or C-Student, after which we used their resulting classification to efficiently train them to advance to a higher-order classification (eg. from B student to A student). This comprised our CAL system which features compatibility between the administered questions and the learner's current ability, efficiency in learning to advance to the next higher classification, and flexibility for the educator to use the system to generate serious games for their educational content with an adaptive difficulty feature.

Additionally, we have implemented another version of the system with an alternative randomized selection algorithm which operates in the traditional manner of increasing the questions' difficulty sequentially. We have also embedded a different adaptivity feature into this version which observes the learner's emotions using the SAM questionnaire redesigned with visual emotional indicators, and adapts the game mechanics and UI according to the recognized emotion.

Afterwards, we designed a pool of MCQs on a topic on Pathology from a university-level course and empirically calibrated it according to difficulty, discrimination, and utility parameters. Then, we held a comparative experiment in which two groups of Pharmacy students were exposed to the two different versions of the system and conducted a series of 3 tests on the two groups to compare between them in terms of (1) learning gain, (2) exposure efficiency, and (3) engagement level.

Finally, we analyzed the results of the tests and reached the conclusion that the CAL system is better for students of all categories (A, B, or C) in terms of exposure efficiency

and learning gain. However, there is no significant difference between the two adaptivity features with regard to their effect on the engagement level of the student.

This draws the important conclusion that the proposed CAL algorithm is effective in optimizing the time spent on learning in comparison to the traditional non-adaptive study techniques in which the learner answers questions with increasing difficulty.

6.1 Limitations

The most prominent limitation to our work was during the calibration phase. Calibration following the adopted CAT expert system model needs a minimum of 50 students to properly estimate the item parameters, however, due to the limited number of students who attended the calibration session (23 students), the calibration results were not as reliable as they should be.

Another problem encountered during the experiment which might have affected the learning gain results, were that the results heavily depended on how seriously the student would take the post-test, as the question pool consisted of lengthy MCQs of higher-order cognition.

Finally, a problem recognized during the experiment as discussed in the results section 5 was that the emotional adaptivity trait was latent for the most part of the experiment as most of the participants would always report that they are feeling happy/excited. Also, the time of the experiment did not give the user a chance to experience the different adaptivity features.

6.2 Future Work

The utmost focus in future research should be to provide better definitions of efficiency and to design tests that evaluate the learning efficiency of the proposed algorithms as the amount of knowledge gained per unit time, instead of just the efficiency of exposure.

Additional features will also be implemented from both the educator's side and the learner's side to enhance the usability of the system.

6.2.1 Developer Tools

For the educator, an extra tool is to be developed which visualizes the calibration results for the educator, as well as enables them to run computer simulations on the calibrated pool. For example, the computer simulation could be a virtual student that randomly responds to the questions being administered to them by the CAL engine. The educator can then observe how the CAL operates and how many questions it takes for it to terminate and classify the student. The purpose of this is to provide the educator with more information to enable them to hard-code operative parameters such as the test's error probabilities and test length constraints.

6.2.2 Save Progress

For the learner, an important question to ask would be what if the learner wants to stop training and save their current progress for the future. In this case, we plan to provide the learner with two options that gives them conscious control over the session, a *save* option in which they can save their current progress and later pick up right from where they left off, and a *re-take the test* option where they can choose to stop training and re-take the test in hope of a different classification and subsequently a different cluster of questions for training. In the latter case, strategic randomization techniques will have to be implemented at the starting point to ensure that the initial sequence of questions is different each time, in addition the difficulty of the first item should be chosen according to the learner's classification the last time they took the test.

Appendix

Appendix A

Engagement Questionnaire

1. I felt in control of what I was doing
2. I was absorbed intensely by the activity
3. I found the activities enjoyable
4. I found the activities interesting
5. I was frustrated by what I was doing
6. The activities bored me
7. The activities excited my curiosity
8. I knew the right thing to do
9. It required me a lot of effort to concentrate on the activities

List of Figures

Bibliography

[1] Assessing by multiple choice questions. https://teaching.unsw.edu.au/assessing-multiple-choice-questions. Accessed: 2019-06-10.

[2] International association for computerized adaptive testing. www.iacat.org/.

[3] Luigi Anolli, Fabrizia Mantovani, Linda Confalonieri, Antonio Ascolese, and L Peveri. Emotions in serious games: From experience to assessment. *International Journal of Emerging Technologies in Learning (iJET)*, 5(2010), 2010.

[4] Albert Bandura. *Self-efficacy: The exercise of control*. Macmillan, 1997.

[5] Albert Bandura and Daniel Cervone. Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of personality and social psychology*, 45(5):1017, 1983.

[6] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[7] Walter B Cannon. Again the james-lange and the thalamic theories of emotion. *Psychological Review*, 38(4):281, 1931.

[8] Susan DeAngelis. Equivalency of computer-based and paper-and-pencil testing. *Journal of Allied Health*, 29(3):161–164, 2000.

[9] Pieter Desmet. Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Funology*, pages 111–123. Springer, 2003.

[10] Theo JHM Eggen. Computerized adaptive testing item selection in computerized adaptive learning systems. *Psychometrics in Practice at RCEC*, page 11, 2012.

[11] Andrew J Elliot and Markus A Maier. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65:95–120, 2014.

[12] Elaine Fox, Riccardo Russo, Robert Bowles, and Kevin Dutton. Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of experimental psychology: General*, 130(4):681, 2001.

[13] Roy O Freedle and Richard P Durán. *Cognitive and linguistic analyses of test performance*, volume 22. Ablex Pub, 1987.

[14] Theodore W Frick. Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research*, 5(1):89–114, 1989.

[15] Theodore W Frick. Analysis of patterns in time: A method of recording and quantifying temporal relations in education. *American Educational Research Journal*, 27(1):180–204, 1990.

[16] Theodore W Frick. Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research*, 8(2):187–213, 1992.

[17] Elissavet G Georgiadou, Evangelos Triantafillou, and Anastasios A Economides. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, 5(8), 2007.

[18] Noriko Hara. Student distress in a web-based distance education course. *Information, Communication & Society*, 3(4):557–579, 2000.

[19] Jen Harvey and Nora Mogey. Pragmatic issues when integrating technology into the assessment of students. *S. Brown, P. Race, & J. Bull (Eds.), Computer-assisted assessment in higher education*, pages 7–20, 1999.

[20] Nina Keith and Michael Frese. Effectiveness of error management training: a meta-analysis. *Journal of Applied Psychology*, 93(1):59, 2008.

[21] G Gage Kingsbury and David J Weiss. A comparison of irt-based adaptive mastery testing and a sequential mastery testing procedure. In *New horizons in testing*, pages 257–283. Elsevier, 1983.

[22] Gary P Latham. *Work motivation: History, theory, research, and practice*. Sage, 2012.

[23] Mariana Lilley, Trevor Barker, and Carol Britton. The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education*, 43(1-2):109–123, 2004.

[24] JM Linssen. Adaptive learning in an educational game–adapting game complexity to gameplay increases efficiency of learning. Master's thesis, 2011.

[25] Edwin A Locke and Gary P Latham. *A theory of goal setting & task performance.* Prentice-Hall, Inc, 1990.

[26] FM LORD. Application of item response theory to practical testing problems. first. *Hilsdale, New Jersey, EUA: Lawrence Erlbaum Associates*, 1980.

[27] Ruth H Maki, William S Maki, Michele Patterson, and P David Whittaker. Evaluation of a web-based introductory psychology course: I. learning and satisfaction in on-line versus lecture courses. *Behavior research methods, instruments, & computers*, 32(2):230–239, 2000.

[28] B Jean Mason, Marc Patry, and Daniel J Bernstein. An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational computing research*, 24(1):29–39, 2001.

[29] Phil Mollon. Cognition and emotion. e. eich, jf kihlstrom, gh bower, jp forgas and pm niendenthal. oxford university press, new york, 2000. no. of pages 259. isbn 0-19-511334-9. price£ 17.95 (paperback). *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 16(4):487–488, 2002.

[30] GT Plew. A comparison of major adaptive testing strategies and an expert systems approach. *Unpublished doctoral dissertation, Indiana Uni-versity, Bloomington*, 1989.

[31] Norma Pritchett. Effective question design. *Computer-assisted assessment in higher education*, pages 29–37, 1999.

[32] MARK D RECKASE. A procedure for decision making using tailored testing. In *New horizons in testing*, pages 237–255. Elsevier, 1983.

[33] Ute Ritterfeld, Michael Cody, and Peter Vorderer. *Serious games: Mechanisms and effects*. Routledge, 2009.

[34] Lawrence M Rudner. An examination of decision-theory adaptive testing procedures. In *annual meeting of the American Educational Research Association*, 2002.

[35] Lawrence M. Rudner. Param calibration software logistic irt models (freeware). http://echo.edres.org:8080/irt/param/, 2012.

[36] Jailan Salah, Yomna Abdelrahman, Ahmed Dakrouni, and Slim Abdennadher. Judged by the cover: Investigating the effect of adaptive game interface on the learning experience. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, pages 215–225. ACM, 2018.

[37] Samuel A Schmitt. *Measuring uncertainty: An elementary introduction to Bayesian statistics*. Addison-Wesley, 1969.

[38] Judith A Spray. Multiple-category classification using a sequential probability ratio test. 1993.

[39] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[40] Nathan A Thompson and David J Weiss. A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, 16, 2011.

[41] Abraham Wald. Sequential analysis, john wiley & sons. *New York, NY*, 1947.

[42] M Yoes. An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter irt model. *Saint Paul, MN: Assessment Systems Corporation*, 1995.

[43] Dongsong Zhang, J Leon Zhao, Lina Zhou, and Jay F Nunamaker Jr. Can e-learning replace classroom learning? *Communications of the ACM*, 47(5):75–79, 2004.