Content Agnostic Game Based Stealth Assessment

by

Vipin Verma

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2021 by the
Graduate Supervisory Committee:

Scotty D. Craig, Co-Chair
Ajay Bansal, Co-Chair
Ashish Amresh
Roy Levy
Tyler Baron

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Serious or educational games have been a subject of research for a long time. They usually have game mechanics, game content, and content assessment all tied together to make a specialized game intended to impart learning of the associated content to its players. While this approach is good for developing games for teaching highly specific topics, it consumes a lot of time and money. Being able to re-use the same mechanics and assessment for creating games that teach different contents would lead to a lot of savings in terms of time and money. The Content Agnostic Game Engineering (CAGE) Architecture mitigates the problem by disengaging the content from game mechanics. Moreover, the content assessment in games is often quite explicit in the way that it disturbs the flow of the players and thus hampers the learning process, as it is not integrated into the game flow. Stealth assessment helps to alleviate this problem by keeping the player engagement intact while assessing them at the same time. Integrating stealth assessment into the CAGE framework in a content-agnostic way will increase its usability and further decrease in game and assessment development time and cost. This research presents an evaluation of the learning outcomes in content-agnostic game-based assessment developed using the CAGE framework.

DEDICATION

In loving memory of my DEAREST BABA, from whom I inherited most of my character and calmness. I wish you had stayed a little longer to see me become a Doctorate.

I dedicate my dissertation to the four most important people in my life:
my parents, my sister, and my husband.

Without your love, support, and advice, I would not be who I am today. You have made me a better person by giving me your strengths and helping me overcome my weaknesses. Thank you with all of my heart.

"We cannot teach people anything. We can only help them discover it within themselves."

~ Galileo Galilei

ACKNOWLEDGMENTS

First and foremost, I have to thank my research supervisors, Scotty Craig, and Ajay Bansal. Without their assistance and dedicated involvement in every step throughout the process, this dissertation would have never been accomplished. I would like to thank you very much for your support and understanding over the past years.

I would like to immensely thank Ashish Amresh for accepting me into this degree program. I would like to thank him again for both being on my committee and for providing me the research questions that formed the basis of this dissertation.

I would like to thank Roy Levy for being on my committee and helping me with the Bayesian Networks required for my studies.

I would like to thank Tyler Baron for being on my committee and for providing me the foundation, the CAGE architecture, which was required for my dissertation.

I would like to thank Michael Edwards for an excellent class on Psychometric Methods and Multivariate data analysis which helped me in creating and validating the assessments used in the game, specifically the item response analysis.

Lastly, I would like to thank god and my parents who made me capable enough to earn a PhD.

TABLE OF CONTENTS

Page

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

The video game industry expanded a lot during the pandemic especially during the lockdown imposed by the coronavirus and is expected to surpass both film and sports combined (Gilbert, 2020). It generated about $180 billion in revenue, of which $2.6 billion was the revenue generated by serious games alone (Adkins, 2016). Revenue yielded by serious games is expected to increase to 8.1 billion by 2022 (Adkins, 2017), with the highest revenue obtained from China, followed by the US, and India, indicating the extent to which serious games are being used across various countries in the world. Among the various sectors, industrial corporations have the highest demand growth rate for serious games, followed by preschools and higher educational institutions. One of the key factors which are promoting the growth in serious game sales is the increasing demand for early childhood learning games, followed by the decrease in resistance to serious games. In earlier days, parents used to stop their children from playing games (Steinberg, 2012). But in the present era, 70% of the parents feel that video games have a positive influence on their children and about 67% of them play with their kids at least once a week (Entertainment Software Association, 2020). With the availability of fast network connectivity, serious games can use location-based services, virtual reality, augmented reality, and have low latency rates, enabling online multiplayer games with ease. Thus, there are many factors that are operating simultaneously to boost the serious games sector.

Video games are used for a wide variety of purposes, ranging from recreation (Biddiss & Irwin, 2010), education (Squire, 2003), training (Rosser et al., 2007), to platforms for advertisements (Schmierbach, 2017). For example, the game *Need for Speed* has been used by various companies to advertise their brand (Subani, 2009). In the United States, 67% of households own a video game console or a variant of a gaming device, with an average of 1.7 gamers in each household (Entertainment Software Association, 2017).

1.1 The Current Problem

Development of an educational game and assessment takes a significant amount of time, and once the development is complete, the developers may well have to start over to create another game (Moreno-Ger et al., 2014). Baron (2017) designed a content-agnostic architecture called Content Agnostic Game Engineering (CAGE) for creating multiple educational games that rely on the same game mechanics, leading to lower time and cost requirement for building several games at once. However, the architecture did not implement a content-agnostic student model of assessment built into it, and the study employed survey questionnaires to assess the engagement, which Baron (2017) noted are interruptive in nature and leads to a reduction in the motivation level of players.

Previous research has used commercially available games for educational purposes (Van Eck, 2006) and has tried to integrate stealth assessment in an existing game (Shute & Wang, 2015). Further, Baron (2017) has provided an architecture that helps develop multiple educational games at once. But no research has been done regarding the use of stealth assessment in a content-agnostic way. To address this problem, stealth assessment will be built into the CAGE framework as it helps in sustaining the motivation level of the students (Shute & Ventura, 2013).

1.2 Research Questions

This research seeks to answer the following five research questions which will be discussed in more detail following the literature review.

1. Is there evidence of validity for the use of Bayesian networks to model learner beliefs in the CAGE based games?
2. Does game adaptation using affect assessment help in improving the learning and engagement of the player?
3. Does adding stealth assessment based adaptive game design improve learning?
4. Does adapting the game using stealth assessment enhance the engagement of the players?
5. Does CAGE (same game mechanics for different content in multiple domains) with adaptation help in sustaining student engagement and promote learning performance?

Chapter 2

THEORETICAL FOUNDATION

Motivation is to be moved to do something or stimulated to achieve an end (Ryan & Deci, 2000). People possess various levels and kinds of motivation, which are distinguished by the goals or reasons that cause action. The ones concerning serious games are the intrinsic and extrinsic motivation which can lead to different performance and quality of experience.

2.1 Intrinsic and Extrinsic Motivation

Intrinsic motivation is characterized by an innate desire to achieve something for personal satisfaction rather than to attain an outcome (Ryan & Deci, 2000). It is particularly important in the field of education, where it can facilitate better and high-quality learning. This is because the task is fun and challenging. This natural motivation plays a significant role in the overall development of an individual because they can acquire the knowledge and skills when acting on their natural tendencies (Ryan et al., 2005). While someone may be intrinsically motivated towards an action, others may not find the same task motivating (Ryan & Deci, 2000). Intrinsic motivation usually occurs when the task at hand is inherently interesting for an individual, comprising challenge, novelty, or aesthetic value to them. After early childhood, the freedom to be intrinsically motivated is shaped by tasks that require a person to assume responsibilities for disinteresting tasks. For example, it appears that intrinsic motivation to learn tends to weaken with one's advancing grade in school. Video games being inherently more interactive than the static classroom material can help in sustaining intrinsic motivation (Freire et al., 2016) provided that the engagement is kept intact, using unobtrusive assessment techniques like stealth assessment (Shute & Ke, 2012).

While intrinsic motivation is a personal phenomenon, extrinsic motivation is external and driven by external rewards, like money, to accomplish a task (Ryan & Deci, 2000). Tasks performed for extrinsic rewards may cause disinterest, resistance, and resentment among students. Sometimes the work required by students may be inherently laborious and boring, and in such cases, educators are required to motivate the students extrinsically to teach them, such as using scores and grades in a classroom setting.

Deci (1971) conducted various experiments to find the effect of external rewards on the intrinsic motivation of an individual. In each experiment, subjects performed an activity that was observed for their motivation level during three periods. During the second period, external rewards were given to the test subjects while control subjects received no reward. Differences in the motivation level were observed during the first and the third period for both the test and control groups. It was found that money used as an external reward tends to cause a decrease in intrinsic motivation. However, when positive feedback and verbal approval were used as a reward, the level of intrinsic motivation got enhanced as it is less likely to be treated as a control mechanism by the subjects. This behavior is explained with the help of Self-Determination Theory (SDT) in the following section.

These observations were termed as the over-justification hypothesis by Deci (1971) and were further tested by Lepper et al. (1973) with preschool children in a field experiment. The children who were selected for the experiment showed an initial intrinsic interest in a drawing activity during the in-class drawing sessions. These children were divided into three groups, namely: expected reward, unexpected reward, and no reward which is also the control condition. It was found that the students in the expected reward condition show less intrinsic interest in the subsequent activities than the students in the other two conditions. On the other hand, students in the second group in which they were presented with an unexpected reward at the end of the activity showed a substantial increase in their intrinsic interest in the activity. These observations suggest that carefully providing the extrinsic motivation to a learner can strengthen the intrinsic motivation associated with the task, and if not done properly may ruin it altogether. This is an important observation that should be taken into consideration when designing any kind of learning activity, classroom learning, intelligent tutoring system, or educational video game.

The over-justification hypothesis has direct implications in the field of education (Shute & Ventura, 2013). Students may or may not be intrinsically motivated in learning, but the education system tends to reduce the intrinsic value of learning by attaching extrinsic rewards like grades. Moreover, there has been a change in the learning model over the past years from learning by listening to learning by doing which requires more intrinsic motivation, and instructional games are primarily seen as a way to improve it (Garris et al., 2002). Challenge, curiosity, and fantasy are a few key factors that can make a video game intrinsically motivating. In a study carried out by Williams et al. (2008), they found the average playing time of a video game player to be around twenty-six hours per week. It is unlikely that they will spend so much time studying. Since video games are usually intrinsically motivating, it is highly desirable to leverage those motivating factors in the field of learning.

2.2 Self Determination theory

SDT is related to the magnitude to which a person's actions are self-determined or motivated (Ryan & Deci, 2000). It involves an examination of an individual's natural psychological needs and inbuilt growth tendencies that are the basis for their personality and self-motivation. SDT also involves the investigation of surrounding factors that inhibit personal well-being and self-motivation. Intrinsic and extrinsic motivation are the basis of SDT (Lepper et al., 1973). Deci and Vansteenkiste (2003) claimed that there are three essential components of SDT. First, humans have the innate power to work on and command their internal forces rather than being passively controlled by them, being constitutionally proactive. Second, they have an inbuilt propensity towards integrated functioning and development and are inclined to pursue the means that promote their outcomes and positive processes. Third, actions and optimal growth, although being integral to humans, do not occur automatically but require sustenance from the social environment. An absence of these three nutrients from the social environment may lead to negative outcomes like alienation and passivity. Three innate psychological needs that foster motivation and well-being have been identified – the needs for autonomy (DeCharms, 1968), relatedness (Reis, 1994), and competency (Harter, 1978). These needs can be seen operating across time, culture, and gender, encouraging their optimal functioning (Chirkov et al., 2003).

The need for autonomy is related to the integrated sense of self, and the feeling of control over one's surroundings (Deci & Vansteenkiste, 2003). Autonomy pertains to an individual's desire to be a causal agent and willingness to support their actions at the highest level. The need for relatedness revolves around peoples' inclination to connect and interact with others, to achieve a sense of belonging. Finally, the need for competency relates to the innate desire of achieving mastery in dealing with the environment and seeking control of one's surroundings. These needs are crucial, and individuals are found to be favoring situations that allow gratification of these needs as opposed to the ones that thwart them. SDT explains the behavior observed by Deci (1971) in his experiments regarding the over-justification hypothesis where he found that expected extrinsic rewards for an intrinsically motivated activity undermines the intrinsic motivation associated with it. Extrinsic rewards being treated as a control mechanism weakens the autonomy leading to a reduction of intrinsic motivation. Also, the unexpected positive feedback accomplishes an individual's need for competence, enhancing their intrinsic motivation.

SDT is widely popular and has been applied to a variety of domains from parenting (Soenens et al., 2007), teaching (Roth et al., 2007), sports (Fortier et al., 2007), workplace (Fernet et al., 2004) to health (Kennedy et al., 2004). This theory has also been applied to explain the motivational pull in video games by Ryan et al. (2006). They carried out four experiments using single-player and online multi-player games to find out the effect of the three needs of autonomy, relatedness, and competency in independently predicting their enjoyment and future game play. In the first experiment, they administered a questionnaire as a pre-test and post-test to a group of 89 undergraduate students who played a game called Super Mario 64 (1996). The questionnaire employed the Player Experience of Need Satisfaction (PENS) scale to measure the degree of autonomy, competence, and relatedness satisfaction (Ryan et al., 2006). Experiment results confirmed the hypothesis that the needs of autonomy and competence can account for the motivation and enjoyment level within a game (Ryan et al., 2006). In their second and third experiment also, they administered a questionnaire as a pre-test and post-test to a group of 50 and 58 undergraduate students respectively. In the second experiment, two games were used, The Legend of Zelda: The Ocarina of Time (1998), and A Bug's Life (1999) while in the third experiment, four games were used, Super Mario 64 (1996), Super Smash Brothers (1999), Star Fox 64 (1997), and San Francisco Rush (1997). In both these experiments, participants played the games during their visits which were separated by two to seven days each. These two experiments corroborated the results from the first experiment. For the fourth experiment, they assessed 730 people from an online multi-player gaming community to account for the needs of relatedness. An online survey was used for this purpose. In addition to supporting the results from previous studies, this study found that the need for relatedness is a key factor that contributes to game enjoyment and intentions of future play. These results suggest that the role of SDT in video games can be very useful. Thus, creating a gaming environment that is autonomy-friendly, competence-evoking, and relatedness-invoking can help keep up motivation levels in a video game (Sørebø & Hæhre, 2012).

Chapter 3

GAME MECHANICS AND CONTENT DOMAIN IN GAME DESIGN

It is important to understand what a game is before going deeper into their literature. Kelley (1998) defined game as a kind of play governed by a set of rules or mechanics that specify an objective and the ways in which it can be achieved. A video game is a type of game that involves audio-visual apparatus and possibly a story (Esposito, 2005). A report by Hines et al. (2009) stated that games as an educational medium offer powerful affordances for learning and encouraged for expanding the research in this domain. The boom in research that followed found the digital games to be more effective for learning as compared to their non-gaming counterparts (Clark et al., 2016).

A game mechanic is a control mechanism, a rule of game play used by a player for interactions within the game world to achieve the goals of the game (Sicart, 2008). In Angry Birds[1], the player can fire a bird into the sky by dragging them off a catapult using touch and drag on screen, and then release to launch, a mechanism called sling-shotting, depicted in Figure 1. The content domain of a game is the subject knowledge that the game is intended to impart (Baron & Amresh, 2015). Consider a game designed to teach chemical equation balancing skills to its players. Chemistry would be the content domain for such a game. While game mechanics are important for any video game, the content domain is considered only with regards to an educational game. Commercial games do not usually define a content domain as they are not trying to teach anything specific. Educational games, however, need to

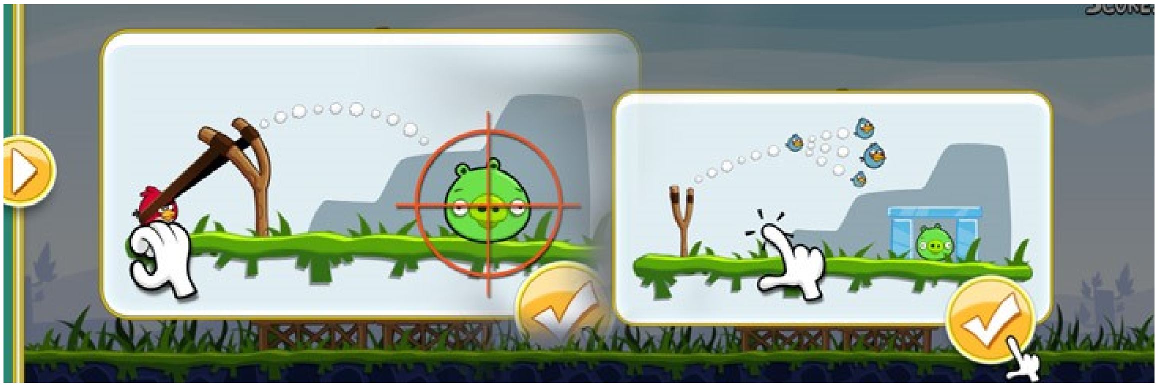


Figure 1. An example showing how to launch an angry bird using catapult.
Source: Angry Birds (2021)

define a content domain to make sure that the game is designed to impart skills in that domain.

Commercial games adapted for teaching have been used in the past by some educators (Van Eck, 2006). Although this approach has shown to be effective (McFarlane et al., 2002), it poses various challenges (Van Eck, 2006). In these games, the content is not integrated well with the game mechanics since they were not made to teach a subject, in the first place. Further, teachers may not possess the ability to alter the game and modify the content, leading to finite or inaccurate content. On the other hand, planning the content domain and game mechanics from the beginning provides a deeper connection of the content with mechanics, making the game fun to play while being educative as well.

3.1 Game-Based Assessment

Chin et al. (2009) described the assessment as the procedure used to decide if the learning goals are met or not, with the help of data. Consider a game designed to teach cryptographic encryptions to its player. Then the role of the assessment would be to identify if the player has gained the knowledge of how to use the encryption methods like Caesar cipher.

Assessment of the student knowledge is as important as setting up the content and mechanics of the game. In a level-based game, the student will be allowed to progress to the next level only if they demonstrate through their game-play that they have learned the knowledge required to progress to the following stage. If there is no assessment, then the level of progress will not be an indicator of the skill level or knowledge gained by the player, and they will be stuck on the current level forever and get frustrated.

Plass et al. (2013) have identified three variables of interest during an educational assessment: general trait variables, general state variables, and situation-specific variables. Trait variables such as executive functions and spatial abilities of players are more or less stable but are not typically targeted in educational video games, although they can be impacted by game-play. State variables such as knowledge in an area are the ones that are targeted in serious games. Engagement, cognitive load, affective state, are the situational variables and are there because of the player's interaction within the gaming environment. A typical game would thus be governed by a player's trait variables and should be designed to level up their state variables while keeping their situation variables in an optimum range for best results.

Digital games are gaining attention extensively owing to three factors which are arousing interest in games as an instrument of learning (Van Eck, 2006). First, the widespread research conducted by the advocates of Digital Game-Based Learning and the growing volume of literature in the area. Second, the current generation who have turned away from traditional forms of education and prefers active interaction along with inductive reasoning. Third, the gaining popularity of entertainment video games over the last few years. Building upon these factors, video games are moving away from being associated with a stigma of being just meant for play. This does not imply that video games can teach everything to everyone. Designing a video game may need a professional game designer and developer, as opposed to just academicians, and a meaningful context for effective and engaged learning. If an academician makes a game for learning, a likely problem is that the game is too focused on learning and not on fun. When developers build an educational game, it is often too entertaining, but not enough learning.

One of the most pertinent problems in this context is the evaluation of the learning outcomes that a game can offer (Bellotti et al., 2013). Learning assessment is of two types: formative and summative (Boston, 2002). Formative is used during the learning process, while summative is used after the learning has taken place. In the context of educational video games, formative ones would mean continuous assessment during the game play, while summative ones would mean an assessment at the end of game play. Formative ones evaluate the learner's strengths and weaknesses and help educators in tailoring their practice accordingly while summative ones are carried out at the end of the learning session and provide detailed feedback to the learners (West & Bleiberg, 2013). Summative assessments are usually high stakes while formative ones may or may not be. Survey questionnaires and teacher evaluations are the most common and simplest methods of summative assessment of a player's knowledge (Bellotti et al., 2013). Boyle et al. (2009), Baron (2017), and many others have used this method of survey questionnaires in their studies. Various questionnaires have been developed to assess players' engagement during the game, such as the Game Experience Questionnaire (IJsselsteijn et al., 2008) and the revised User Engagement

Scale (Wiebe et al., 2014). The Game Experience Questionnaire is available in four languages and measures the seven aspects of a player's experience: Tension, Positive Affect, Negative Affect, Sensory and Imaginative Immersion, Competence, Challenge, and Flow (IJsselsteijn et al., 2008). The revised User Engagement Scale (UESz) consists of 28 items that measure engagement in video games using a 5-point Likert scale. It is composed of four factors: Aesthetics (8 items), Perceived Usability (8 items), Satisfaction (5 items), and Focused Attention (7 items) (Wiebe et al., 2014).

Formative assessments yield data which is critical to high-quality teaching and provide a chance to rectify the mistakes during the learning process itself without any serious penalties (West & Bleiberg, 2013). They are a powerful but highly resource-intensive teaching tool. They can replace the questionnaires and traditional approaches to evaluation which interrupts the learner's flow and has negative effects. When employed in a game, formative assessment provides people with the ability to adjust the difficulty of the game in real-time and adjust the game to the pace of the learner and can be embedded into the game itself, such as in the case of stealth assessment (Shute & Ventura, 2013).

Lee et al. (2013) carried out two experiments to investigate the effect of in-game or formative assessments on the learner's task completion speed and engagement. They used a game called Gidget, which teaches programming to its players. In the first experiment, they assessed the engagement of 200 participants by measuring the total game play time and the number of levels completed during the game. In the second experiment, they assessed a total of 30 participants for their speed by measuring the quickness with which players completed the levels. Subjects for the study were non-programmers who had never done coding before. In both experiments, treatment was the inclusion or exclusion of assessment levels in the game. In the test condition, each set of levels was followed by two assessment levels which were built aesthetically similar to other game levels to flow with the game story. The assessment levels administered explicit questions, like a teacher evaluation at the end of a course, but they were followed by immediate feedback to the response. Lee et al. (2013) found a strong effect suggesting that the in-game assessment can help enhance the learner's speed and engagement during the learning process, and thus impact the learning process positively. However, participation involved a small monetary reward, which impacts the intrinsic motivation of the learner as discussed previously, but, at the same time, immediate feedback helped sustain the motivation during the game.

Kiili and Ketamo (2017) investigated the fairness of game-based assessments with 60 sixth grade students using the Semideus research engine for teaching rational numbers. They conducted paper-based and game-based tests to evaluate the student performance and found a significant correlation between the two. They further observed that the game-based tests were associated with significantly lower anxiety as compared to the paper-based tests and led to better flow experience and test performance. Their results suggested the use of game-based assessments as fair, meaning that the game flow and player anxiety levels will be similar for all players, independent of their earlier playing experience and gender, unlike the test anxiety associated with the paper-based tests.

The assessment is useful for game designers as well since it provides useful information about the student pain points, engagement levels, feature usage, etc., and helps to refine the game for future players (El-Nasr et al., 2016). However, much like content-agnostic mechanics, there is a need for content-agnostic assessment, so that the developers don't have to create an assessment for every content being taught by the game (Baron, 2017). Thus, CAGE will incorporate a content-agnostic stealth assessment embedded within to sustain the player flow while assessing them at the same time.

3.2 Game Mechanics and Assessment dependent on Content

Previously, commercial games have been adapted for educational purposes, but they pose several challenges (Van Eck, 2006). While some of the problems arise because of the inability of the educators to make required modifications to the game (Tang et al., 2009), many issues occur because the content being taught is not tied to the mechanics of the game. This suggests the ideal solution is to link the mechanics with the game content (Van Eck, 2006). However, linking mechanics and game content could cause other problems.

Consider an educational video game that is designed to teach cipher-text to its players. A development studio makes a successful game that teaches cipher-text and embeds an assessment into it to evaluate the learning as the game progresses. Over time, as user needs change, the studio may decide to make a new game for teaching chemistry. The problem that they will come across is that how can the example game which is used to teach cipher-text can also be used to teach chemical equation balancing while having a valid assessment at the same time?

It would be rather difficult to efficaciously teach chemical equation balancing using the mechanics of the cipher-text game. It would be equally difficult to assess the learning of chemical equation balancing with the assessment that was developed for a cipher-text game. Developers may need to make a lot of adjustments to the game mechanics and assessment, spend significant time in coding the game or start an entirely new project from scratch.

3.3 Disconnecting the Three

As mentioned previously, the mechanics and assessment are not transferable across various content domains if they are heavily tied to them. But if they are transferable then it may pose two problems. The first one is that it can lead to inaccuracy in the content and assessment and thus pose difficulty using it as a good educational tool, the same problem which is encountered when using commercial games for educational purposes (Van Eck, 2006). However, CAGE architecture can be used to palliate this, as the game design will incorporate learning and assessment strategies from the inception of the game (Baron et al., 2016).

The second problem is the over-generalization that this may cause. Mechanics that are omnipresent are hard to enjoy and could be detrimental to learning (Baron, 2017). It would become boring to play many games all of which employ the same game mechanics while teaching different contents. Thus, it can have serious ramifications for the learning. There exist many specialized skills, for example, operating a nuclear power plant, that requires focused training. It will be extremely hard to build a universal set of mechanics and assessment which can be used to teach and assess any type of content. However, keeping this in mind from the beginning while developing a game and trying to accommodate it using stealth assessment and student model for dynamic game adaptation and feedback will help alleviate this problem to a considerable extent. Further, mechanics and assessment that can work across several domains would be better over the current state where a dedicated game is required for each type of content and assessment.

3.4 Impediments to Integration of Game-Based Learning in Classroom

It is quite difficult for educators to incorporate educational video games in a regular classroom (Del Blanco et al., 2012). This is due to several reasons. First, the games may not be properly synchronized to the educational standards, as they vary across countries and regions. Second, the technology required for gaming is costly to buy and maintain, and the developmental cost associated with producing a video game is also high. Serious games being financed by non-governmental organizations and research projects have a smaller budget in comparison to commercial games, hence they are unable to leverage the economies of scale where high development cost is paid off by the huge volume of sales (Freire et al., 2016). Thirdly, games may be interrupting educators because of the challenges it offers in aligning the game with the curriculum goals and evaluating the learning imparted by the games (Del Blanco et al., 2012). In order to integrate video games into the classroom, it is necessary to have the least possible impact on the educator. Kenny and McDaniel (2011) argued that lack of proper teacher training methods, poor infrastructure, and complex technology are the major barriers to integrating games in regular classrooms.

Del Blanco et al. (2012) suggested a framework to aid smoother integration of video games in a classroom. This framework has three goals that aim to reduce the overhead on educators. The first goal is to define the objectives of the game and measuring the student learning during the gameplay. Games involve a large number of player interactions, and thus even a small gameplay session can generate data enough to inundate the system (Freire et al., 2016). Del Blanco et al. (2012) thus advised using short games having short completion time which can ease game development, analytics, and maintenance while aiming for a single goal aligned with the learning outcome. These short educational games can be coupled with learning with repetition to form a burst game (Baron & Amresh, 2015). Burst games keep the learning experience short, making sure that the hours of work won't be lost due to failure while keeping the volume of data generated to a minimal extent. Del Blanco et al. (2012) indicated four variables that can be used for assessing student learning in a serious game, namely: global score, game completion status, total time, and play time. The second goal of the framework is to adapt the game to each student's needs suitable for his learning style. For example, if a student is performing poorly, then adjusting the game difficulty or providing additional help content to ease their learning. Third and the last goal is to encourage collaboration and reuse of successful game design among researchers and educators. Baron (2017) presented such a framework called CAGE in his dissertation which aims to reuse the development code of the game across various educational contents, encouraging reusability of the code and minimizing developmental cost and time.

Tüzün (2007) conducted a study to investigate video games from an international perspective, aimed to identify core challenges faced while integrating games in the classroom. In this project, Tüzün (2007) used three video games for teaching units on geography, first aid, and basic computer hardware to primary, secondary, and higher education students respectively. In the first game, students were asked to identify the country of origin of non-playable characters which were lost in the game world. The second game consisted of first aid in the context of fields and hospitals, and the third game was about fixing a giant malfunctioning computer. More than 3000 students participated in the three studies combined. Their findings reported issues in various domains, discussed in the following sections.

Tüzün (2007) found five issues related to the video game environment design.

First, it took a lot of time to design the games, with an average of 60 hours per game. Second, creating a background story in the context of games was a demanding task. Third, students had high expectations from the aesthetics of the game, as they were comparing it with the commercial and professional video games which are financed by billion-dollar industries. Fourth, students were required to complete orientation to help them understand the game play and mechanics before they could play the game, which reduced the time available for actual learning. This could be mitigated using a common orientation session across different subject matters, rooting out the need of doing orientation every time a game is played in a classroom. Fifth, game play time was limited by the class time and needed to cover the entire subject matter in a brief period. The first, second, and fourth problems were addressed by Baron (2017) in his dissertation thesis. He used a framework called CAGE to mitigate these issues in his study. Using this framework, a video game needs to be designed only once and the game mechanics can be re-used, eliminating the problem of re-designing, repeat background story creation and re-orientation of students.

Tüzün (2007) further found three issues owing to the school's infrastructure. First, to save time and have more game play time, it was required to set up the games on computers in advance, before the class began. A solution to this problem was suggested by Freire et al. (2016), which involves selling serious games as a service, deployed on the web, instead of selling it as a product, removing the overhead of deployment on individual computers again and again. Second, technical issues like firewalls and crashing of computers interfered with the access to games (Tüzün, 2007), and needed immediate technical support. Third, an absence of prompt technical support to rectify the technical faults may hamper the class, wasting a lot of teaching time. Further, he found balancing the entertainment and learning aspect of games to be difficult. Teachers were often found re-directing the students to learning tasks when they diverted from it.

Van Eck (2006) suggested integrating commercial off-the-shelf games into learning as the most suitable approach, but it may suffer from the finite or inaccurate content of the subject matter as it involves using existing games and customizing them for learning. Although his approach may fix some issues, others like the multiple orientations and limited game play time from the first set of issues regarding the game environment design are still not fixed, nor does it address any of the concerns raised by Tüzün (2007) regarding the school infrastructure. However, if games are designed from scratch like Tüzün (2007) did in his study, these limitations disappear, but the previous set of restrictions mentioned by Tüzün (2007) may re-appear. Thus, it is advisable to evaluate both the options and proceed with the best approach for using educational games in a particular classroom.

Chapter 4

STUDENT ASSESSMENT

Assessment is essential to gauge the current understanding of the student and to provide the interventions that are required for further learning (Gronlund, 1998). This chapter reviews the available techniques which have been used for assessment in the past. Stealth assessment which is one of the foundations of this dissertation is based on Evidence-centered Design and therefore it is reviewed as well.

4.1 Evidence-centered Design

Evidence-centered Design (ECD) was created by Mislevy et al. (2003) to support assessment developers in designing assessments. It helps assessment developers in explicating the rationales, choices, and consequences reflected in their assessment design (Rupp et al., 2010). ECD is a comprehensive framework suitable for the development of performance-based assessments which are created in the absence of easily definable test specifications (Mislevy et al., 2012).

External knowledge representations (EKRs) are used to recognize, represent, transform, store, share, and archive information and they are key to attaining proficiency in almost every discipline (Mislevy et al., 2010). An assessment is an EKR that elicits the knowledge and the ways to use that knowledge. For the assessment of learning in a domain, it is mandatory to explicate the EKRs in it. As opposed to internal knowledge representation in the brain, EKRs are designed to surpass the finite working memory and defective long-term memory over time, making it easier to capture and organize information. There are several examples of EKRs, such as simulations, formulas, maps, bus schedules, and the periodic table. EKRs, like maps and graphs, take advantage of the human strengths in understanding spatial relationships and identifying patterns to encode information. EKRs shape the instructional and assessment design in a domain by linking the proficiencies in that domain with the learning and assessment. (Mislevy et al., 2010) outlined the four components of EKRs. The first component is a world that is being represented, called the represented world. The second is a world that carries the representations, called the representing world, which includes only relevant entities and relationships and not everything contained in the represented world. The third is the set of rules that map these two worlds together, called the representing rules. The fourth, the process, uses these representations, which helps in defining the potential of the system being represented.

There are five layers in the ECD framework, governed by various EKRs (Mislevy et al., 2012). The types of activities and thinking that occur in each layer during the operation and development of an assessment system are shown in Figure 2. Each layer consists of representations, entities, and processes, the EKRs which are proper for the activities that occur in that layer. Table 1 from Mislevy et al. (2010) sums up these layers explaining their roles and key entities, such as concepts and building blocks, and the EKRs that aid in accomplishing each layer's purpose. Although ECD facilitates the creation of performance-based assessments, its layered architecture is too expensive to implement in a full-scale assessment model, hence its full-scale model is usually not used in practice (Crisp, 2014).

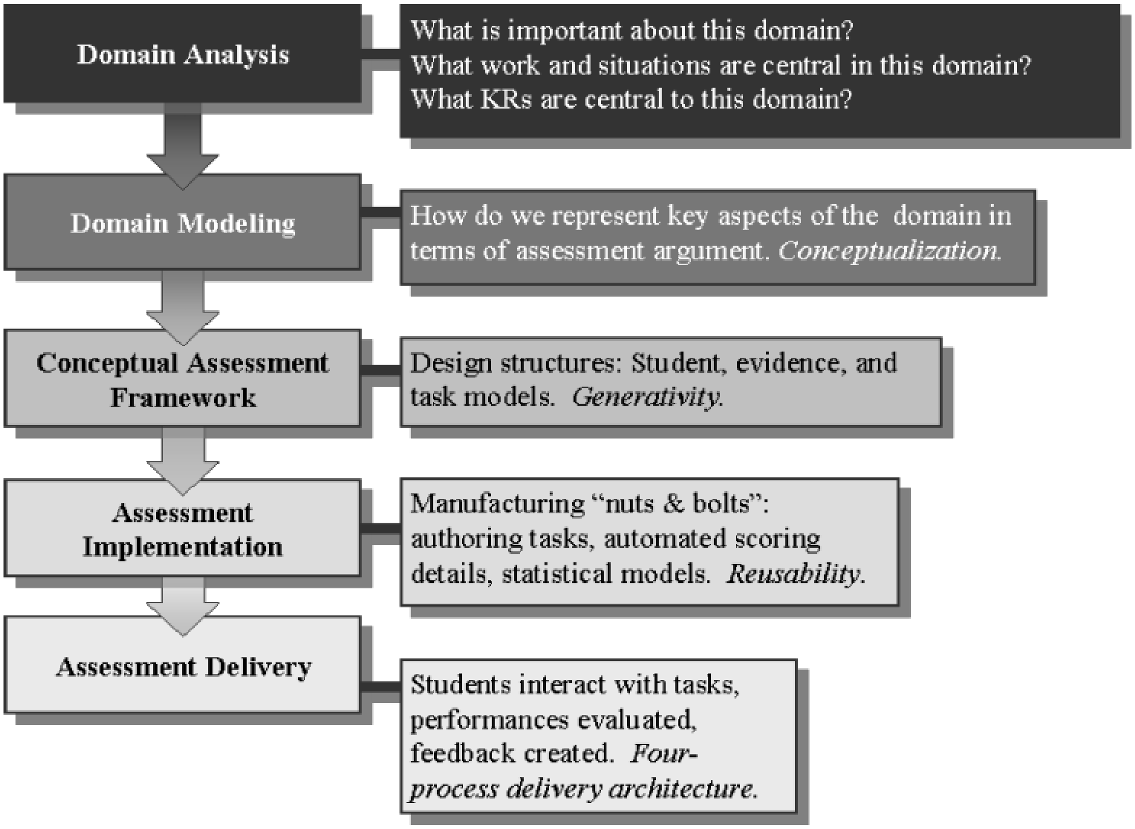

Domain Analysis
What is important about this domain?
What work and situations are central in this domain?
What KRs are central to this domain?
Domain Modeling
How do we represent key aspects of the domain in terms of assessment argument. Conceptualization.
Conceptual Assessment Framework
Design structures: Student, evidence, and task models. Generativity.
Assessment Implementation
Manufacturing "nuts & bolts": authoring tasks, automated scoring details, statistical models. Reusability.
Assessment Delivery
Students interact with tasks, performances evaluated, feedback created. Four-process delivery architecture.

Figure 2. Layers in the evidence-centered assessment design framework.

Source: Mislevy et al. (2012)

Table 1. Layers of Evidence-Centered Design.

Layer	Role	Key entities	Selected EKRs
Domain analysis	Gather substantive information about the domain of interest that has direct implications for assessment: how knowledge is constructed, acquired, used, and communicated.	Domain concepts, terminology, tools, knowledge representations, analyses, situations of use, patterns of interaction.	Content standards, concept maps (e.g., Atlas of Science Literacy, American Association for the Advancement of Science, 2001). Representational forms and symbol systems of domain of interest, e.g., maps, algebraic notation, computer interfaces.
Domain modeling	Express assessment argument in narrative form based on information from domain analysis.	Knowledge, skills and abilities; characteristic and variable task features, potential work products and observations.	Assessment argument diagrams, design patterns, content-by-process matrices.
Conceptual assessment framework	Express assessment argument in structures and specifications for tasks and tests, evaluation procedures, measurement models.	Student, evidence, and task models; student model, observable, and task model variables; rubrics; measurement models; test assembly specifications.	Test specifications; algebraic and graphical External Knowledge Representations of measurement models; task template; item generation models; generic rubrics; automated scoring code.
Assessment implementation	Implement assessment, including presentation-ready tasks, scoring guides or automated evaluation procedures, and calibrated measurement models.	Task materials (including all materials, tools, affordances); pilot test data for honing evaluation procedures and fitting measurement models.	Coded algorithms to render tasks, interact with examinees, evaluate work products; tasks as displayed; IMS/QTI representation of materials; ASCII files of parameters.
Assessment delivery	Coordinate interactions of students and tasks: task-level and test-level scoring; reporting.	Tasks as presented; work products as created; scores as evaluated.	Renderings of materials; numerical and graphical score summaries; IMS/QTI results files.

Source: Mislevy et al. (2012)

4.1.1 Domain Analysis Layer

This is the first layer of the ECD framework and it lays the foundation for the other layers (Mislevy et al., 2010). It mobilizes representations, beliefs, and modes of discourse for the target domain (Mislevy et al., 2012). How people think, what they do, and the situations in which they do it matters as much as the content of the domain. Knowledge, skills, and abilities that are to be inferred, student behavior the inference is based on, and the situations that will evoke those behaviors are defined in this layer (Mislevy et al., 2010). Maps and computer interfaces are some representational forms for the domains of interest. Identification of the EKRs is vital to domain analysis and involves activities like literature reviews and cognitive task analysis. Domain analysis helps in comprehending the knowledge systems used in a domain, and in identifying the valuable knowledge, task features, and performance outcomes (Hamel et al., 2006).

4.1.2 Domain Modeling Layer

This is the second layer of the ECD framework (Mislevy et al., 2012). In this layer, insights about the domain from its analysis are organized by the assessment developers in the form of assessment arguments. Domain modeling structures the outcomes of domain analysis in a form that reflects the structure of an assessment argument, using EKRs (Mislevy et al., 2010). A design pattern is a kind of EKR that was developed by Mislevy et al. (2003) for domain modeling and can be employed across various domains taking advantage of the same pattern. A design pattern is essentially a set of problems, solutions, and consequences that usually exist together.

It capitalizes on the recurring phenomenon in one's surroundings, which can be used later, providing a high level of reuse of experience and structures.

4.1.3 Conceptual Assessment Framework Layer

Conceptual Assessment Framework (CAF) layer is the nuts and bolts of assessment and responsible for the formal specifications of its operational elements (Mislevy et al., 2010). Information about constraints, goals, and logistics is combined with domain information by designers to create a blueprint for the assessment (Mislevy et al., 2012). The blueprint is created in terms of schemas for tasks, psychometric models, specifications for evaluating students' work, and definitions of the interactions that will support the operation and delivery of the assessment. This layer provides a structural bridging between the previous two layers and the actual objects and processes that will constitute the assessment. Blueprints for tasks, evaluation procedures, statistical models, delivery, and operation of the assessment are provided by the objects and specifications of the models from this layer as depicted. It involves EKRs like the measurement model.

The student model, evidence model, and task model are the main models in CAF, depicted in Figure 3 (Mislevy et al., 2010). There are other models as well, namely, assembly model, presentation model, and delivery model, but they are not central to CAF. The assembly model dictates the assembly of tasks into tests. The presentation model specifies the requirements of interaction with the learner, and the delivery model governs the requirements for the functional setting.

The student or competency model represents the variables that are being measured, which is a collection of skills that need to be assessed (Mislevy et al., 2012).

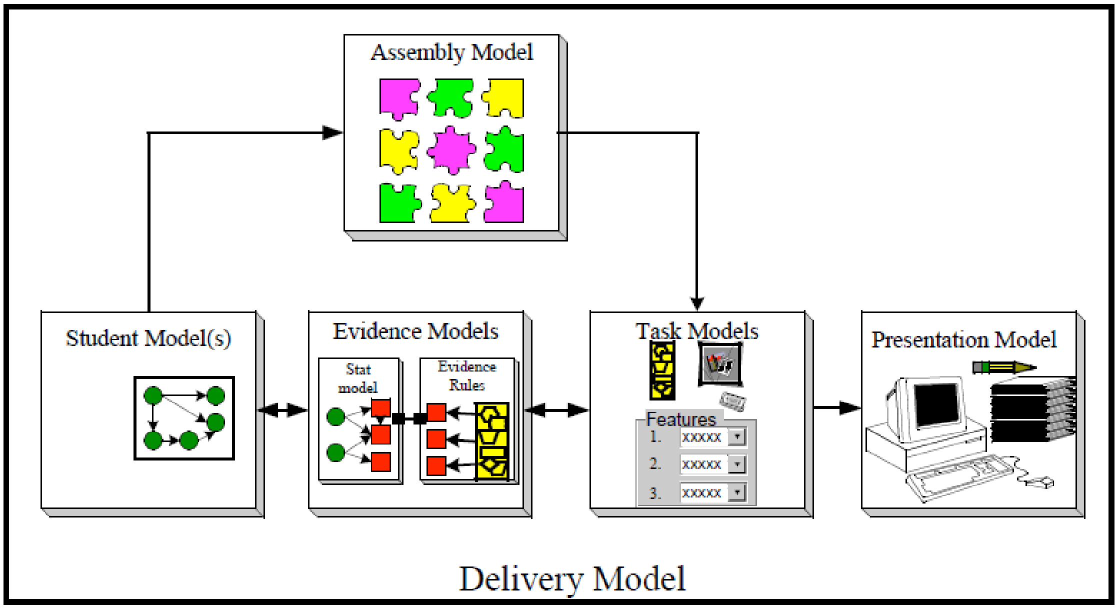


Figure 3. Central models of Conceptual Assessment Framework.
Source: Mislevy et al. (2010)

It is comprised of variables that target the aspects of students' skills and knowledge, at a grain size in accordance with the purpose of the assessment. The structure of these variables is used to capture information about the aspects of students' proficiencies and is a representation of the student's knowledge (Conrad et al., 2014). The student model reflects the claims which are used to define variables that can describe facets of knowledge, traits, and skills of a student (Kim et al., 2016).

The task model describes the task and environment features that should elicit the behaviors that can support evidence to measure the student's competencies defined in the student model (Shute & Spector, 2008). Specification of the cognitive units, affordances required to aid the student's activity, and the forms in which students' performances will be captured are the key design elements in the task model (Mislevy et al., 2012). It is composed of scenarios that can draw out the evidence to update the student model. Tasks usually correspond to quests or missions in games whose main purpose is to elicit the observable evidence about unobservable competencies (Shute & Spector, 2008). In some games, tasks may correspond to game levels, where the game play is divided into levels, although it may sometimes be challenging to define the task boundary (Kim et al., 2016).

The evidence model is the one which bridges the student and the task model and is a way to measure competencies (Conrad et al., 2014). It governs updates of the student model, given the student shows evidence of learning. It answers the question about why and how the observations in a given task constitute evidence about the student model variables (Shute & Spector, 2008). For example, in a video game, the player achieving a certain amount of score could be the evidence that can testify that learning has taken place. The task model further has two components: evidence identification component and statistical component (Shute & Spector, 2008). The evidence identification component provides the rationale and specifications for identifying and evaluating the salient aspects of work products, which will be expressed as values of observable variables. The statistical component is responsible for synthesizing the data generated in the evaluation component across tasks. This could be as simple as summing the percentage correct score or could be more complicated.

4.1.4 Assessment Implementation Layer

This is the fourth layer of the ECD framework where assessment developers make operational actualization of the models defined in the CAF layer (Mislevy et al., 2012). Its basic role is to implement the assessment and calibrate the models which will be used for measurement (Mislevy et al., 2010). Evaluation procedures and scoring guides are implemented here. Model fit is checked using the field test data, which is also used to estimate the parameters of the functional system (Mislevy et al., 2012). The tasks and parameters follow the data structures specified in the CAF model. The scoring rules and parameterized jobs are adjusted with the help of field test data. For passing the data from the design system to the calibration and scoring engine, and gauge student proficiency, the University of Maryland created a data management tool called Gradebook (Mislevy & Haertel, 2006). Assessment implementation interacts in both directions with other ECD layers (Mislevy et al., 2012). With the implementation, unexpected results and new findings may lead to better understandings of the domain or improvements in the CAF model.

4.1.5 Assessment Delivery Layer

This is the fifth layer of the ECD framework, in which students' interaction with the tasks takes place (Mislevy et al., 2012). Based on these tasks, their performance is measured and feedback reports and graphical summaries, which are key EKRs in this layer, are produced. It could be carried out by computers, humans, or both. For example, paper tests, computer-based tests, etc. Tasks, work products, and scores are the key entities here and score summaries are the EKRs (Mislevy et al., 2010). The assessment delivery is governed by four processes (Hamel et al., 2006). The activity selection process either utilizes the existing templates to create a task or selects one from the library. The presentation process presents the task to the student, handles the interaction, and captures the work product. The evidence identification process uses these work products to generate the scoring in accordance with the evaluation criteria established in the evidence model. The evidence accumulation process accumulates the observed evidence over time and summarizes it as a probability distribution for the competency model variables.

ECD has been used widely for educational testing and test development (Hamel et al., 2006; Zieky, 2014), simulation-based assessment (Bauer et al., 2003; Mislevy, 2011), modeling and assessing a school (Shute & Torres, 2012), psychometric modeling (Mislevy & Levy, 2006), and stealth assessment (Shute, 2011). An educational game that embeds stealth assessment into it must draw out behaviors that carry evidence affirming claims about competencies. However, the assessment must be done at an appropriate grain size, small grain size would mean added complexity and high resource requirement for the assessment. High grain size would mean less specific evidence for finding student competency. Moreover, applying ECD for scoring qualitative work, such as essays that involve a high degree of subjectivity would be difficult.

4.2 Educational Data Mining

Educational Data Mining (EDM) is data mining in the context of educational data which largely emerged from learner-computer interaction log analysis (Baker & Yacef, 2009). EDM involves researching and developing methods to discover patterns in large volumes of educational data generated during the student's interactions while learning (Scheuer & McLaren, 2012). Large amounts of educational data are attracting the developers' interest in creating new analysis methods mainly because of the boost in computational power as a result of the advances in educational technology (Nithya et al., 2016). The number of articles on EDM has grown tremendously in the past few years owing to the forming of peer-reviewed Journal of

Educational Data Mining (JEDM) in 2009 and various books on EDM (Romero & Ventura, 2010). The Pittsburgh Science of Learning Centre (PSLC) DataShop, a public educational data repository established in 2008, provides a lot of educational data making EDM feasible, leading to its further growth (Nithya et al., 2016). PSLC DataShop contains constantly growing student-computer interaction data stored as log files and offers online visualization tools for the data (Koedinger et al., 2010). The data is fine-grained, longitudinal, and extensive. However, most of it comes from a similar kind of tutoring environment (Romero & Ventura, 2010). There is a need to obtain data from other kinds of educational systems as well.

EDM is applied as well as pure research-oriented (Romero & Ventura, 2010). It aims to improve the learning process and guide students' learning and achieve a deeper understanding of the educational phenomenon. EDM is helpful not only to teachers but is equally beneficial for the students (Merceron & Yacef, 2005). Teachers can use it to figure out the students' learning process and manage the course while analyzing their own teaching methodology, and support themselves in decision making. They can also use it to improve student learning by providing them with feedback, making it a useful tool for summative assessment. EDM has been used to predict students' performance and knowledge using several types of data mining algorithms (Romero & Ventura, 2010). Its application also involves using the learner's style, motivation, behavior, meta-cognition, and emotional state to create a model of their skills and knowledge, called the student model, indicating that it can be used with the ECD model of assessment. A Bayesian Network (BN) is one of the most popular methods used for this purpose (Romero & Ventura, 2010) which will be discussed in further sections below. EDM supports sensing undesirable characteristics in students, such as low motivation, cheating, and has been used to prevent student dropouts (Romero & Ventura, 2010). It also helps the formation of student groups that have similar characteristics so that they can be provided with a personalized learning plan, raising their effective group learning.

A review by Baker and Yacef (2009) identified the four key areas in which EDM can be applied. The first area is the prediction of student learning with improvements of the student model, which can help cater to students' individual differences, enhancing their learning. The second area is the identification and amendment of domain models from data. For example, finding out the best order in which the instructions should be delivered to assist the student's learning style. The third one is about examining educational support, to find the most effective type of didactic support and promote learning. The fourth is the better realization of factors that affect learning, to create an improved environment that facilitates learning. This involves a persistence lookup for empirical evidence to improve and expand existing educational phenomena and theories.

Peña-Ayala (2014) conducted an extensive review of 240 EDM works published from 2010 until the first quarter of 2013, using data mining itself. He claimed that EDM is in its adolescent stage as 98% of the works that are cited, date to 2000 and above. Of the 240 EDM works that were reviewed, forty-three were related to student modeling, forty-eight were about student behavior modeling, forty-six were related to student performance modeling, and the rest belong to the other categories of EDM techniques, such as, assessment, student support and feedback, and teacher support. This suggests that a lot of work has been done to model student data using EDM techniques. Student modeling which is the first key area identified by Baker and Yacef (2009) will be one of the focus areas in this literature review.

D'Mello and Graesser (2010) conducted a study using binary logistic regression to predict the affective states of students using their posture while they were seated on a chair. One may associate high pressure exerted by the student on the seat with a high attention level and high pressure on the back of the chair with a low attention level (Bull & Argyle, 2016). D'Mello and Graesser (2010) examined the posture patterns and affective states of 28 students during their learning session with a narrative intelligent tutoring system. The pressure exerted by the student on the back and seat of the chair was measured to account for their postural configuration. A video of their posture and face was captured during the tutoring session. Assessment of the student affective state was made by the students themselves, an untrained peer, and two trained judges. Results indicated leaning back on the seat to be associated with disengagement and boredom from the learning session, accompanied by a heightened rate of pressure change exerted on the seat. Leaning forward was found to be related to a state of frustration or delight, depending on the angle of inclination at which they leaned forward. This study modeled the affective state of the student using their postural configuration.

Vocal dialogues (Litman & Forbes-Riley, 2006), body language, facial features (Baron, 2017; D'Mello & Graesser, 2010), and a combination of sensors (Muldner et al., 2010) have been used in the past to predict the affective states of learners. However, approaches using sensors are limited by their cost and applicability for schools as the application of sensors is restricted to the data sets for which they were used (Baker et al., 2012). Baker et al. (2012) modeled the affective states of students from their interaction logs within the tutoring system of Cognitive Tutor Algebra I. They used the data available in PSLC DataShop to create a system for automatic detection of the affect, free from any kind of sensor. They used different Machine Learning algorithms to make sensor-free detectors for the four affective states, namely boredom, confusion, engaged concentration, and frustration. Based on their detectors, which performed better than chance, they suggested the features for each construct. For example, frequent guessing suggested boredom, asking for more hints indicated confusion, and making a wrong answer instead of asking for help signaled frustration. Their study lacked generalizability as the model which was developed represented a homogeneous population, yet their work shows EDM to be a useful research methodology for generating a student model.

EDM is a powerful tool, yet it is complex for educators to use and mostly beyond their scope (Romero & Ventura, 2010). With the vast number of EDM methods available, it is difficult to choose which one to use, and expertise is needed to find the right algorithm. Further, the tools available are very specific to the educational system it is designed for, there is no standard do-it-all tool, causing the issue of reusability. Therefore, a great deal of care should be taken when using EDM tools while designing any kind of assessment, specifically within the CAGE architecture as the assessment needs to be made content-agnostic.

4.3 Bayesian Nets and Student Model

A Bayesian Network (BN) approach uses probabilistic graphical modeling in which several variables have conditional dependence on each other (Friedman et al., 1997). A BN represents a graphical structure made up of nodes and directional links or edges between them. The nodes represent continuous or discrete variables, and the link represents conditional dependence between them. Each node in the structure has a probability distribution attached to it dependent on the nodes that have directed links flowing into it. These probabilities can either be learned from data or assigned by an expert rater.

García et al. (2007) used a BN in a web-based learning system to predict the learning style of students. This can help in presenting the information in a personalized way to students, compatible with their learning style, for enhanced learning (Felder & Brent, 2005). Students can learn and absorb knowledge in a variety of ways depending on their learning styles, and the teacher can present their content in various teaching styles (Felder & Silverman, 1988). The compatibility of these two greatly impacts student learning in a classroom. Incompatibility may lead to boredom, attention loss, and inferior performance. Learning involves obtaining the external information through senses, followed by its processing. Learning style identifies how a student takes in and processes information presented to them. Felder and Silverman (1988) introduced five dimensions, each having two attributes that can characterize the learning style of a student. They are sensing/intuition, visual/auditory, inductive/deductive, active/reflective, and sequential/global. Based on these dimensions, there are 32 possible learning styles. Trying to address all these styles in one class may appear daunting at first, but it is manageable.

The first dimension includes sensing and intuition which are the two ways in which students perceive information (Felder & Silverman, 1988). Sensing refers to using the human senses to collect the information around them while intuition is an indirect way of perceiving by using imagination and insights. Sensors prefer facts, standard methods, and details while intuitors like theories, innovation, and complications. Sensors tend to struggle with symbols as they are slow in translating them, while intuitors do not struggle. The second dimension, visual and auditory, are the two ways to receive information. People use either of the two and usually ignore the other. Visual learners best recall the things that they saw, while auditory learners remember most of what they heard. Visual is the most common modality among people. The third dimension, induction and deduction are the two ways to organize information. In induction, which is an innate learning style in humans, one observes the surroundings to make sense of the governing laws, while in deduction, which is the natural human teaching style, principles are used to create an understanding of surroundings. Inductive students require motivation, they need to see it to believe and appreciate the information presented to them. The fourth dimension, active and reflective are the two ways to transform the information received into knowledge. Active learning involves transformation through physical engagement or discussion active participation to grasp the knowledge, while reflective learning calls for introspective examination. The fifth dimension, sequential and global are two ways in which students move in a direction of greater understanding. Sequential learners take a step-by-step approach to solve a problem, while global learners take intuitive leaps. Sequential learners have good convergent thinking and analytical abilities, while global learners are good at divergent thinking and synthesis. Global learners usually struggle in a classroom where the teaching methods are mostly sequential. They are quite important to society, owing to their multi-disciplinary and creative abilities, hence should not be lost in the education process.

All the learning styles may not be relevant in a particular domain (García et al., 2007). Video games usually have audio as well as visual elements. This suggests that video games accommodate both the visual and auditory learning styles of students. Hence, detecting them is not very useful in the context of educational video games. Therefore, dimensions that are not applicable or are always present should be excluded. García et al. (2007) conducted a study in which they analyzed the interaction logs of students for creating a model of their learning style using BN. Using BN, they modeled only the dimension of perception, processing, and understanding, and the remaining two dimensions were discarded. Probability values for the nodes in the BN were initially assigned random or equal values, which were updated with the student's interactions within the system. Probabilities were updated until the difference threshold was reached, after which the state of the network corresponded to the student's model. The model predicted by the BN was compared with Felder's scale obtained using the Index of Learning Styles questionnaire (Soloman & Felder, 2005). Results of this study with 27 students yielded 77% accuracy in perception, 63% in understanding, and 58% in the processing dimension for predicting the learning styles with BN. This suggested that BN can be a powerful tool in modeling a student and it can help in adjusting the teaching style according to the learning style of a student. Moreover, BN does not require the participants to explicitly answer questionnaires and thus help maintain the game flow.

A Dynamic Bayesian Network (DBN) is a kind of BN in which variables have probabilistic dependence over a period called lag or time-steps, allowing for modeling sequences and time-series (Reichenberg, 2018; Reye, 2004). Figure 4 shows a simple DBN called knowledge tracing (Corbett & Anderson, 1994). The network shows a 2-quiz series that have four performance parameters and three nodes associated with it. These parameters are prior knowledge, slip rate, learn rate, and guess rate. The three nodes are: a participant node (S), a knowledge node (K), and a question node (Q).

The participant node represents an individual learner and governs the prior knowledge parameter P(L). The prior knowledge parameter describes the initial

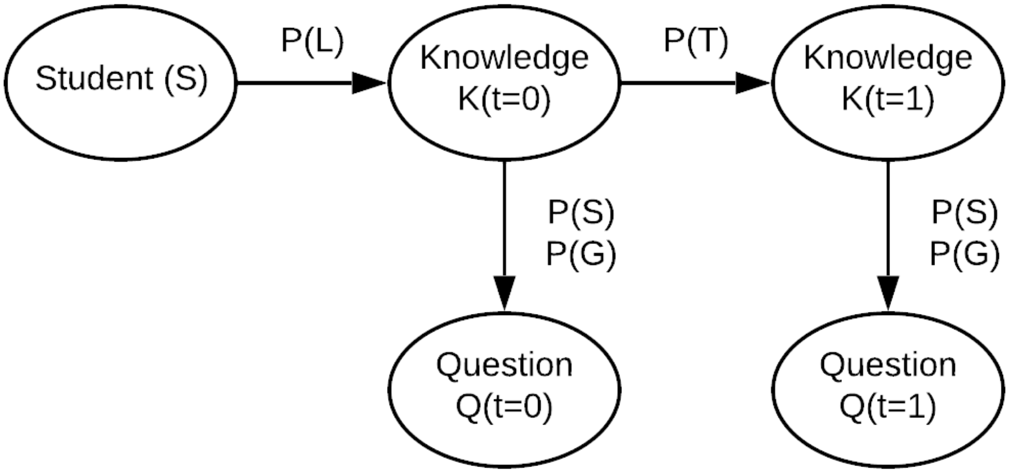


Figure 4. Bayesian knowledge tracing model showing two time slices.
Source: Pardos and Heffernan (2010)

knowledge level of a participant that they possess before playing the game. It can be obtained through a diagnostic assessment.

The knowledge node depicts the state of participant knowledge at any point in time and is dependent upon the prior knowledge that the participant has. In typical applications of knowledge tracing, it is a discrete node that has two states, namely true and false which represent the possible states of the participant having or not having the knowledge, respectively. The knowledge node is time-dependent, also called a temporal node, and is therefore replicated across both the time-steps. The knowledge node at any time-step is directly dependent on the knowledge node from the previous time-step. This conditional dependence of knowledge with itself based on time lag is represented using transition or learn rate P(T), which expresses the probability that a participant will transition from an unlearned to learned state in the next time-step. In the knowledge tracing model, it is typically assumed that it is not possible to lose knowledge on transition, and therefore the probability to go from a learned state to an unlearned state in the next time-step is zero (Corbett & Anderson, 1994).

The question node denotes the question which is asked to gauge the knowledge of the participants. It has two states, true and false, which corresponds to the participant's answer being correct or incorrect and is modeled as being dependent on the time-specific knowledge level of the participant. The question node is also temporal and therefore replicated across both the time-steps as shown in Figure 4. It has two associated parameters, guess P(G) and the slip rate P(S). Guess rate models the probability of guessing correctly when a participant does not have the knowledge, while slip rate accounts for answering incorrectly despite having the required knowledge. The current study employs a more complex model, in which there are multiple observables at each time-step, linked to the time-specific knowledge node.

4.4 Stealth Assessment

Stealth assessment is an unobtrusive ECD-based assessment technique embedded deeply within the game, utilizing the enormous data generated during the game play for inferring player performance at several grain sizes (Shute et al., 2010; Shute, Ventura, Small, et al., 2013; Ventura et al., 2014). It has been used for the assessment of creativity (Kim & Shute, 2015), persistence (Ventura et al., 2014), physics knowledge (Shute, Ventura, & Kim, 2013), problem solving skills (Shute & Wang, 2015), systems thinking (Shute, 2011), causal reasoning (Shute & Kim, 2011), and team performance (I. Mayer et al., 2013), in the past. ECD supports embedding the assessments holistically into the game play, with the main disadvantage being the associated high cost for enforcing a full-scale model (Crisp, 2014). A more befitting approach is to accommodate just the design framework of ECD instead of carrying out the full implementation, with key elements being the student, task, and evidence models.

Numerous factors are actuating the research on stealth assessments. The first one is the disproportionate change in the education system as compared to the rapidly changing and evolving world which causes student drop-outs from school (Shute et al., 2010; Shute, Ventura, Small, et al., 2013). Second, the need to assess higher-order thinking skills and evolve the assessment beyond multiple-choice questions, which are not enough to measure learning in complex scenarios. Third, video games employ non-cognitive skills like creativity and persistence (Shute et al., 2015) to excel in them, and thus it is easier to assess the performance of these competencies in a game rather than in a regular classroom. Fourth, stealth assessment in games maintains the flow of players and keeps them engaged in the content presented to them (Chen, 2007), unlike the survey questionnaires in which a student is interrupted for assessment followed by delayed feedback which is not of much use as the new learning has already started by that time (Crisp, 2014). This is one of the main goals of stealth assessment, to ultimately obliterate the line between learning and assessment (Shute, 2011). Fifth, stealth assessment considers the changes happening during the learning process, and thus provides a deeper insight into the abilities of a learner and their learning process (Eseryel et al., 2011). Sixth, it can provide real-time feedback to players and help adapt the game difficulty level to the player's skill and affective states (Baron, 2017; Shute, Ventura, Small, et al., 2013). Last, it can help alleviate test anxiety among students to a considerable extent, which is caused by traditional tests, enhancing student engagement (Shute & Wang, 2015).

To measure the understanding of physics principles among secondary school students, and to test the stealth assessment approach itself, Shute, Ventura, and Kim (2013) created a video game called Newton's Playground with assessment built into it. The game involved navigating a green ball to a red balloon using inanimate simple machines like levers which come to life once drawn, governed by gravity and Newton's three laws of motion. They used the logs generated by player interactions within the game for a summative assessment of the player's skills. Results indicated an overall improvement in the physics understanding of the students over time. However, the assessment could only help in inferring the physics understanding of students but not the knowledge of the formal language used in physics. In another study with Newton's Playground, Ventura et al. (2014) experimented with 70 students to measure their persistence levels. They used the player log data to find out the time expended on hard stages of the game, to assess their level of persistence. Anagrams and picture comparison tasks were used as external measures of persistence. Results indicated a significant correlation between the in-game measure of persistence and the external measures. There are several other studies employing the use of Newton's Playground to assess several constructs. However, creating a video game for measuring every skill, in every learning environment and embedding the relevant assessment into it is a cumbersome task, as there are a variety of skills and learning environments. Thus, it makes sense to create a universal assessment that can measure every skill in every learning environment (Shute, Ventura, Small, et al., 2013). This problem can be solved by creating a content-agnostic stealth assessment, which can adapt itself to content within any domain. However, it would require a lot of work to embed the assessment for all the domains in a single framework.

Some ethical concerns may arise when so much data is collected covertly during game play. Fairness of the assessment with respect to students should be considered, if they are evaluated without their information, then they are being deceived (Walker & Engelhard Jr, 2014). Laws regarding the storage of data, privacy, and anonymization should be taken into consideration, especially when the game is used internationally, where each country has its own set of laws (Freire et al., 2016). This becomes more critical when dealing with sensitive data, like grades and performance. Adoption of a clear ethics policy covering all the stakeholders is the key here.

4.5 Affect in Serious Games

Educational games are more interactive and engaging than traditional classroom material, hence they can help to sustain the intrinsic motivation of students (Amresh et al., 2014; Amresh et al., 2019; Freire et al., 2016). As stated by Aristotle, people seek personal happiness and pleasure (Bartlett, Collins, et al., 2011). Therefore, a player-oriented educational game would evoke players' positive feelings, if used with a personalized adaptive design.

Ekman (1999) identified anger, disgust, fear, happiness, sadness, surprise, and contempt as seven basic emotions that are universally experienced across cultures. These seven basic Ekman emotions are simply referred to as emotions in this dissertation. Even though these emotions are universally experienced across cultures, they are not as functionally useful within a learning context, as they do not occur frequently during the learning process (Craig et al., 2008; Russell, 2003).

Boredom, flow, and frustration were found to be more useful to predict learning than the Ekman emotions (Craig, Graesser, et al., 2004; Pekrun et al., 2002). Pekrun et al. (2002) found that academic learning had high correlation with boredom, flow, and frustration. They indicated the extent to which these affective states correlated with the study interest, learning strategy, irrelevant thinking, and self-regulation during the learning process. Therefore, they were chosen as the affective states of interest for this dissertation.

Ekman and Friesen (1978) adopted a taxonomic system known as the Facial Action Coding System (FACS) which categorizes basic emotions based on the movement of muscles on the face. The movement of muscles is called an Action Unit (AU) in FACS. Although AUs can be coded to identify the basic emotions, efforts that try to predict the cognitive-affective states (affective states) of boredom, flow, and frustration have been carried out in the past (Craig et al., 2008).

Among the new non-intrusive approaches, the focus of current research is facial emotion tracking using Affdex Software Development Kit (SDK) from Affectiva. This approach utilizes a webcam, which is easily accessible hardware, to record the changes in facial features using the facial feature detection SDK that quantifies the changes so that they can be used to assess the emotional states of users. Provided that the tracking environment is set up correctly and the user is front-facing the camera, it is possible to achieve a relatively high detection rate for facial features (Magdin & Prikler, 2018), which then can be used to predict the basic emotional states of users. While research has provided successful results in the detection of facial features, only a few methods exist that can accurately predict the affective states and provide these as inputs to curate and personalize the user experience (Harley, 2016; Tadayon et al., 2018). Such efforts have high value to users in fast-paced interactive and immersive environments, as real-time adaptation becomes critical to the success of such environments. As seen with the increase in the use of online and interactive methods to improve educational outcomes during pandemics such as the COVID-19 crisis (Zhou et al., 2020), the need to build robust, scalable, and multi-setting methods to accurately measure and adapt based on affective states of the users becomes paramount.

4.6 Adaptability in educational games

According to the current literature, no real proof exists which indicates adaptive educational games to be better than non-adaptive ones. Limited and contradictory research exists that compares the two. In a study, Sampayo-Vargas et al. (2013) used the responses to game objectives to manipulate the game difficulty. In the adaptive version, they decreased the game difficulty for incorrect responses, and increase it for correct responses. They found better learning outcomes as a result of adaptation. In another experiment, Holmes et al. (2009) assessed the impact of game adaptation on the performance of the working memory. To keep the players at the edge of their working memory limits, game difficulty was matched to their performance which led to significant improvement in their working memory and mathematical abilities. van Oostendorp et al. (2014) also found the adaptive version to be better in terms of learning gains. Their game adapted to a complexity level that was governed by the player's previous scores. Likewise, Ali and Sah (2017) found the better and faster performance of participants when the game was adapted based on the current knowledge level of the player.

Contrarily, there exists some research studies that say otherwise. In a recent study by Vanbecelaere et al. (2020) no impact of game adaptation was observed for the cognitive and non-cognitive factors. Their game consisted of several exercises whose number was based on the player's performance in former exercises. The adaptive version also had a threshold of 65% score needed to clear the current level and process to the next, which was not present in the non-adaptive version of the game. Similarly, Shute et al. (2020), found no substantial difference on participant learning due to adaptation. They manipulated the order of game levels based on an algorithm. Orvis et al. (2008) and Plass et al. (2019) affirm these results and found no impact of adaptation.

Likely based on how the adaptation is built into the game, there is conflicting evidence to the utility of game adaptation. Even though D'Mello et al. (2010) investigated the part that affect plays in an Interactive Tutoring System, there is very finite research that examines the role of affect in serious game adaptation. D'Mello et al. (2010) found affect-sensitive systems to be more useful for a learner who possess lower domain knowledge as compared to the high knowledge learners. Current research is intended to expand on this finding in the context of an educational video game.

Chapter 5

CONTENT AGNOSTIC GAME ENGINEERING

Usually, the game mechanics are tightly tied to the educational content being taught by the game, which renders the programming code of the game unusable for further development (Baron, 2017). Therefore, it requires a major overhaul of the game program for future projects, and often the code is discarded as starting over is more cost and time-efficient. CAGE is a model for designing educational games which alleviates this problem by separating the game mechanics from the educational content of the game. This is beneficial for both industry and academia as it will help in the rapid creation of educational games and savings in terms of time and money. Only the first game project will require full-scale expenses, all the subsequent games can be rapidly developed by re-using the code of the first game.

5.1 Current model

The current model for game-based learning is shown in Figure 5 (Baron, 2017). In this model, the player inputs the commands via an input device. These commands are passed to the mechanics component, translating them into in-game actions. For example, the player presses the S key, moving their game character backward. This action is passed to the content component, which will evaluate if appropriate. The content component after evaluating the action as right or wrong, passes the result of the assessment along the loop where it appears as feedback to the player. The player

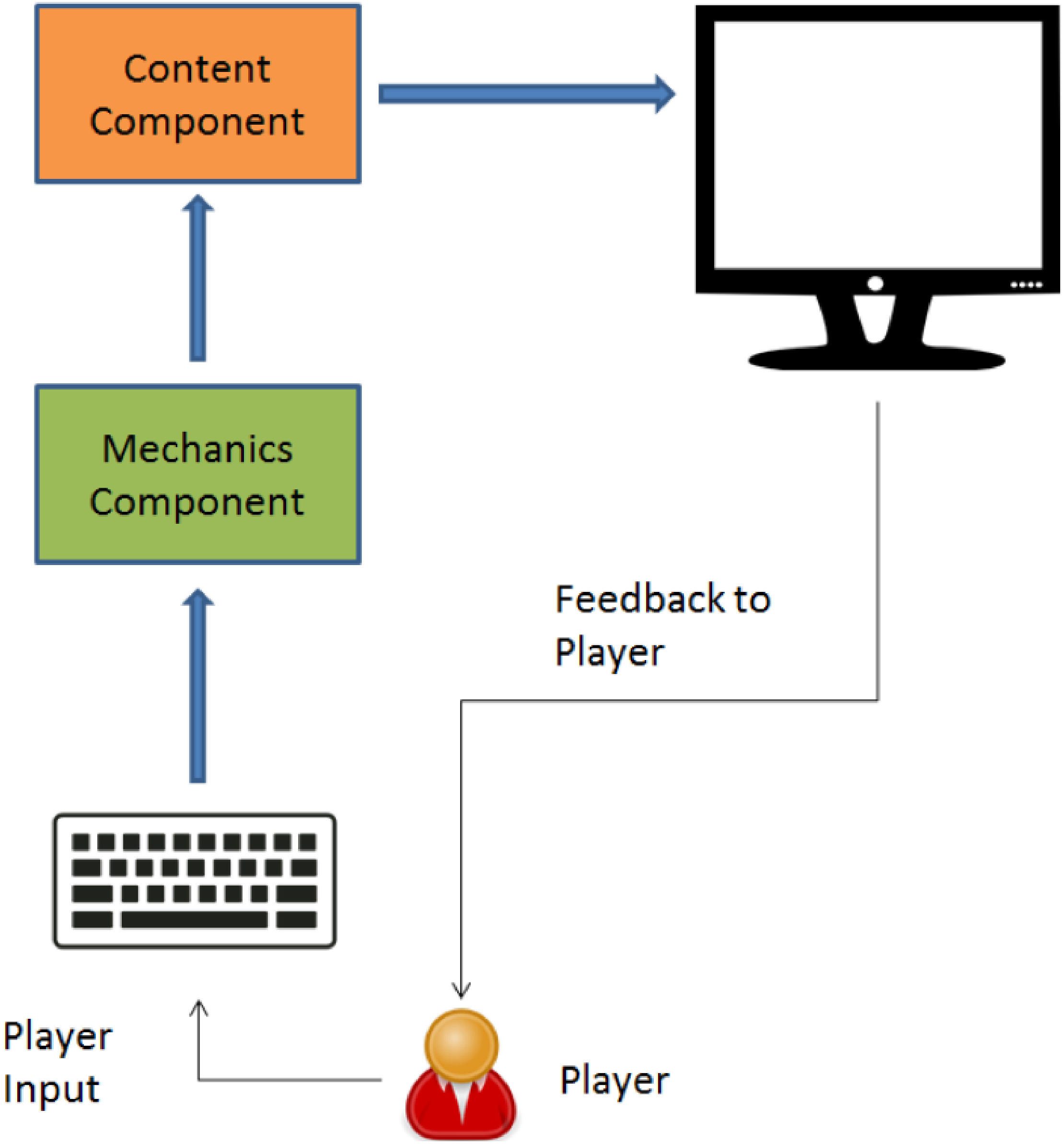


Figure 5. Model currently in use for educational game development.

Source: Baron (2017)

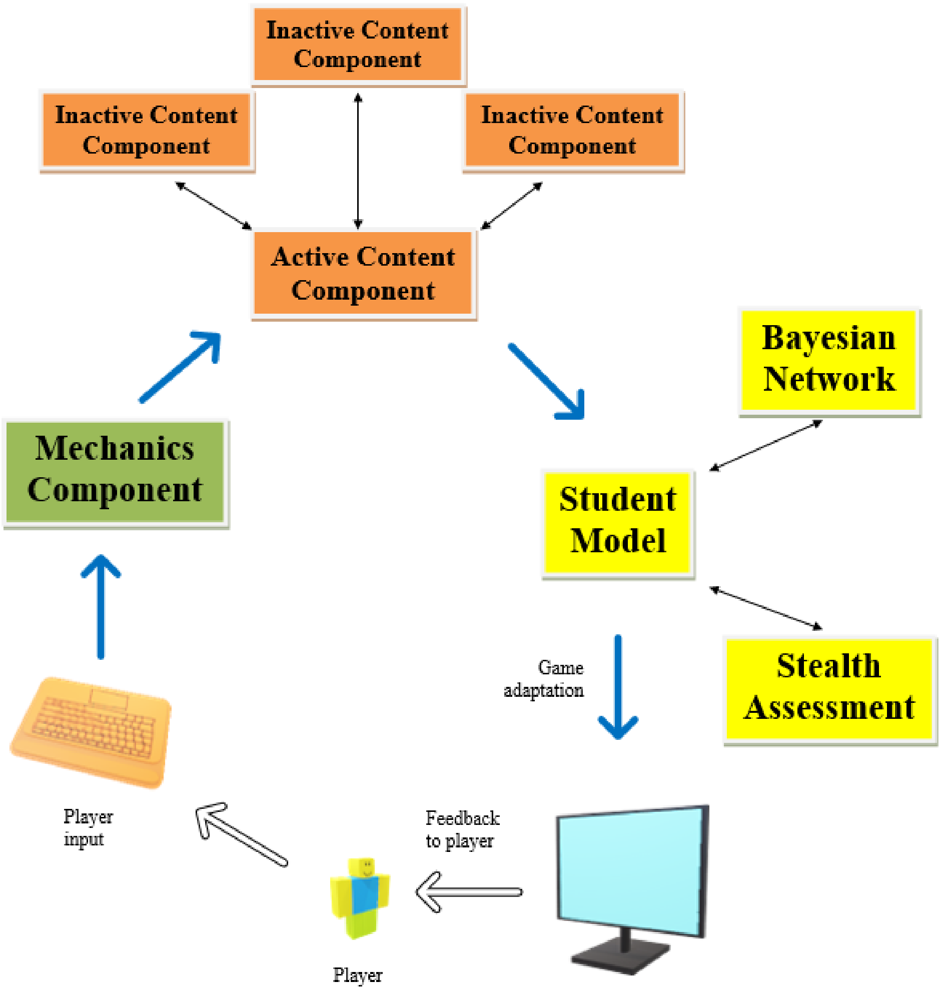


Figure 6. CAGE Model for educational game development.

Source: Baron (2017)

upon seeing the feedback, acts accordingly, which incorporates the feedback in their next action.

5.2 The CAGE model

The CAGE model in Figure 6 is similar to the model in Figure 5 with a few changes (Baron, 2017). It follows the ECD approach of assessment and a component-based architecture. A new step for the student model is added in the framework, which will be updated with every student action, and may not be changed when the game switches content. The student model can be built using multiple techniques such as BNs and player log tracking. The content component is the part that will be switched for another content. CAGE has several content components instead of one, but only one of them will be active at a time. CAGE is composed of the following four components: framework, the mechanics component, the content component, and the student model.

The framework is a static part responsible for gluing the components together (Baron, 2017). It connects player input with game mechanics, which is linked to the content component. Evaluation from these components updates the student model, which passes the feedback to the player through the framework. The mechanics component upon receiving the input from the player, interprets it into a corresponding action in the game. In CAGE, this component is designed to be content-agnostic, independent of the content being taught by the game, giving CAGE its name. The content component evaluates the action to update the student model and pass the corresponding feedback to the player. This component is dynamic and can be switched for teaching different content using the same game mechanics. The student model is the one that corresponds to the state of knowledge of a student at any point in time during the game play. It should be pliable enough to be able to assess any domain.

5.3 CAGE outcomes

Baron (2017) conducted a study with eleven students from a graduate-level course in game-based learning for the development of the CAGE framework. Participants were asked to develop a game from scratch, in two weeks using the CAGE architecture. On completion of their first game, they were asked to make another version of the game using CAGE, but this time with different content and within a week. Upon finishing the second version of the game, participants were administered a questionnaire regarding the process. The results indicated a natural tendency of the participants for re-using their code. It was found that CAGE led to the reduction of the number of lines of code and hours spent from the first game to the second game creation. Development of the second game consumed less than half the time and reduced the amount of code needed by two-thirds on an average compared to the first game, speeding up the entire process. This study addressed the first issue mentioned by Tüzün (2007) regarding the game design. However, the study suffered from external validity issues owing to the small sample size, and absence of teamwork as the task was to be performed individually (Baron, 2017).

5.4 CAGE and engagement

In another study conducted by Baron (2017) regarding the effectiveness of CAGE using the eleven games created using the CAGE framework, participants were required to play two versions of the game chosen randomly. They were surveyed using a questionnaire at the end of each game to assess their cognitive load and engagement after the completion of the game. Results indicated a decrease in cognitive load from the first game to the second one, irrespective of the order in which the two versions were played, which is desirable. This can be attributed to the same mechanics being used for both games, eliminating the need to orient a student multiple times towards the game play, addressing the second and fourth issues reported by Tüzün (2007) regarding the game design. Further, it was found that engagement levels decreased from the first game to the second one, regardless of their play order, owing to the same game mechanics being employed in the second version of the game (Baron, 2017). This is something that needs to be worked upon if the CAGE games are to be employed in a regular classroom. In this study, Baron (2017) conducted a questionnaire in between the two games. This interruption can break the participant flow, and they may lose motivation during the learning exercise. To deal with this issue, stealth assessment should be embedded into the game play to maintain the learner's flow and sustain their engagement in the learning process.

5.5 CAGE and affect

Baron (2017) conducted another study with CAGE and affect detection, to find the effect of dynamic difficulty on engagement and cognitive load, using another game that was made specifically for this purpose. The study involved seventeen graduate-level students in a game engine architecture class, who were required to play the two versions of this game in random order, like Baron's (2017) previous study regarding CAGE effectiveness. It was found that the affect detection had no effect on the cognitive load or engagement either, contrary to what was observed by Baker et al. (2010). However, this study suffered from external validity issues owing to the small number of participants, and lack of the calibration of the affect detection (Baron, 2017).

CAGE can further be used for plugging the assessment into the game, independent of the game mechanics (Baron, 2017). This will subside the problem of creating an assessment for every content and mechanics of the game, and thus reduce the assessment re-design effort for each version of the game. However, one limitation of CAGE is that it did not implement the student model. This dissertation intends to expand upon the CAGE architecture for the creation of the student model independent of the content being taught by the game.

Chapter 6

RESEARCH QUESTIONS

This dissertation seeks to answer the following five research questions.

6.1 Validity of Dynamic Bayesian Network

The first research question is stated as follows: "Is there evidence of validity for the use of Bayesian networks to model learner beliefs in the CAGE based games"? It seeks to explore the validity of the DBN when applied to the assessment of student knowledge in a CAGE based game. Previous work (García et al., 2007) has shown BNs to be able to precisely determine the learning styles of a student in a web-based course to create a student model. Based on the previous model (Pardos & Heffernan, 2010), the current research used a DBN as a form of stealth assessment to model the student's knowledge and provide necessary remediation if required. However, the validity of the DBN needs to be established before it can be used to create a student model because there are multiple network structures that could be used to implement a DBN within a game. A given network structure while seemingly plausible, may not be valid. In the current study, results from the DBN were compared with an external measure of the student skill. A post-test for the game contents was used to validate the output of the DBN.

6.2 Affect adaptation in games

The second research question is stated as follows: "Does game adaptation using affect assessment help in improving the learning and engagement of the player"? The present literature does not explicitly demarcate student learning due to adaptation within a serious game. The effectiveness of game adaptation has been found to vary depending on the way it was implemented in the serious game (Orvis et al., 2008; Sampayo-Vargas et al., 2013; Shute et al., 2020; van Oostendorp et al., 2014). Further, previous research (Pardos et al., 2014) has shown that the affective states can be used to predict the learning outcomes. Therefore, affect detection was used to adapt the game play with an aim to provide interventions that can be used to alleviate the negative affective states and promote the affective state that can provide better learning outcomes. It was hypothesized that adapting the game using the player's affect will help in improving player engagement and learning. To investigate this, the current research used a game that had embedded affect detection built into it. Two groups were used, the test group with the dynamic game adaptation and the control group devoid of the adaptation capabilities. The results from the two were compared to evaluate the hypotheses.

6.3 Multi-modal adaptation vs. learning

The third research question is stated as follows: "Does adding stealth assessment based adaptive game design improve learning"? The previous research question investigated the impact of adapting the game using affect detection only. However, there are various other ways that could be used to stealthily assess the state of a player and provide a suitable intervention (Verma et al., 2019). As stated earlier, the existing literature does not agree on the effectiveness of the adaptation for providing better learning performance. Therefore, the current research implemented further adaptative interventions which used player log data and DBN in addition to the affect assessment. It was hypothesized that using a multi-modal adaptive technique will help in improving student learning. A test group having these multi-modal adaptive abilities and a control group that doesn't were employed. A pre-test and post-test were administered using the external measure of skill to investigate this research question.

6.4 Multi-modal adaptation vs. engagement

The fourth research question is stated as follows: "Does adapting the game using stealth assessment enhance the engagement of the players"? Similar to the learning outcomes, the engagement is expected to be dependent on the way the adaptation is built into the game. Very limited research exists that compares engagement with and without game adaptation. Sharek and Wiebe (2015) found that the engagement did not change significantly due to adaptation. They adapted the game to decide the subsequent levels of the game that the player should play, but the adaptation played no role in adapting the game within a level. The fourth research question employed the multi-modal technique of adaptation which was used to enhance learning and hypothesized that it will help in enhancing student engagement. To examine the hypotheses, the UESz scale (Wiebe et al., 2014) was utilized. Two groups were used for the purpose, a test group with multi-modal adaptive capabilities and a control group devoid of it.

6.5 Multi-modal adaptation vs. learning and engagement in CAGE

The fifth research question is stated as follows: “Does CAGE with adaptation help in sustaining student engagement and promote learning performance”? There is no research at present that evaluates the learning outcomes in a CAGE game. CAGE has been shown to speed up the development process of an educational video game leading to savings in terms of time and money (Baron, 2017). However, it leads to reduced engagement on playing the subsequent games that have different content but use content agnostic mechanics. Lower engagement then results in poorer learning outcomes (Halm, 2015; Park, 2003). It was hypothesized that using a multi-modal adaptation in a CAGE game will help sustain student engagement and promote learning performance. A 2×2 factorial design experiment was used to evaluate these two hypotheses regarding learning and engagement. Content order and adaptivity were the two independent variables of interest. Content order had two factor levels which were used to denote which content was played first in the CAGE game. Adaptivity had two levels that signified the presence or absence of the multi-modal game adaptation. Pre-and-post tests were used to evaluate the hypotheses regarding learning and UESz scale for engagement.

Chapter 7

CHEM-O-CRYPT

Content Agnostic Game Engineering (CAGE) architecture (Baron, 2017; Baron et al., 2016) was used to create a 2D platformer game called "Chemo-o-crypt" in Unity3D (v2018.1.9f2). A recent pilot-study (Atmaja et al., 2020) showed the advantages of using CAGE in higher education, especially among undergraduate students. Recent years have indicated a rise in educational games that follow this architecture as the economics of tightly connecting the content to the design make it inefficient and time-consuming for educators (Baron et al., 2016).

7.1 Game Mechanics

In Chemo-o-crypt, the game mechanics allowed left and right player movement, ladder climbing, and jumping. There were three different types of patrolling enemies which reduced a partial portion of the player's health on collision. There were also two types of environmental hazards, which were spikes and water, shown in Figure 7. It would reduce the available player health to zero when they fell into these hazards. Also, these penalties were determined based on the game difficulty that ranged from one to four. For example, full life was reduced if a player collided with an enemy when the game difficulty was set at five, but only 25% of health was reduced if the difficulty level was set at one. The game could be played either for chemistry or cryptography content learning. Each content had four levels, which were distinct from the game difficulty levels. Later levels featured moving platforms, which were

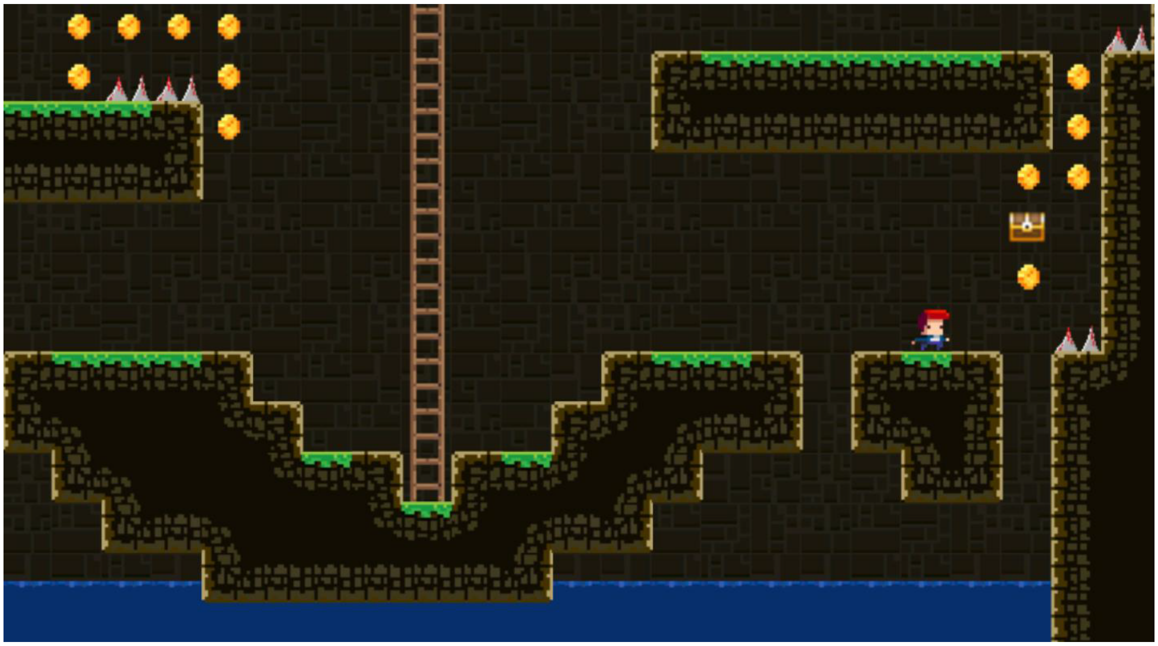


Figure 7. Screen capture showing the spike and water hazard in Chem-o-crypt.

either moving by default or started moving when a player jumped onto them. The moving direction could be horizontal or vertical. There were coins and heart-shaped items (1-up) scattered throughout the game map. A player initially had three lives which could be increased by collecting one hundred coins or a 1-up.

Each game level in Chemo-o-crypt was divided into 4 navigable chunks that lied next to each other in a sequence. Governed by the game difficulty, every chunk held a game scene in it. Consequently, each chunk could have four possible scenes that it could be populated with. Therefore, there were 4 × 4, i.e. 16 maximum possible layouts for the game level environment at any point in time which was dependant on the game difficulty. A player could easily move between the chunks as they were continuous, but only if the player avatar was on the ground level. During the first content level, players spawned in the first chunk, they spawned in the second chunk for the second content level, and so on.

For the adaptive version of the game, the layout of a given chunk only changed when the player crossed a chunk boundary that was not adjacent to that chunk. For example, if the player is moving from chunk 2 to chunk 3, then the environment layout for chunk 1 and chunk 4 may change but not for chunk 2 and chunk 3. Similarly, when they are moving from chunk 3 to chunk 4, the layout may change for chunk 1 and chunk 2 since they are not located next to this boundary. This was done to avoid the distortion of the gaming world in front of the learner's eyes. However, this layout change depended on the game difficulty only. If the adaptive algorithm determined that the game difficulty should increase when a player moved from chunk 1 to chunk 2, then the layout for chunk 3 and chunk 4 will change corresponding to that difficulty level. Screenshots of the possible layouts of all four chunks are available in Appendix K.

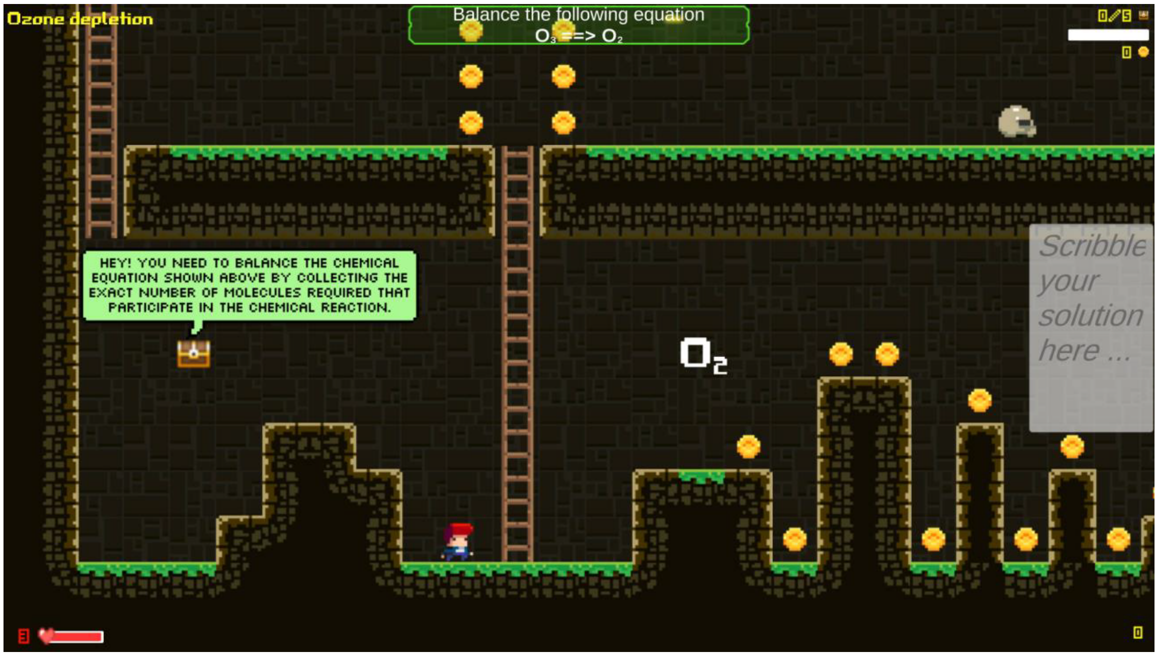
7.2 Game Content

For the chemistry version of the game, players were required to collect the correct number of elements and molecules that take part in the chemical reaction to balance it. Consider the chemical equation represented in Figure 8, it required 3 Oxygen (O_2) and 2 Ozone (O_3) molecules to balance this equation. However, there were be 3 distractors present in the game environment, which were the excess of these molecules. For example, for the equation shown in Figure 8, more quantity of Oxygen or Ozone than needed would act as a distractor. This was done to make the game more challenging and to keep in check if the players were collecting everything instead of collecting only the required quantities. All the collectible elements were initially displayed in a static white color whether they were distractors or not.

However, distractor elements become red when picked up and the rest were displayed in a glowing green color indicating that they were not distractors. The player received a kickback and possible health loss depending on the content level they were playing on picking up a distractor. This behavior was in accordance with the operant conditioning, as it punished the player for collecting distractors (Skinner, 2019). When the player came in the proximity of a collectible, it randomly became either a required molecule or a collectible with an equal probability. The "GO" (completion text) text appeared once all the required molecules were collected. However, if there were some distractors that were not yet collected, then there was a 50% chance that the completion text would show up and 50% chance that the distractor would be displayed. When the player collected the completion text the same equation appeared (as a quiz) which they had balanced with the help of game play mechanics (Figure 9). On hitting the submit button, the next content level was loaded irrespective of the wrong or right answer. However, they were given 1 more attempt before submitting if the answer was wrong. The game consisted of four content levels, each having its own background music that got more intense as the player moved to higher content levels. The balanced equation for each content level is enumerated below:

1. $2\,O_3 \longrightarrow 3\,O_2$
2. $N_2 + 3\,H_2 \longrightarrow 2\,NH_3$
3. $ZnS + 2\,HCl \longrightarrow ZnCl_2 + H_2S$
4. $Al_2O_3 + 6\,HCl \longrightarrow 2\,AlCl_3 + 3\,H_2O$

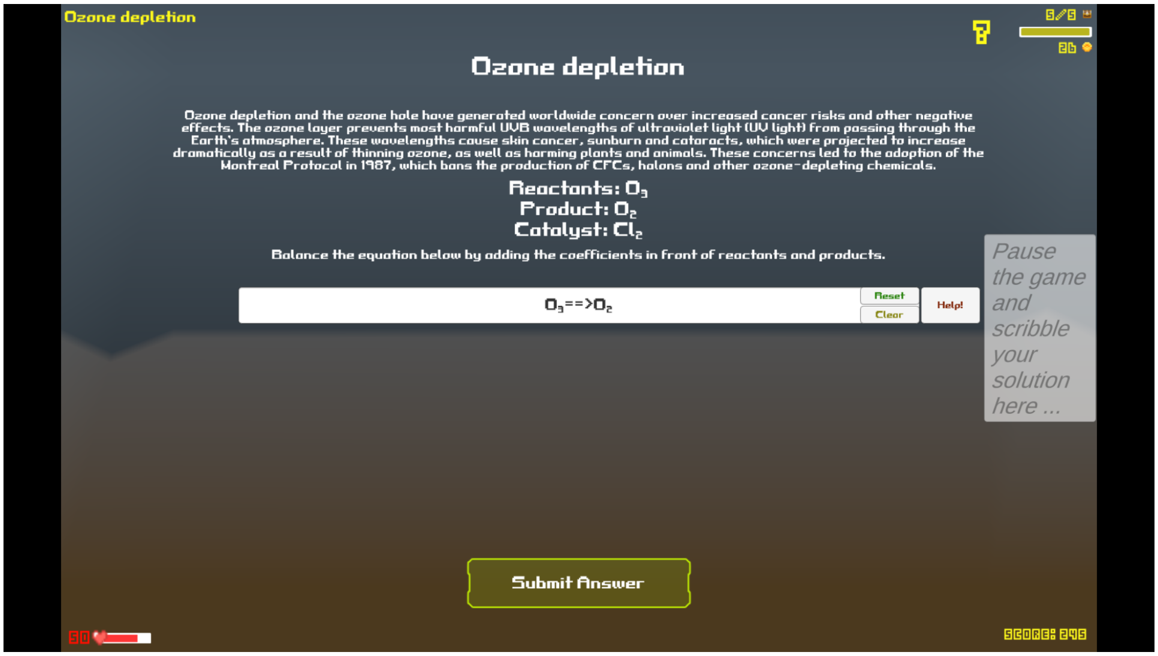
For the cryptography version of the game, each content level aimed to encode or decode a piece of text using the encryption key provided to the player. Similar to the chemistry version, there were either different or excess letters present that would act



Ozone depletion
Balance the following equation
O_3 ==> O_2
0/5
0
O_2
Scribble your solution here ...

HEY! YOU NEED TO BALANCE THE CHEMICAL EQUATION SHOWN ABOVE BY COLLECTING THE EXACT NUMBER OF MOLECULES REQUIRED THAT PARTICIPATE IN THE CHEMICAL REACTION.

Figure 8. Screen capture of the goal of the game Chem-o-crypt.



Ozone depletion

Ozone depletion and the ozone hole have generated worldwide concern over increased cancer risks and other negative effects. The ozone layer prevents most harmful UVB wavelengths of ultraviolet light (UV light) from passing through the Earth's atmosphere. These wavelengths cause skin cancer, sunburn and cataracts, which were projected to increase dramatically as a result of thinning ozone, as well as harming plants and animals. These concerns led to the adoption of the Montreal Protocol in 1987, which bans the production of CFCs, halons and other ozone-depleting chemicals.

Reactants: O_3
Product: O_2
Catalyst: Cl_2

Balance the equation below by adding the coefficients in front of reactants and products.

O_3==>O_2

Reset
Clear
Help!

Submit Answer

Figure 9. Screen capture of the level-end task or quiz which appeared on collecting the completion text.

as a distractor. The task and its corresponding solution for each content level are listed below:

1. Encrypt the Plain Text: “ATTACK AT DAWN” using the Key: 2
   Resulting encryption = “CVVCEMCVFCYP”
2. Decrypt the Cipher Text: “EFGFOE UIF DBTUMF” using the Key: 1
   Resulting decryption = “DEFENDTHECASTLE”
3. Encrypt the Plain Text: “PURA VIDA” using the Key: 13
   Resulting decryption = “CHENIVQN”
4. Decrypt the Cipher Text: “URON RB KNJDCRODU” using the Key: 9
   Resulting decryption = “LIFEISBEAUTIFUL”

7.3 Affdex Software Development Kit

Affdex Software Development Kit (SDK) from Affectiva (Magdin & Prikler, 2018) was integrated into the Chemo-o-crypt game. SDK tracked the facial features of the players to output the probabilities for their emotions with a sampling rate of 20 Hz. A template size of 640 by 480px (height by width) was used to capture their face. When the player moved out of the field of view of the camera, then the game paused itself, asking the player to re-orient themselves so that the camera could detect their face. SDK traced the seven basic Ekman emotions of Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Contempt, in real-time. The output probability ranged from 0 (emotion absent) to 100 (emotion fully present). The SDK also tracked the physical properties of 15 different facial features (facial expressions) which included Attention, BrowFurrow, BrowRaise, ChinRaise, EyeClosure, InnerBrowRaise, LipCornerDepressor, LipPress, LipPucker, LipSuck, MouthOpen,

Table 2. Description of expressions obtained from Affectiva's Affdex SDK

Expression	Description
Attention	Measure of point of focus of the user based on the head position
BrowFurrow	Both eyebrows moving lower and closer together
BrowRaise	Both eyebrows move upward
ChinRaise	The chin boss and the lower lip pushed upwards
EyeClosure	Both eyelids closed
InnerBrowRaise	The inner corners of eyebrows are raised
LipCornerDepressor	Lip corners dropping downward (frown)
LipPress	Pressing up the lips together without pushing up the chin boss
LipPucker	The lips pushed forwards
LipSuck	Pull of the lips and the adjacent skin into the mouth
MouthOpen	Lower lip dropped downwards
NoseWrinkle	Wrinkles appear along the sides and across the root of the nose due to skin pulled upwards
Smile	Lip corners pulling outwards along with other cues from the face (e.g. eyes) combine to indicate a true smile
Smirk	Left or right lip corner pulled upwards and outwards
UpperLipRaise	The upper lip moved upwards

Source: iMotions Inc. (2018)

NoseWrinkle, Smile, Smirk, UpperLipRaise. These expressions correspond to the AUs from the Ekman and Friesen's Facial Action Coding System (Ekman & Friesen, 1978). A description of these expressions obtained from iMotions website (iMotions Inc., 2018) is available in the Table 2.

7.4 Student Model

Figure 10 shows the DBN employed in the game Chem-o-crypt. Bayes Server 8.17 (BayesServer, 2020) was used to create the network. The network consisted of eleven nodes, which included five temporal or time-series nodes. Table 3 presents a description of each node. The prior knowledge of a student was modeled using the

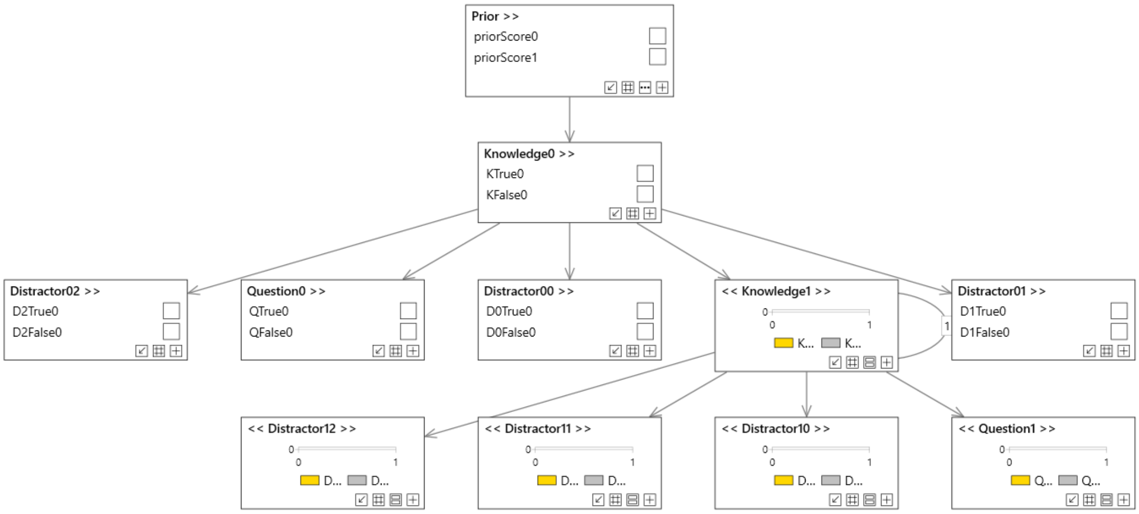


Figure 10. DBN implemented in the game Chem-o-crypt

Prior node. A pre-test that consisted of twenty questions was used to gauge the prior knowledge and assign the value to the Prior node. For example, a score of 13 on the pre-test (i.e. 65%) would mean that the Prior node will be assigned a state of priorScore6 out of the 11 [0-10] states that it could possibly have. This would then serve as a piece of evidence for the latent node Knowledge0. Knowledge0 node expressed the probability that the student possessed the required content-specific knowledge at the time-step $t = 0$, i.e. content level 1. As an example for the cryptography content, it represents the probability that the student knows how to use Caesar cipher for encoding a piece of textual data. Knowledge0 consisted of two states, true and false, which indicated the presence or absence of required knowledge.

Distractor00, Distractor01, Distractor02 symbolized the probability that the participant picked up the three respective distractors on the 1st content level ($t = 0$). They consisted of two states, true and false, which corresponded to the events that the distractors were collected or not. For example, a value set of {True, True, and False}

Table 3. Description of nodes in the DBN.

Node name	Type	Conditional dependency	States	Description
Prior	Initial ($t = 0$)	None	11 states indicating score in the range [0–10]	A state of 0 denotes a student who scored a 0 in the pre-test, while a state of 10 represents a student who scored 100% in the pre-test
Knowledge0		Prior	True and False	State of true denotes the possibility that the student has the required knowledge at timestep $t = 0$
Distractor00		Knowledge0		True denotes the evidence that the student has collected this distractor at timestep $t = 0$
Distractor01				
Distractor02				
Question0				True denotes the evidence that the student answered the quiz correctly at timestep $t = 0$
Knowledge1	Temporal ($t = 1, 2, 3$)			True denotes the possibility that the student has the required knowledge at timestep $t = 1, 2, 3$
Distractor10		Knowledge1		True denotes the evidence that the student has collected this distractor at timestep $t = 1, 2, 3$
Distractor11				
Distractor12				
Question1				True denotes the evidence that the student answered the quiz correctly at timestep $t = 1, 2, 3$

would mean that only the first two distractors were collected. Question0 denoted the probability that the participant answered the level 1 quiz correctly. A state of false would mean that they answered it incorrectly and true would mean the opposite.

As shown in Figure 10, the Knowledge1 node was linked to itself with a temporal order of 1, which meant that knowledge on the next level (current time-step, $t = t$) depended on the knowledge acquired in the previous level ($t = t - 1$). Knowledge1

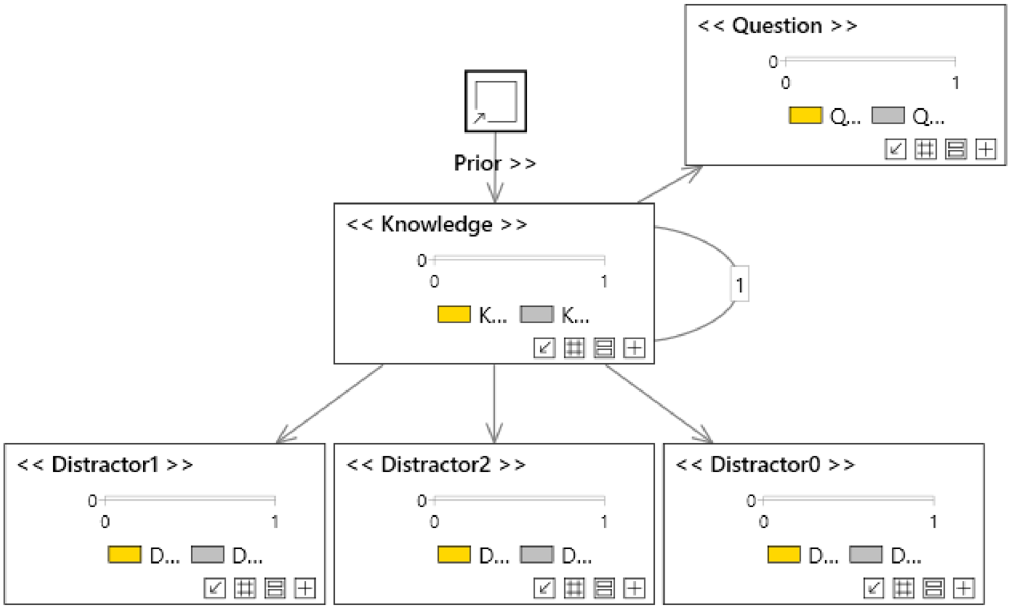


Figure 11. A simplified version of the DBN used in the game.

represented the probability of having knowledge at the content levels 2, 3, and 4 (i.e. time steps $t = 1, 2, 3$). Distractor10, Distractor11, Distractor12 represented the probabilities that the students picked up the three respective distractors on later levels ($t = 1, 2, 3$). Question1 denoted the probability that the student answered the level 2, 3, and 4 quizzes correctly or incorrectly.

Figure 11 depicts a simplified version of the DBN shown in Figure 10. Except for the difference that the former does not show the content level 1 nodes (for $t = 0$), the two are similar. The missing nodes were merged in their temporal counterparts. As a result, the probability of knowledge, answering the questions, and picking up the distractors was the same across all four content levels. Although they were different, by design, for the complex version of the network from Figure 10. The probabilities were distinct for the first level ($t = 0$), compared to the rest of the levels ($t = 1, 2, 3$). This distinction was deliberately done to allow for the students to learn the mechanics of the game during the first content level. Initially, the students did not know that they were supposed to avoid distractors. In doing so, they got a kickback and a health loss which depended on the content level. Therefore, at content level 1, while they were learning the mechanics of the game, the probability of picking up the distractors is expected to be higher as compared to the rest of the levels.

Initially the DBN is not fully specified since the joint probability distribution of the nodes are not known because of the absence of any prior data. It is required to determine the probability distribution, called parameters, for each conditional upon its parents. These probabilities can either be assigned with the help of an expert rater, obtained using parameter learning, or assessed using a combination of both approaches (Neapolitan, 2004). Parameter learning is a process in which the conditional probability distributions are learned with the help of data. Current research will involve collecting the data to learn these probabilities which will be used for subsequent experiments to better adapt the game with the help of DBN and evaluated the research questions posed in the first chapter of this document.

7.5 Learning design

The Chem-o-crypt game employed several learning design principles to create an educational game that was directed towards achieving a balance between the learning and entertainment aspects of serious games. These learning principles are described in detail below.

7.5.1 Contiguity effect

The game followed the contiguity principle, which states that the learning material should be designed in a way that keeps the related elements and ideas temporally and spatially close to each other (R. E. Mayer, 2005). The feedback from collecting a distractor appeared immediately when it is collected, at the same location where the distractor element was present.

7.5.2 Perceptual-motor grounding

The game grounded the new concepts in perceptual-motor experiences in the beginning (Glenberg & Kaschak, 2002). For example, in the chemistry content, a list of all the chemical elements used in the game was introduced to the player before they could start playing the game. This made sure that they know that 'Al' stands for aluminium, and is a single element that would need balancing. If this grounding were lacking, they may assume that they are two disparate elements, and need to be balanced separately.

7.5.3 Dual code

Learning material was designed to deliver the instructions in multiple modes (R. E. Mayer, 2005). The content was presented as text with supporting images that explained the content. This is accomplished using rotating disks to illustrate the use of Caesar cipher. For the chemistry content, images were used as examples that took them through the balancing process step-wise.

7.5.4 Testing effect, spaced effect, and negative suggestion effect

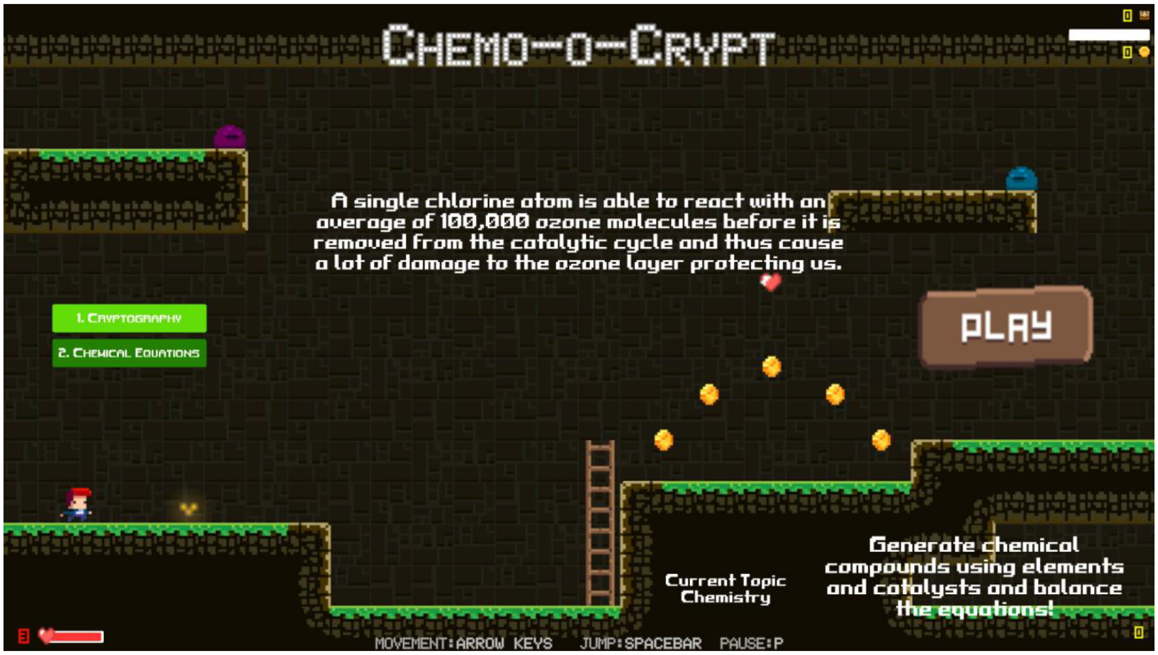
Players were frequently tested on their knowledge and the task that they accomplished during the content levels with the help of a quiz (Roediger III & Karpicke, 2006). These tests were spaced equally and appeared at the end of each content level. This allowed for long-term retention of the information in memory (Cepeda et al., 2006). The testing was followed by immediate feedback with the help of a star rating system that was rewarded at the end of each content level, along with a textual response that indicated their response to be correct or incorrect. This helped in reducing the negative suggestion effect by recalling the incorrect responses (Roediger III & Marsh, 2005). The testing results were also used to dynamically determine if they required remediation during the game.

7.5.5 Segmentation principle

Segmentation principle was used to present the information in the reading material. The content was presented in discrete chunks to prevent the learners from getting overwhelmed during the process (R. E. Mayer & Moreno, 2003). Appendix C and D show the information that was placed in each chunk that was presented to the user.

7.5.6 Deep questions

The tests consisted of deep questions which facilitated comprehension or learning material (Craig et al., 2006). For example, questions 16 and 17 of the cryptography



A single chlorine atom is able to react with an average of 100,000 ozone molecules before it is removed from the catalytic cycle and thus cause a lot of damage to the ozone layer protecting us.

Generate chemical compounds using elements and catalysts and balance the equations!

Figure 12. Screen capture of the main menu showing a fun fact about Chemistry.

pre-test (Appendix H) are deep questions. They promoted deeper understanding by asking what went wrong during the encryption process, rather than directly asking them to encrypt information.

7.5.7 Gagne's Taxonomy of Learning

Along with the principles stated above, Gagne's nine principles (Gagne et al., 2005) were used to scaffold the learning in the game Chem-o-crypt. It consists of a sequence of steps that need to be followed to facilitate learning. These steps are detailed below.

7.5.7.1 Capture the learner's attention

The game started with surprising statistics or facts about the game content that the learner chose to play with, to gain their attention and arouse their interest in the learning process. For example, when choosing cryptography, a learner may see the fun fact about it: "The film, The Imitation Game (2014) tells the story of Alan Turing and his attempts to crack the Enigma machine code during World War II". These facts appeared on the main menu of the game screen. Figure 12 shows a fun fact that appeared when chemistry content is selected. There were twenty-two fun facts about cryptography and thirteen about the chemistry content. Every-time the learner selected any of the content, one of these fun facts was displayed at random. See Appendix A and B for a full list of fun facts that were used in the game.

7.5.7.2 Introduce the learning objectives

Goal and outcomes were displayed in the game on collecting the treasure on every level. For the chemistry content, the goal was to learn to balance a chemical equation, and for cryptography content, the goal was to learn to encode/decode a piece of text with the given key using Caesar cipher. Figure 8 shows the goal of the first level for the chemistry content.

7.5.7.3 Induce prior learning

The game activated the recall of prior knowledge with the help of a pre-test administered before the game play. The pre-test consisted of twenty questions for each content and did not allow skipping any question. The pre-test questions are listed in the Appendix (E, F, and H). The pre-test score was also used to calculate the skill threshold for a learner which was used to show remediation in certain cases. It was obtained by dividing the pre-test score by the number of questions, i.e. $threshold = score \div 20$.

7.5.7.4 Present the content

Learning content was displayed in small chunks to the learner, with the help of text and images. The chunks were set as timed pagination which could be navigated in a sequence. The reading material was timed according to an average reading speed of 350 words per minute. Therefore, a learner could not skip through the reading chunks unless the timer expired. At the end of the timer, a “Next” button appeared at the bottom right of the screen that would enable the learner to go to the next chunk of reading material.

For the cryptography content, reading material consisted of the explanation about the origin of Caesar cipher and how to use it for basic encryption and decryption of textual data. See Appendix D for the reading material for cryptography. The chemistry content talked about the structure of a chemical equation and how to balance it (Appendix C).

7.5.7.5 Guide the learning process

The learner was guided through their game play. A tutorial level was made to gauge their game play skill, while explaining the mechanics of the game to the learners, helping them navigate the game environment. During the game play, after collecting the required molecules (for chemistry) or letters (for cryptography), a speech bubble appeared, guiding the learner to collect the “GO” text to proceed further to the next level of the game.

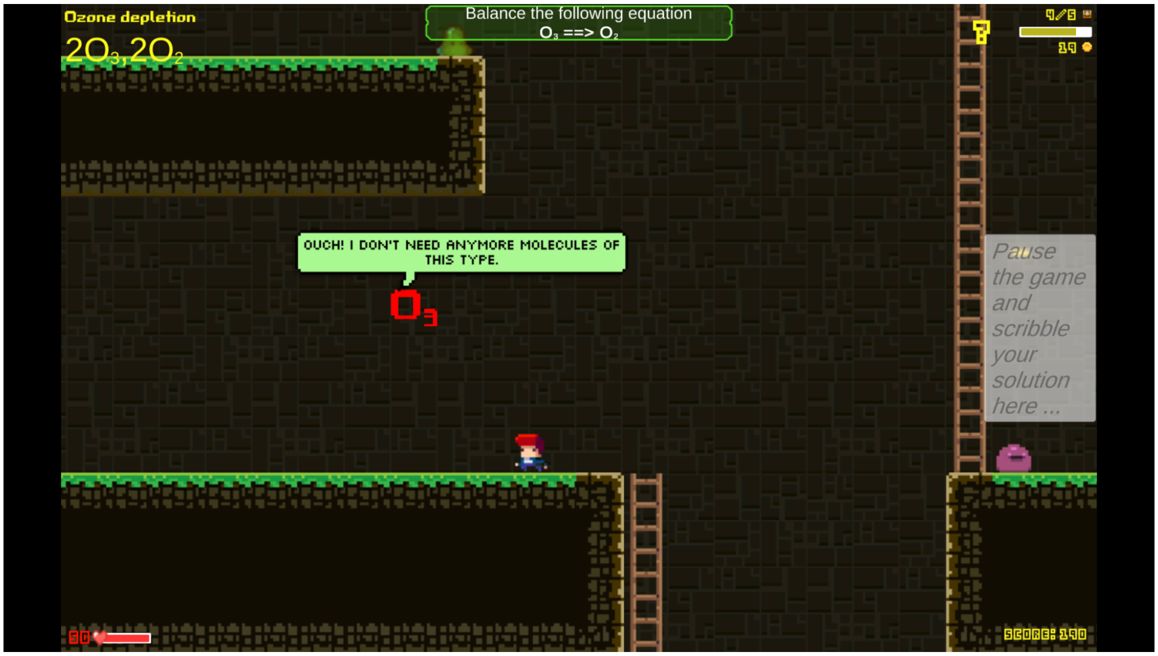
7.5.7.6 Allow practice time

The game had four content levels with increasing difficulty and allowed the learner to play as long as they would like to and proceed at their own pace as there was no time limitation. They were asked a question about the task (example, Figure 9) they performed at the end of each content level. This was done in accordance with the testing, spacing, and negative suggestion effect described earlier.

7.5.7.7 Provide feedback in a timely manner

Providing feedback during the learning process involved displaying remediation options whenever learner’s skill fell below their threshold calculated from the pre-test score. Learner’s skill was inferred from the knowledge node of the DBN. For the first content level ($t = 0$), the Knowledge0 node represented the learner’s skill and for later content levels ($t > 0$) Knowledge1 node governed it. Whenever the current skill determined from DBN fell below the threshold, the next collectible that spawned would be displayed as a help symbol (?) which when collected paused the game and showed the reading material that was presented to the learner at the beginning of the game.

Further, the player received a kickback, health loss, and feedback on picking up a



Balance the following equation

O_3 ==> O_2

OUCH! I DON'T NEED ANYMORE MOLECULES OF THIS TYPE.

Pause the game and scribble your solution here ...

Figure 13. Message that appears when the player collects a distractor element.

distractor element. The kickback and feedback were present on all the content levels. However, the health loss increased with the content level. For the first content level, there was no health loss to allow the player to learn the mechanics of the game. However, it increased with each level and caused instant player death on content level 4, which was the last and most difficult content level. For the chemistry version, feedback said that they do not need any more molecules of this type. For the cryptography version, the feedback said that they do not need any more letters of this type, or the letter does not exist in the resulting encoded or decoded text. As an example, Figure 13 shows a screenshot of the message that appeared when the player collected an excess of the Ozone (O_3) molecules. Lastly, at the end of each content level, the learner was given a star rating based on the current skill level, along with the feedback about their quiz response being correct or incorrect. The star rating was rewarded out of three stars based on the current knowledge level output of the DBN.

7.5.7.8 Assess performance

Learner performance was assessed throughout the game play through the DBN which updated itself whenever new evidence concerning learner knowledge was made available. Collecting the distractor elements and responses to the level-end quizzes served as the evidence for the DBN. Remediation was carried out as stated earlier if the new belief about the learner skill fell below the threshold value determined from their pre-test score. Apart from the learner skill, their affect was continually monitored using facial emotion tracking. The game play was adapted based on the learner's affect, depending on the treatment group. The adaptation was built differently for distinct experiments and will be discussed in the following chapters regarding individual experiments.

7.5.7.9 Promote external knowledge transfer

This is arguably the most difficult part of Gagne's taxonomy of learning (Gagne et al., 2005). The game did not implement any tactics which were used to enhance the retention and transfer of learning to real-world situations.

Chapter 8

EXPERIMENT 1

Experiment 1 was conducted to investigate the influence of affective adaptation on the engagement and learning of an individual learner. As stated earlier, the current literature does not completely support or negate the effect of adaptation in serious games. Therefore, this experiment was designed to potentially explore if the adaptation will back the second hypothesis regarding affect assessment and engagement. To this end, only affect assessment was used to adapt the game play in the game Chem-o-crypt, ignoring the other forms of stealth assessment and their prospective role in adaptation. The purpose of the experiment was three-fold. Firstly, to test the second hypothesis from the set of research questions stated in chapter 1. Secondly, to collect the data required to develop an algorithm for more efficient affect detection. Thirdly, to use the data for the parameter learning of the DBN which was used in Chem-o-crypt.

8.1 Method

The experiment involved a randomized pretest-posttest control group design. Control group participants played the game at a constant difficulty throughout the game regardless of their affective state. While in the treatment group, they played in a dynamic difficulty game environment which was adapted using facial emotion tracking. In other words, the control group played without adaptation while the test group played the game with adaption built into it. The adaptation is the measure that was manipulated in the treatment group to gauge its impact on learning and engagement.

8.1.1 Participants

A total of 107 undergraduate students (78 male, 29 female, $M = 18.9\ years$, $SD = 2.6\ years$) were recruited to take part in this experiment. Their participation lasted up to 1.5 hours ($M = 42.1\ minutes$, $SD = 8.7\ minutes$) and they were given a 1.5-course credit. Sixty-one participants reported having played games with an average game play time of six hours per week and a standard deviation of 8.82 hours.

8.1.2 Material

Chem-o-crypt game with chemistry content was used for the purpose of this experiment. The game had four difficulty levels and it started with a default difficulty level of one when players started playing the game. It remained the same for the control group participants. For the test group participants, it increased when they were bored and decreased when they felt frustrated, based on the detected state of their affect. Participants began in the first chunk when the game started. A flowchart depicting the participant workflow during their participation is shown in Figure 14.

8.1.3 Affect detection

This experiment adopted a process similar to the emote-aloud procedure, which was used by Craig et al. (2008) to capture the changes in the affective states.

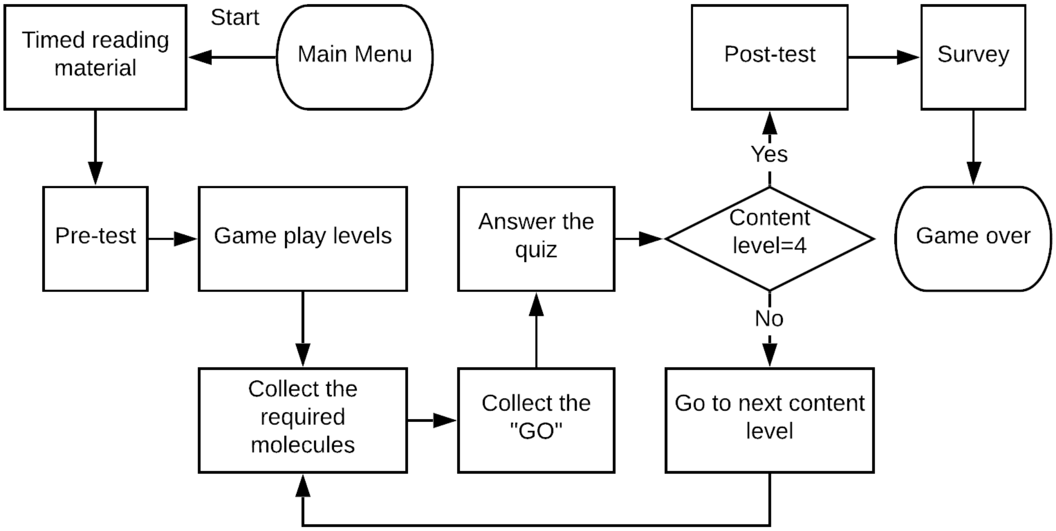


Figure 14. Flowchart depicting a typical participant workflow for Experiment 1.

Whenever the detected value for any emotion surpassed the threshold value of 40 (the maximum value was 100) during the game play, a pop-up message appeared at the bottom of the screen (Figure 15) which acted as a self-reported measure of the affective state. It asked the participants to pick one of the four available choices, which were, bored, frustrated, engaged, or other. The self-emote pop-up disappeared when the learner selected one of the options, until then it kept blinking at the rate of 2 Hz. Further, the pop-up appeared only appeared after the 30 seconds were elapsed in the game play. The interval between two intermittent pop-ups was kept at 90 seconds at the minimum, to reduce the potential interruptions that it can cause.

Apart from the self-reported affect, the player emotions were continuously tracked using the Affectiva SDK (McDuff et al., 2016). The observed emotions were categorized in real-time into the affective states of boredom, flow, and frustration using the following algorithm adopted from Baron (2017).

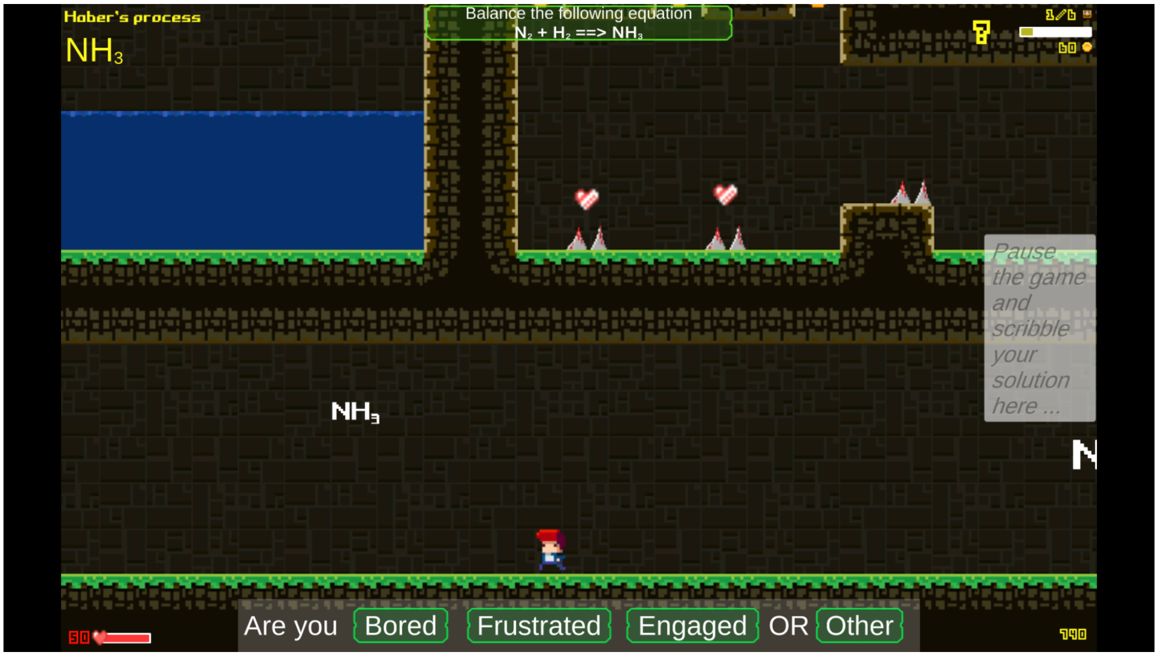


Figure 15. Screen capture of the emote aloud pop-up.

- If all the emotions are below the threshold, then the player is classified in a BORED state unless they were in a state of FLOW previously.
- If any of the emotions is above the threshold, then the player is in a non-bored state.
  - If anger is above the threshold and happiness is below the threshold, then the player is classified to be in a FRUSTRATION state.
  - If surprise is above the threshold and sadness is below the threshold, then the player is classified to be in a FLOW state.
- If the above rules fail, then the player is classified to be in a state called NONE.

A threshold of 10 was used for anger, while for all the other emotions a threshold of 20 was used. This was deliberately done due to the difficulty in detecting anger (Craig et al., 2008). When the player moved between the chunks, crossing the chunk boundary, the entire affect data from previous time-frames was aggregated and the most frequent affective state was assigned to the player at that point. In sum, this state represented the overall affective state of the player which they had while exploring the current chunk before moving to another. Therefore, the aggregation was reset on crossing the chunk boundary. This classification was then used to dynamically change the game difficulty depending on the test and control groups. If a state of frustration was detected in the test group, the difficulty was decreased by 1 when it was higher than 1. On detecting boredom, the difficulty was incremented by 1 if it was less than 4 (which was the maximum value of the difficulty in Chemo-o-crypt).

8.1.4 Procedure

The experiment took place in a computer lab with maximum participation of 20 participants at any point in time. Participants signed the consent form electronically and sat approximately 60 cm away from the monitor. To avoid any hindrance in the facial emotion detection process, participants were requested to remove their caps and glasses and to abstain from masking their faces with their hands while playing the game. Upon consenting to the participation, they downloaded the game and started it. As the game started, it assigned the participants into either test or the control group randomly. Then they played the game as per the workflow depicted in Figure 14 until they finished it and were rewarded course-credits upon game completion. There were four content levels and the experiment ended when the player cleared all four levels.

8.2 Analysis

8.2.1 Item-response analysis

One-parameter logistic model (1PL) based item response analysis was carried out to analyze the pre-test instrument which modeled the response of each participant to each question (or item) in the test.

8.2.2 Inter-rater reliability

The affect classification obtained from the affect algorithm used for automatic affect detection was validated using the instance of self-emote elicited by the participants. To this end, data were gathered for the duration represented by each self-emote instance. Then the affect detection algorithm was used to predict the affect from emotions for each data-frame that occurred during that duration. For example, a participant indicated that they were engaged during the time period from 584.93 to 675.28 seconds. So the emotion data (7 columns representing 7 emotions) corresponding to that participant for that duration (which consisted of 1592 rows or data frames) was reduced to a single column containing affect classification. These 1592 values for affect classification were then aggregated and the most frequent value was assigned as the affect of that participant for that duration. The same process was used for all the self-emotes for every participant resulting in a data matrix comprising of 1030 values. The self-emote data was then used for validating the affect classification obtained from the algorithm. For calculating cohen's kappa of boredom, engagement, frustration, and others were relabelled as ‘nonboredom’ states. The same procedure was used for engagement and frustration.

8.2.3 Learning and Engagement

ANCOVA was used to compare the post-test score across two groups by keeping the pre-test score as the covariate. An independent sample t-test was used to contrast the effect of stealth assessment on engagement and the other sub-scales of user engagement scale. Data for two participants were removed as there were some missing values for the UESz items, leading to a total of 105 values that were available for comparison.

8.2.4 Parameter Learning

The pre-test score (prior), evidence of collecting distractors, and the quiz responses of the participants were gathered in a log file. This data was then used for parameter learning in the Bayes Server 8.17 using Log-Likelihood as the convergence method and a rolling time-series mode (BayesServer, 2020). This allowed for learning the conditional probabilities of the nodes which are part of the DBN which can then be compared across the test and the control groups. The analysis is done for the cumulative data set, as well as the test and the control groups.

8.2.5 Affect Detection Algorithm

The observed emotions and expressions from Affectiva SDK were averaged across all time points between two neighboring pop-ups, with an exception that, for the first pop-up message, the averaging was performed across the time points between the start of the game and the time when the first self-emote pop-up appeared. This process was used to create a data matrix that consisted of the averaged emotion and expression indices for each time when the pop-up message appeared for all participants. As a result, this data consisted of 1030 observations.

The data were then fitted using the step-wise binomial logistic regression. Each pop-up instance could have had one of the four classes (e.g. boredom, flow, frustration, and other). Therefore, boredom, frustration, and others were relabelled as 'nonflow' states to be able to apply the binomial logistic regression resulting in the two classes of flow and nonflow states. The same procedure was also applied to the boredom and frustration classes, but not to the 'other' class, to examine if the tracked emotion and expression data could predict the states of flow, boredom, and frustration. Emotion (Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Contempt) or facial expression (Attention, BrowFurrow, BrowRaise, ChinRaise, EyeClosure, InnerBrowRaise, LipCornerDepressor, LipPress, LipPucker, LipSuck, MouthOpen, NoseWrinkle, Smile, Smirk, UpperLipRaise) was set as the predictor. The cognitive-affective state of flow (flow, nonflow), boredom (boredom, nonboredom), or frustration (frustration, nonfrustration) was set as the outcome, resulting in a total of six regression models. The 'other' class was beyond the scope of the study mainly because it is an ambiguous notion or state, thus was not reported.

To fit the model for each cognitive-affective state, the data was split into test and training datasets. 80% of the entire data was randomly assigned to training data and the rest was assigned to the test data. The R software (R Core Team, 2013) and the glm package (Jackman, 2017; Zeileis et al., 2008) were used to fit the binomial logistic models from the training data, with a binomial family and logit as model link functions. As the last step, the accuracies of the models were evaluated by examining the correlations between the predicted and the actual cognitive-affective states using the test data.

8.2.6 Qualitative analysis

Thematic analysis was used to analyze the feedback from participants. After filling the demographics survey, participants were asked to provide feedback (optional) if they had any. A total of 51 participants, 27 from the test group and 24 from the control group submitted the feedback. The feedback was then coded as positive, neutral, or negative, by two independent raters. Twenty percent of the data were randomly selected for inter-rater reliability resulting in a cohen's kappa of .75 ($p < .001$). An inductive approach was then used to discover the sub-categories within the three main categories, with the help of feedback data. The sub-categories identified were “suggestion”, “fun”, “liked graphics”, “liked music”, “interesting”, “confused”, “liked game”, “frustrated”, “disliked music”, and “difficult”.

8.3 Results

8.3.1 Item response analysis

Mean pre-test score of all the participants was 16.5 ($SD = 3.25$, $\alpha = .81$, Mdn=18) and for post-test it was 16.07 ($SD = 3.78$, $\alpha = .85$, Mdn=18). For both the pre-test and the post-test, item 20 had the least while item 1 had the most proportion of correct answers discriminating between subjects with different abilities. This is evident from the item characteristic curve shown in Figure 34. The Item information curve in Figure 35 displays the range of ability over which individual items contributed. For example, it suggests that item 1 was futile to say anything about the ability of the people who possessed more than average ability to answer the pre-test. Similarly, item 20 did not provide much information about the low-ability ones. Further, the test response function in Figure 37 shows that participants having more than average ability scored very high on the pre-test indicating that the test was very easy. Finally, the test information function in Figure 37 suggests that the pre-test provided most information regarding the participants who scored 2 standard deviations below the mean. Therefore, the pre-test did not provide enough information regarding the participants who had high ability and was re-designed for the later experiments.

8.3.2 Inter-rater reliability

Self emote-data consisted of 205 observations for boredom, 321 for flow, 395 for frustration, and 109 for other. The data obtained from the classification algorithm consisted of 988 observations for boredom, 35 for flow, 4 for frustration, and 3 for other. Analysis of inter-rate reliability yielded a negative cohen's kappa value of $-.002$ ($p = .80$) for boredom. Reliability of frustration ($\kappa = -.008$, $p = .11$), flow ($\kappa = .04$, $p = .01$), and other ($\kappa = -.006$, $p = .55$) were also very low. To overcome these reliability issues, another affect detection algorithm was built whose results are described in the sections below.

8.3.3 Learning and Engagement

Pre-test score was a significant co-variate for the post-test score. The post-test score mean was lower in stealth assessment group ($M = 15.94$, $SD = 3.77$) as compared to the non-stealth group ($M = 16.14$, $SD = 3.86$). However, there was no significant effect of stealth assessment on the post-test score, $F(1, 102) = 3.52$, $p = .06$, $\eta_p^2 = .03$.

The independent samples t-test performed on the data revealed that the mean UESz score was significantly different for the two groups, $t(103) = 3.07$, $p = .003$. The UESz score was much higher in non-stealth group ($M = 87.18$, $SD = 21.45$) compared to the stealth group ($M = 73.93$, $SD = 22.65$). The score on the sub-scales of UESz was also significant for focused attention, [$t(103) = 2.34$, $p = .02$], perceived usability [$t(103) = 4.64$, $p < .001$], and satisfaction [$t(103) = 2.08$, $p = .04$]. However, it was not significant for aesthetics [$t(103) = -.42$, $p = .68$].

8.3.4 Parameter Learning

The test group (n=55) played the adaptive version of the game and 52 in the control group played a static version of the game. In the test group, a total of 41 participants exhausted the available player lives and therefore did not complete all the four content levels. However, 31 participants from the test group still managed to reach content level 4. Exhausting the available lives instead of finishing all the four content levels would mean that less data is available for parameter learning. Due to the presence of the latent nodes (Knowledge0 and Knowledge1), two equivalent solutions were obtained as a result of parameter learning. These solutions suggested the phenomenon of label switching (Jasra et al., 2005). However, the most interpretable solution is presented in the findings.

The parameter learning for the entire data set using Log-Likelihood converged in 23 iterations. Conditional probabilities thus obtained from the parameter learning are summarized in Tables 16, 17, and 18. The probabilities of the Prior node (Table 16) show that most of the participants scored 80% and above in the pre-test, with almost none answering more than fourteen questions incorrectly. However, there were only a handful of participants who answered all the twenty questions correctly. Therefore, the data set consisted of more people who had a high level of initial knowledge according to the pre-test scores.

The parameter learning for the test data using Log-Likelihood converged in 54 iterations. Conditional probabilities thus obtained from the parameter learning are summarized in Tables 16, 17, and 19. The probabilities of the Prior node (Table 16) show that 80% of the participants in the test group scored 80% and above in the pre-test.

The parameter learning for the control data using Log-Likelihood converged in 25 iterations. Conditional probabilities thus obtained from the parameter learning are summarized in Tables 16, 17, and 20. The probabilities of the Prior node (Table 16) show that 72% of the participants in the control group scored 80% and above in the pre-test.

Table 18 shows the conditional probabilities of various nodes given the knowledge node at the first content level ($t = 0$). As an example, the probability that a participant picked up the second distractor (Distractor01) given they have the knowledge of the content is 0.01. Similarly, the probability that they answered the quiz incorrectly despite having the knowledge is 0.58. Further, the probability that they have the knowledge on the next level (Knowledge 1 is true), given that they possess the knowledge on level 1 (Knowledge 0 is true) is 0.53. Table 18 also shows the results for the higher content levels (two and above). For example, the probability that the participant will pick up the third distractor (Distractor12) above level 1 given that they do not have the knowledge on that level is 0.20. Similarly, looking at the last column of data, the probability that they will lack the skill on the next level (Knowledge 1 is false) given that they possess the skill on the current level (Knowledge 1 is true) is 20%. Table 19 shows the learned probabilities for the test group, while table 20 is for the control group.

Parameter learning of the structured DBN revealed the probabilities associated with various events happening in the game. Results of the analysis at time $t = 0$ for level 1 are as expected. There is a 52% chance to pick up the first distractor despite having the knowledge, and a 99% chance if the knowledge is missing. On picking up the first distractor, players get a kickback and a feedback message not to pick them up, and therefore the probability of picking up the second distractor went down to

1%, provided they have the knowledge. However, it remained high (0.94) for the less skilled ones who do not have the knowledge yet. The probability of collecting the third distractor went even further down to almost 0% for knowledgeable ones, suggesting the possible effectiveness of the feedback system. However, it could also be possible that the third distractor did not show up in the player's surroundings before they could finish the level, since it is a random process. Slip rate for answering the question incorrectly when players possess the knowledge remained low (3%), while the guess rate for guessing the answer correctly despite no knowledge was high (58%). However, the conditional probabilities of having the knowledge on level 2, given the knowledge on level 1 were unexpected. Ideally, it is assumed that once a player has gained knowledge, they are going to retain it 100% (Pardos & Heffernan, 2010). But the results show sustenance of about 53% only, a loss of 47%. Results show a low probability of 34% that a player who is not skilled on level 1 will transition to get the skill on the next level.

Conditional probabilities obtained for $t > 0$ were as expected. The probability of picking up the first distractor, given the player has the knowledge, decreased to 34%. For the second and third distractors, they were almost zero. However, for the players with a low knowledge level, the probabilities of picking up the distractor remained high at ~100%, 64%, and 20% respectively. The slip rate for answering the question increased from 3 to 8%, while the guess rate also increased from 56% to 75%. The knowledge retention rate increased from 53% to 80%, while the transition rate decreased from 34% to 11%.

Separate parameter learning for the test and the control groups revealed some differences between the two. Conditional probabilities obtained for $t = 0$ were similar for the test and control groups, as compared to the overall data, the only difference being the Knowledge1 node. While the probability of retaining the knowledge from level 1 to level 2 was comparable, the probability of transition from no knowledge state to knowledge state was much higher in the test group (71%) compared to the control group (32%). However, the probability of picking up the first distractor at $t > 0$ was higher (62%) in the test group compared to the control group (30%). At $t > 0$, the slip rates and guess rates were also higher for the test group (18% and 94%) in comparison to the control group (7% and 70%). The transition rate at $t > 0$ was also higher for the test group (35%) than the control group (14%).

8.3.5 Affect Detection Algorithm

Binomial logistic regression was used to create new model equations for the three affective states of boredom, flow, and frustration. The equations were derived from either emotions or expressions and are detailed below.

8.3.5.1 Flow modeling

The data consisted of 321 rows in which participants reported being in the state of flow. The rest of the 709 rows were coded as the non-flow states.

8.3.5.1.1 Model using emotions

The flow model obtained using the emotion data can be expressed as:

$$ln(\frac{Flow}{NonFlow}) = -0.84 + (0.4 \times Fear) + (0.09 \times Happiness) + (-0.074 \times Sadness)$$

Fear ($p = .03$) and Happiness ($p = .02$) were identified as significant predictors of the flow state. In contrast, Sadness ($p = .12$), Anger ($p = .99$), Disgust ($p = .47$), Surprise ($p = .38$), and Contempt ($p = .59$) were non-significant predictors. The difference between the null deviance (1025.9) and the residual deviance (1013.4) was 12.5. The smaller residual deviance, compared to the null deviance, indicates that the model can better predict the outcomes compared to the null model, which is a model with only the intercept. The model was significant, p=.006, and correctly classified 72.8% of the cases with an area under the receiver operating characteristic curve (AUC) of .57. The predicted and actual values were positively correlated, $r_s(204) = .25$, $p < .001$.

8.3.5.1.2 Model using expressions

The flow model obtained using the expression data can be written as:

$$\begin{aligned} ln(\frac{Flow}{NonFlow}) = 1.5 + (-0.02 \times Attention) + (-0.025 \times EyeClosure) + \\ (-0.037 \times InnerBrowRaise) + (0.02 \times LipPucker) + \\ (-0.02 \times LipSuck) + (0.02 \times MouthOpen) + (0.08 \times Smile) \end{aligned}$$

InnerBrowRaise ($p = .01$), LipPucker ($p = .05$), MouthOpen ($p = .01$), and Smile ($p = .02$) were identified as significant predictors of the flow state. In contrast, Attention ($p = .10$), EyeClosure ($p = .06$), LipSuck ($p = .12$), BrowFurrow ($p = .33$), BrowRaise ($p = .14$), ChinRaise ($p = .83$), LipCornerDepressor ($p = .81$), LipPress ($p = .31$), NoseWrinkle ($p = .43$), Smirk ($p = .65$), and UpperLipRaise ($p = .88$) were non-significant predictors. Increasing LipPucker, MouthOpen, and Smile were associated with an increased likelihood of the immersion in the flow state, but increased InnerBrowRaise caused a reduction in the likelihood. The difference between the null deviance (1025.9) and the residual deviance (994.51) was 31.39 which was better compared to the model built using emotions. The model was significant, p<.001, and correctly classified 72.33% of the cases, with an AUC of .63. The predicted and actual values were positively correlated, $r_s(204) = .22$, $p < .001$.

8.3.5.2 Boredom modeling

The data consisted of 205 rows in which participants reported being in a state of boredom. The rest of the 825 rows were coded as the non-boredom states.

8.3.5.2.1 Model using emotions

The boredom model obtained using the emotion data can be expressed as:

$$ln(\frac{Boredom}{NonBoredom}) = -1.24 + (-1.13 \times Fear) + (-0.38 \times Happiness) + \\ (0.15 \times Sadness)$$

Happiness ($p = .01$) and Sadness ($p = .01$) were identified as significant predictors of the boredom state. In contrast, Fear ($p = .08$), Anger ($p = .63$), Disgust ($p = .90$), Surprise ($p = .19$), and Contempt ($p = .99$) were non-significant predictors. The difference between the null deviance (836.18) and the residual deviance (811.37) was 24.8. The model was significant, $p < .001$, and correctly classified 83% of the cases, with an AUC of .6. The predicted and actual values were positively correlated, $r_s(204) = .15$, $p = .03$.

8.3.5.2.2 Model using expressions

The boredom model obtained using expression data can be written as:

$$ln(\frac{Boredom}{NonBoredom}) = -8.44 + (0.07 \times Attention) + (0.02 \times BrowFurrow) + \\ (0.06 \times BrowRaise) + (0.02 \times InnerBrowRaise) + \\ (-0.028 \times MouthOpen) + (-0.03 \times Smile)$$

Attention ($p = .001$), BrowFurrow ($p = .04$), BrowRaise ($p = .001$), MouthOpen ($p = .019$) and Smile ($p = .01$) were identified as significant predictors of the boredom state. In contrast, InnerBrowRaise ($p = .09$), ChinRaise ($p = .29$), EyeClosure ($p = .87$), LipCornerDepressor ($p = .30$), LipPress ($p = .52$), LipPucker ($p = .60$), LipSuck ($p = .97$), NoseWrinkle ($p = .31$), Smirk ($p = .86$), and UpperLipRaise ($p = .55$) were not significant predictors. Increased Attention, BrowFurrow, and BrowRaise were associated with an increased likelihood of exhibiting boredom, but increased MouthOpen and Smile decreased the likelihood. The difference between the null deviance (836.18) and the residual deviance (791.52) was 44.65 which is better compared to the model built using emotions. The model was significant, $p < .001$, and correctly classified 83% of the cases, with an AUC of .64. The predicted and actual values were positively correlated, $r_s(204) = .15$, $p = .03$.

8.3.5.3 Frustration modeling

The data consisted of 395 rows in which participants reported being in a state of frustration. The rest of the 635 rows were coded as the non-frustration states.

8.3.5.3.1 Model using emotions

None of the emotions significantly predicted the state of frustration. Fear ($p = .14$), Anger ($p = .54$), Disgust ($p = .73$), Happiness ($p = .8$), Sadness ($p = .97$), Surprise ($p = .75$), and Contempt ($p = .72$), were non-significant predictors. The difference between the null deviance (1087.1) and the residual deviance (1084) was 3.1. The model was marginally significant, $p = .08$, and correctly classified 54.4% of the cases, with an AUC of .55. The predicted and actual values were negatively correlated, $r_s(204) = -.06$, $p = .38$.

8.3.5.3.2 Model using expressions

The frustration model obtained using expressions can be written as:

$$\begin{aligned} ln(\frac{Frustration}{NonFrustration}) = 1.85 + (-0.02 \times \mathit{Attention}) + (-0.03 \times BrowFurrow) + \\ (0.02 \times \mathit{EyeClosure}) + (-0.067 \times \mathit{LipPress}) + \\ (-0.03 \times \mathit{LipPucker}) + (0.03 \times \mathit{LipSuck}) \end{aligned}$$

BrowFurrow ($p = .01$), LipPress ($p = .01$), LipPucker ($p = .03$), and LipSuck ($p = .02$) were identified as significant predictors of the frustration state. In contrast, Attention ($p = .09$), EyeClosure ($p = .09$), BrowRaise ($p = .39$), ChinRaise ($p = .58$), InnerBrowRaise ($p = .20$), LipCornerDepressor ($p = .20$), MouthOpen ($p = .14$), NoseWrinkle ($p = .57$), Smile ($p = .93$), Smirk ($p = .75$), and UpperLipRaise ($p = .68$) were not significant predictors. Increased LipSuck was associated with an increased likelihood of being in the frustration state, but increased BrowFurrow, LipPress, and LipPucker decreased the likelihood. The difference between the null deviance (1087.1) and the residual deviance (1060.8) was 26.3 which is better

Table 4. Thematic analysis for feedback from experiment 1

Category	Total count	Test	Control
Positive	23	11	12
Negative	12	9	3
Mixed	3	0	3
Neutral	13	7	6
Sub-category	**Total count**	**Test**	**Control**
Fun	8	4	4
Liked graphics	4	2	2
Liked music	5	2	3
Interesting	4	1	3
Liked game	6	3	3
Confused	4	2	2
Frustrated	6	5	1
Disliked music	2	1	1
Difficult	5	4	1
Suggestion	9	5	4

compared to the model built using emotions. The model was significant, p<.001, and correctly classified 61.1% of the cases, with an AUC of .57. The predicted and actual values were positively correlated, $r_s(204) = .18$, $p = .01$.

8.3.6 Qualitative analysis

In the control group, 28 participants did not give any feedback regarding the game, while 24 did. In the test group, 27 gave feedback and 28 chose not to. Results from the thematic analysis are displayed in Table 4. Overall, participants gave less positive and more negative or neutral feedback in the test group. They found the game to be difficult and were more frustrated in the test group than the control group.

8.4 Discussion

8.4.1 Item response analysis

Median score of 18/20 indicated a ceiling effect (Taylor, 2012) for the pre-and-post tests used in this study. Therefore, they were redesigned for the subsequent experiments.

8.4.2 Inter-rater reliability

The inter-reliability of the affect detection algorithm was low. It was not sensitive to the affective states of flow and frustration. Therefore, another algorithm was developed for the follow-up experiments.

8.4.3 Learning and Engagement

Although there was no significant impact of adapting the game using affect assessment, the overall score was lower in the group that played the stealth version of the game. Further, the player engagement was significantly lower in the stealth group. These results do not support the second research question about the effect of affective adaptation on learning and engagement. However, as indicated earlier, the affect detection algorithm was not reliable enough to support these results. Therefore, the second research question needs to be re-evaluated with a better affect detection algorithm.

8.4.4 Parameter Learning

This analysis examined adaptivity using affect assessment in a serious game to determine its effectiveness for lower and higher domain learners. The rate for the transition from no knowledge state to knowledge state indicates that adapting the game will be more advantageous for low domain learners. The stealth group had a better transition rate (71%) as compared to the non-stealth group (32%) suggesting that the affective adaptation was useful in improving the knowledge levels in the stealth group. In the absence of affective adaptation, the transition rate remained relatively low. The adaptive game offered better learning to individuals with low skill levels, compared to the non-adaptive version of the game. These results signal the value of interventions to palliate the negative affective states such as frustration and boredom.

Learning has been shown to correlate negatively with boredom and positively with the flow, in an Interactive Tutoring System (Craig, Graesser, et al., 2004). Parameter learning from this experiment corroborates this finding in the context of an educational video game. Similar to previous research with interactive affective systems, the current work also showed that adaptivity based of off affect can impact learning. Affect adaptation is more important for learners who have low domain knowledge. These results show that affect adaptability is also useful for improving the performance of low domain learners within a serious game. Therefore, it is important to detect these affective states and provide a way to treat them in a manner that gets a learner more engaged in the learning process. Results from this study are also in agreement with those obtained by D'Mello et al. (2010) using AutoTutor and suggests the importance of affect in the learning process of a learner who has low initial domain knowledge. Although the adaptive system was beneficial for beginners, it did not have any impact on the learners who had high prior knowledge.

These results partially replicate the expertise reversal effect which explains the role of prior knowledge in the learning process. It states that instructional design and techniques that assist learners, are effective for novice learners and can lose their potency, and may even affect the expert learners in a negative manner (Kalyuga, 2007). This pattern was observed in the current results. Although the adaptive game is beneficial for low-knowledge learners as indicated above, it does not affect the learning of high-knowledge learners positively or negatively. The probability for the high skilled learners to keep their knowledge intact is 62% in the test group and 61% in the control group, suggesting that the adaptation did not have any impact on the learning of learners who had high prior knowledge. This is probably because they are disinterested in the learning process as they already possess the skill that the instructional medium is trying to impart. Shernoff et al. (2014) found in a longitudinal study carried out on high school students that they lacked engagement in the classroom if the learning task was not challenging in accordance with their skill level. Such observations could be possible in a serious game as well and could explain the results obtained in this experiment.

Conflicting evidence exists regarding the effect of adaptability on the learning imparted by a serious game. This largely depended on how the adaptation was built into the game (Ali & Sah, 2017; Vanbecelaere et al., 2020). Ali and Sah (2017) used user ontology and semantic rules to adapt the game and found better learner performance for the adaptive version of the game. On the other hand, Vanbecelaere et al. (2020) adapted the game to show a different number of exercises based on the learner's performance in previous exercises. They found no significant improvement as a result of adaptation. Current research uses the affective states of the learner to adapt the game by altering the game difficulty. When boredom kicks in, game difficulty is increased. This causes the game environment layout to change in a way that makes it harder to navigate around posing a challenge to the learner. Further, the health loss from collisions with the enemy and the enemy movement speed is increased to ramp up the challenge. Very finite research exists that adapts the game in such a manner. Present results may explain the current divide in the literature regarding the effectiveness of adaptation in serious games suggesting that it is effective for low domain learners only.

Previous theories such as Zone of Proximal Development (Vygotsky, 1978) and Knowledge Space Theory (Craig et al., 2013; Falmagne et al., 1990) state that adaptation based on the learner's domain knowledge supports learning because prior knowledge indicates what the learner is ready to learn next. However, the outcome of the current study indicates that the adaptation based on learner's affect is useful as well, especially for low domain learners. These findings are useful to keep the learners at the edge of their abilities by detecting their affect unobtrusively. Affect detection can be combined with other forms of stealth assessment to adapt the game play. For example, it can be used along with Dynamic Bayesian Network to assess the current knowledge level of the learner and provide remediation if the knowledge level falls below a certain threshold. It can be used in conjunction with the mouse tracking, player log data, and other forms of stealth assessment indicated in Verma et al. (2019).

Table 5. Comparison of prediction accuracy obtained using similar procedures

Affective State	Current Study	Bosch et al. (2015)	D'Mello et al. (2018)
Flow	.63	.67	.68
Boredom	.64	.60	.61
Frustration	.57	.63	.63

Table 6. Comparison of prediction accuracy obtained using different procedures

Affective state	Current Study	Sabourin et al. (2011)	D'Mello et al. (2008)
	50% chance level	14.29% chance level	50% chance level
Flow	72.8%	–	71%
Boredom	83%	18%	69%
Frustration	61.1%	28%	78%

8.4.5 Affect Detection Algorithm

Affective states predicted using the binary logistic algorithm achieved accuracies which were comparable to previous studies that used similar (Bosch et al., 2015; D'Mello et al., 2018) or different methods (D'Mello et al., 2008; Sabourin et al., 2011) (see Tables 5 and 6). The models built using expressions show comparable accuracy to the models built using emotions.

The analysis revealed the affective states of boredom, flow, and frustration can be predicted using the emotions of sadness, fear, and happiness, and the expressions that these three emotions are composed of. Happiness and Fear were significant predictors of Flow. Of the four expressions that predicted Flow, three were the ones that comprise these emotions. These expressions were Smile (Happiness), InnerBrowRaise (associated with Fear and Sadness), and MouthOpen (Sadness). Sadness and Happiness predicted the Boredom, as well the expressions of BrowRaise

(Happiness), MouthOpen (Sadness), BrowFurrow (Fear and Sadness), and Smile (Happiness). On the contrary, emotions were not able to predict frustration, while expressions of LipPress (Sadness), BrowFurrow (Fear and Sadness), and LipSuck (Sadness) could predict it significantly.

It is noteworthy that the identified predictors of flow and boredom were associated with the emotions of Happiness, Fear, or Sadness. These findings are consistent with claims on the composition of flow and boredom found with the existing literature. Czikszentmihalyi (1990) argued that flow is a state of delight, which suggested that the emotion of Happiness can be used as the proxy for the flow state. Furthermore, Lepp (2018) found that happiness is negatively related to boredom. Altogether, these results suggest that players' affective states and their performance in learning and game play are closely tied with basic emotions of Happiness, Sadness, and Fear. In return, game systems can enhance the learning and game performance of the players if they monitor these emotions and use the observations to personalize the game structure for the individual players.

The results also indicated that expressions were better than emotions in explaining the variance in the affective states. This is probably because the emotions are derived from expressions (iMotions Inc., 2018). Therefore, it is better to use expressions to predict these affective states instead of using emotions, and they are not required simultaneously to predict them. While the models showed promising results, rigorous testing and further research is necessary to determine if applying this model towards monitoring players' emotions and expressions, and adapting the game accordingly, will lead to the facilitation of learning and enhanced game performance.

8.4.6 Qualitative analysis

Feedback was more positive in the control group and more negative in the test group, supporting the results from the quantitative data analysis. Some participants indicated the game to be visually pleasing and having surprisingly good music. Others said it was fun and engaging, but the instructions were not clear. They revealed the game to be interesting yet confusing. One participant said that they will highly suggest the game to someone who is having difficulty in solving chemical equations. On the other hand, some participants found the game to be frustrating with bad music that was irritating and gave them a headache. One participant who was not an avid gamer said that they disliked playing video games and would have preferred plain information to the game. The thematic analysis also revealed the bugs and usability issues that were present in the game and were fixed before any subsequent experimentation as they may have caused frustration during the game play.

8.5 Limitations

A limitation is evident from the prior score distribution in Table 16. The prior score is the evaluation based on the pre-test which determines the participant's knowledge level before the game play. Participants are not evenly distributed across all the groups and most of them have a prior score of 80% and above. The study should involve some participants who have a low level of initial knowledge to prevent the bias that may occur because of this reason. There were not many participants who had extreme scores, i.e. either 0 or 20/20. Therefore, parameter learning did not return the expected probabilities for extreme cases. Table 17 suggests that the probability of knowledge when the participant scored 0 is 50%, which is unexpected. Therefore, these results must be interpreted with caution.

Chapter 9

EXPERIMENT 2

Experiment 2 was conducted to validate the DBN and to evaluate the effect of multi-modal adaptation on learning and engagement. This experiment used affect detection, stealth assessment as well as DBN to adapt the game. Affect tracking was combined with the stealth assessment to change the game difficulty in real-time and DBN was used for remediation purposes.

9.1 Method

The experiment implemented a randomized 2 × 2 factorial design with order (chemistry first or chemistry second) and adaptivity (On or Off) as factors (Table 7). The Order factor consisted of two levels which determined the order in which the contents were played. An order of chemistry first meant that the player played the chemistry content first, followed by the cryptography content. While chemistry second meant cryptography was played first. The adaptivity factor had two levels as well, which were used to denote if the adaptivity was on or off. Adaptivity being on would mean that the game play was adapted using affect and player interactions and remediation was displayed when player skill (governed by the DBN) fell below their skill threshold (determined from their pre-test score). Note that within the adaptivity condition, only game play was adapted, but not the learning content within the game.

Table 7. Experiment 2 factorial design, 2 × 2, with number of participants who completed both, one, none of the contents, respectively.

Play Order Adaptivity	Cryptography first	Chemistry first
On	Chem second, On (19,6,5)	Chem first, On (21,4,11)
Off	Chem second, Off (35,7,7)	Chem first, Off (36,9,12)

9.1.1 Participants

A total of 172 undergraduate students were recruited to take part in this online experiment. This experiment was conducted online due to the pandemic situation which did not allow in-person studies. In the previous experiment, there were no dropouts. However, 35 students quit the game abruptly without completing at least one game content in this experiment. This could be attributed to potential bugs in the game that were not discovered during the game testing or issues with the game user interface (UI) on different screen resolutions. The game was tested on a computer that had a resolution of 1920 x 1080. It was not possible to test it on other resolutions and therefore the game UI might have appeared differently on different resolutions causing some UI elements to go off-screen or scale abruptly. Further, twenty-six participants completed only the first content but dropped out before completing the second one. Consequently, 111 people completed the entire study without dropping out. Table 7 indicates the number of participants in each group (within parentheses) who completed both the contents, completed only one, and dropped out without completing, respectively. Of these 111 participants who completed the study, 91 were male and the rest female ($M = 21.6\ years$, $SD = 6.17\ years$). Their participation lasted up to 2 hours ($M = 95\ minutes$,

$SD = 29.5\ minutes$) and they were given 2-course credit. Seventy-six participants reported having played games with an average game play time of sixteen hours per week and a standard deviation of fifteen hours.

9.1.2 Material

Chem-o-crypt game with both chemistry and cryptography content was used for this experiment. Unlike the first experiment, this experiment had a tutorial level which was designed to gauge the game play skills of participants while simultaneously walking them through the game mechanics. The score on this tutorial level was used to assign the initial difficulty as well as the maximum difficulty for the game. If the participants completed the tutorial level without using all the available lives, then the difficulty cap was set to 4, in all other cases it was assigned using the formula, $\mathit{diffcap} = 4 \times \mathit{score} \div maxscore$, where $maxscore$ is the maximum possible score possible during the tutorial level. A player could earn a score by collecting coins and lives and may lose it when they collide with the enemy. For example, if the player achieved a score of 75% of the maxscore, then the maximum value for difficulty would be set to 3. If the $\mathit{diffcap} > 2$, then the initial difficulty was set to two for the participant. Therefore the player performance during the tutorial level was taken into account to set the game difficulty's initial and maximum value irrespective of the condition they were assigned to. However, for the participants that belonged to the non-adaptive condition, their difficulty remained at the initial level throughout the game play. For the participants who were in the adaptive group, the difficulty may have increased or decreased depending on when they crossed the chunk boundary.

Difficulty increased when the aggregate state detected during the boundary-crossing event was boredom and the player score was at least 40% of the maximum possible score at that instant. Conversely, it decreased when the aggregated state was frustration and the proportional score was less than 20%. Maximum possible score kept updating itself as the chunk layouts changed, as it was governed by the number of coins and lives that were present in the game environment. The score gain for collecting a coin was fixed at 10 points, while the score for collecting lives (or 1-ups shaped like hearts) depended on its location in the environment. It ranged from 100 to 1000 depending on the ease with which it can be collected. Hard to collect 1-ups gave more points than the easier ones. A flowchart depicting the participant workflow during their participation is shown in Figure 16.

9.1.2.1 Assessment design

Pre-test and post-test were redesigned for this experiment. In the earlier experiment, they both asked the same set of questions and the sequence of questions was the same for all the participants. But for this experiment, they were instead made isomorphic by making some questions slightly different from the pre-test and by altering the available answer choices. Further, the updated test included text implicit questions, transfer questions, and had choices that involved misconceptions. The new pre-test and post-test randomized the question and the order of their choices for each participant. They are available in the appendix F to I. Question 9 of the chemistry test had a typographical error in the answer choice, due to which it had to be removed from all the analysis, leaving 19 questions in the chemistry test instead of 20.

To evaluate these tests, separate data was collected online. The evaluation used participants from the same target population of undergraduate students which were

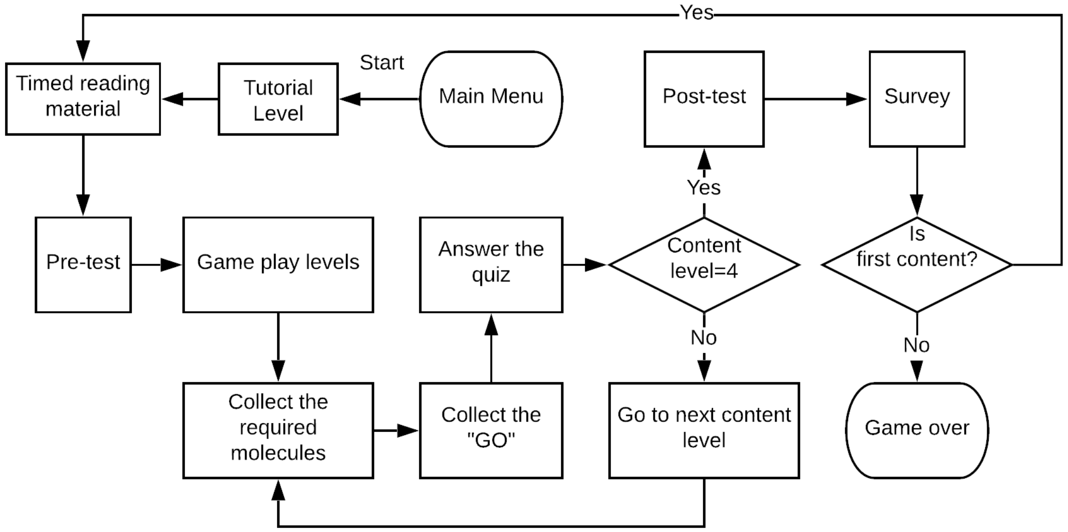


Figure 16. Flowchart depicting a typical participant workflow for Experiment 2.

used for the main experiment. Thirty-three participants completed the chemistry test and 36 finished the cryptography test. Item response analysis was then executed to evaluate the performance of these tests. Although the sample size required for using 1PL model is considerably higher than 33 (Stone & Yumoto, 2004), it was still the best model that could be used for this analysis.

9.1.2.2 Affect detection

This experiment utilized the model equations obtained from the previous experiment. Expressions were used to predict the probability value of each affective state for a given time-frame. The one which had the highest value was assigned as the affective state for that time-frame. For example, if the predicted values obtained for boredom, flow, and frustration were .23, .53, and .64 respectively, then the affective state of frustration was assigned to that particular time-point. All this data was then aggregated for the entire time period during which the player stayed in the chunk, and then the affect with the most frequent occurrence during that time-frame was assigned to the event of chunk crossing. For example, consider that a player stayed in chunk 2 during the time period 283 to 349 seconds. There were 950 observations of affect for the player during the time, which included 500 observations for flow, 200 for engagement, and 250 for boredom. The flow being the most frequent occurrence was assigned as the affective state of the player during that time period.

9.1.2.3 Dynamic Bayesian Network (DBN) Parameters

The conditional probabilities for the DBN were estimated based on the data collected during the previous experiment. The estimation took into account the probabilities learned from the parameter learning and the expert advice. The two methods were used together to fix the conditional probabilities used for the current experiment. The probabilities for the Prior and Knowledge0 node are shown in Table 8. The first row consists of possible states for the Prior node with the corresponding probabilities in the second row. The rest of the two rows tabulate the conditional probability of Knowledge0 given the states of the Prior node. Table 9 contains the probabilities for the rest of the nodes in the network. It depicts the conditional probability of Knowledge0 and Knowledge1 nodes based on the other dependent nodes in the network.

Table 8. Conditional probabilities used for Prior and Knowledge0 node in Experiment 2.

		Prior Score states (pre-test score)										
		0	1	2	3	4	5	6	7	8	9	10
		.01	.01	.01	.03	.04	.05	.05	.08	.30	.37	.05
Knowledge0	True	.40	.42	.44	.46	.48	.50	.52	.54	.56	.58	.60
	False	.60	.58	.56	.54	.52	.50	.48	.46	.44	.42	.40

Table 9. Conditional probabilities used for Distractor, Knowledge1, and Question nodes in Experiment 2.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.52	.48	.01	.99	.00	1.00	.97	.03	.53	.47
False	.99	.01	.94	.06	.28	.72	.58	.42	.34	.66
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.34	.66	.00	1.00	.00	1.00	.92	.08	.80	.20
False	1.00	.00	.64	.36	.20	.80	.75	.25	.11	.89

9.1.3 Procedure

The experiment took place in an online environment. Upon consenting to partake, participants downloaded the game and instructions from the researcher's google drive. They were asked to calibrate their webcam before starting the game. To avoid any hindrance in the facial emotion detection process, participants were requested to remove their caps and glasses and to abstain from masking their faces with their hands while playing the game. As the game started, it assigned the participants into one of the four groups randomly. Then they played the game as per the workflow depicted in Figure 16 until they finished it and were rewarded course-credits upon game completion. There were four content levels for each content and the experiment ended when the player cleared all the 4 × 2 levels.

9.2 Analysis

Multiple analyses were carried out to validate the DBN and evaluate the effect of multi-modal adaptation within and outside of the CAGE framework. These analyses are detailed below.

9.2.1 Learning and Engagement content

To evaluate the effect of adaptation on learning, percent score change from pre-to-post was analyzed as the dependent variable of interest and adaptivity as the independent variable. There were 19 valid questions in the chemistry test and 20 in the cryptography test. Therefore, the percent score change was used instead of the absolute change in score for the analysis. An independent sample t-test was then used to evaluate the hypothesis. Similarly, to determine the effect of adaptation on engagement, an independent sample t-test was used with UESz as dependent and adaptivity as an independent variable.

To evaluate the effect of adaptation on learning and engagement in a CAGE game, UESz and post-test scores were analyzed as dependent variables of interest with order and adaptivity as independent variables. For analyzing the post-test score, a $2 \times 2 \times 2 \times 2$ mixed ANOVA was performed with chemistry score (pre and post) and cryptography score (pre and post) as within-subject factors, and content order and adaptivity as between-subject factors. A $2 \times 2 \times 2$ mixed ANOVA was performed with participants' UESz scores for chemistry and cryptography as a within-subject factor and content order and adaptivity as between-subject factors. A similar process was used for analyzing the four sub-scales of the user engagement score, i.e. focused attention, perceived usability, aesthetics, and satisfaction.

9.2.2 Parameter Learning

The pre-test score (prior), evidence of collecting distractors, and the quiz responses of the participants were gathered in a log file. Similar to the previous experiment, this data was used for parameter learning in the Bayes Server 8.17 using Log-Likelihood as the convergence method and a rolling time-series mode (BayesServer, 2020). This allowed for learning the conditional probabilities of the nodes which are part of the DBN which can then be compared across two pairs of groups; the adaptive and the non-adaptive groups; and chemistry first and chemistry second groups. The analysis is done for the cumulative data set, as well as the four groups for both the chemistry and cryptography contents.

9.2.3 Qualitative analysis

Similar to the previous experiment, thematic analysis was used to analyze the feedback from participants. A total of 107 participants submitted the optional feedback. The feedback was coded as positive, neutral, negative, or mixed, by two independent raters. 20% of the data was randomly selected for inter-rater reliability resulting in a cohen's kappa of .80 ($p < .001$). An inductive approach was then used to discover the sub-categories within the four main categories, with the help of feedback data. The sub-categories identified were "suggestion", "fun", "liked graphics", "liked music", "interesting", "confused", "liked game", "frustrated", "difficult", "long", "usability issues", and "bugs". A similar analysis was done separately for the 61 participants who dropped out without completing both the contents. Of these 61 participants, 31 gave their feedback.

9.3 Result

9.3.1 Post-test and knowledge correlations

Correlation analysis was conducted to see if there is a relation between the post-scores in the two domains and the knowledge inferred from the DBN. Chemistry post-test score was positively correlated to the knowledge inferred from the DBN, $r_s(111) = .36$, $p < .001$. Cryptography post-test score was positively correlated as well, $r_s(111) = .46$, $p < .001$. The overall post-test scores for the two contents combined were also positively correlated to the knowledge inferred from DBN, $r_s(222) = .31$, $p < .001$. Overall, DBN showed a small but significant correlation of knowledge with the post-test scores for both the contents.

9.3.2 Item response analysis of the test design

9.3.2.1 Chemistry test

Mean test score of all the participants was 11.6 ($SD = 5.16$, $\alpha = .89$). Item response analysis using 1PL model revealed the proportion of correct answers were lowest for item 7 and highest for item 3. This is evident from the item characteristic curve which is shown in Figure 38. The Item information curve in Figure 39 suggests that item 3 did not convey much information about the high ability participants, and item 7 was not able to discriminate between low ability ones. Items 3 and 7 were not removed from the current analysis as they were not deemed to have an extreme enough impact on the test. Further, the test response function in Figure 40 shows that participants having more than average ability scored slightly above average on the test. Finally, the test information function in Figure 41 suggests that the test provided most information regarding the ability of the participants who scored slightly below the average.

9.3.2.2 Cryptography test

Mean test score of all the participants was 13 ($SD = 5.73$, $\alpha = .91$). Item response analysis using 1PL model revealed the proportion of correct answers were lowest for item 10 and highest for item 19. This is evident from the item characteristic curve which is shown in Figure 42. The Item information curve in Figure 43 suggests that item 19 did not convey much information about the high ability participants, and item 10 was not able to discriminate between low ability ones. Items 10 and 19 were not removed from the current analysis as they were not deemed to have an extreme enough impact on the test. Further, the test response function in Figure 44 shows that participants having more than average ability scored much above average on the test. Finally, the test information function in Figure 45 suggests that the test provided most information regarding the ability of the participants who scored slightly below the average.

Table 10. Mean and SD for pre-test and post-test scores by condition.

	Adaptivity	Order	Mean	SD	N
Chemistry Pre-test Score	OFF	Chem second	9.69	4.10	35
		Chem first	10.67	3.67	36
		Total	10.18	3.83	71
	ON	Chem second	8.74	3.41	19
		Chem first	11.86	3.34	21
		Total	10.38	3.68	40
	Total	Chem second	9.35	3.87	54
		Chem first	11.11	3.56	57
		Total	10.25	3.80	111
Chemistry Post-test score	OFF	Chem second	8.86	4.78	35
		Chem first	9.86	5.30	36
		Total	9.37	5.04	71
	ON	Chem second	8.68	4.75	19
		Chem first	11.95	4.09	21
		Total	10.40	4.66	40
	Total	Chem second	8.80	4.72	54
		Chem first	10.63	4.96	57
		Total	9.74	4.91	111
Cryptography Pre-test score	OFF	Chem second	14.29	4.70	35
		Chem first	10.75	5.59	36
		Total	12.49	5.44	71
	ON	Chem second	13.32	5.28	19
		Chem first	12.19	4.82	21
		Total	12.73	5.01	40
	Total	Chem second	13.94	4.89	54
		Chem first	11.28	5.32	57
		Total	12.58	5.27	111
Cryptography Post-test score	OFF	Chem second	13.26	4.57	35
		Chem first	11.86	5.60	36
		Total	12.55	5.13	71
	ON	Chem second	14.05	4.66	19
		Chem first	11.86	5.34	21
		Total	12.90	5.09	40
	Total	Chem second	13.54	4.58	54
		Chem first	11.86	5.46	57
		Total	12.68	5.10	111

Table 11. Mean and SD for pre-to-post score percent score change by condition.

	Adaptivity	Order	Mean	SD	N
Chemistry percent score change	OFF	Chem second	-4.36	17.98	35
		Chem first	-4.24	20.00	36
		Total	-4.30	18.90	71
	ON	Chem second	-0.28	24.53	19
		Chem first	.50	15.06	21
		Total	.13	19.85	40
	Total	Chem second	-2.92	20.39	54
		Chem first	-2.49	18.34	57
		Total	-2.70	19.28	111
Cryptography percent score change	OFF	Chem second	-5.41	22.16	35
		Chem first	5.85	20.39	36
		Total	.30	21.88	71
	ON	Chem second	3.88	17.79	19
		Chem first	-1.75	24.76	21
		Total	.92	21.65	40
	Total	Chem second	-2.14	21.04	54
		Chem first	3.05	22.20	57
		Total	.52	21.70	111

Table 12. Results from the 2 × 2 × 2 × 2 mixed ANOVA for learning.

	$F(1, 107)$	p	η_p^2
Chemistry Score	41.21	$<.001$	.28
Chemistry Score * Adaptivity	.07	.78	.00
Chemistry Score * Order	25.18	$<.001$	.19
Chemistry Score * Adaptivity * Order	.71	.40	.01
Cryptography Score	.23	.63	.00
Cryptography Score * Adaptivity	.76	.39	.01
Cryptography Score * Order	.29	.59	.00
Cryptography Score * Adaptivity * Order	1.80	.18	.02
Chem Score * Crypto Score	1.03	.31	.01
Chem Score * Crypto Score * Adaptivity	.44	.51	.00
Chem Score * Crypto Score * Order	.19	.66	.00
Chem Score * Crypto Score * Adaptivity * Order	2.65	.11	.02

Table 13. Mean and SD for UESz by condition.

	Adaptivity	Order	Mean	SD	N
Chemistry Engagement	OFF	Chem second	81.03	23.61	35
		Chem first	80.11	24.03	36
		Total	80.56	23.66	71
	ON	Chem second	78.05	23.33	19
		Chem first	92.95	15.96	21
		Total	85.87	20.94	40
	Total	Chem second	79.98	23.34	54
		Chem first	84.84	22.16	57
		Total	82.48	22.77	111
Cryptography Engagement	OFF	Chem second	85.94	21.46	35
		Chem first	76.06	27.70	36
		Total	80.93	25.14	71
	ON	Chem second	84.42	23.26	19
		Chem first	79.05	26.20	21
		Total	81.60	24.68	40
	Total	Chem second	85.41	21.90	54
		Chem first	77.16	26.96	57
		Total	81.17	24.87	111

9.3.3 Learning and Engagement content

There were 54 students who played cryptography first and 57 who played chemistry as the first content. 40 students played with stealth adaptation, and 71 played without it.

9.3.3.1 Adaptation vs. learning

The percent change in score for chemistry, when played as the first content, was not significant due to adaptivity, $t(55) = .94$, $p = .35$; $Cohen's\ d = .26$. It was not significant when played as the second content either, $t(52) = .70$, $p = .49$; $Cohen's\ d = .20$. For cryptography, it was not significant when played as either the

Table 14. Results from the 2 × 2 × 2 mixed ANOVA for user engagement and its sub-scales.

	$F(1, 107)$	p	η_p^2
UESz Score	.78	.38	.01
UESz Score * Adaptivity	1.23	.27	.01
UESz Score * Order	14.89	<.001	.12
UESz Score * Adaptivity * Order	2.23	.14	.02
Focused Attention	.01	.91	.00
Focused Attention * Adaptivity	1.11	.29	.01
Focused Attention * Order	22.32	<.001	.17
Focused Attention * Adaptivity * Order	2.62	.11	.02
Perceived Usability	3.35	.07	.03
Perceived Usability * Adaptivity	1.16	.29	.01
Perceived Usability * Order	.03	.86	.00
Perceived Usability * Adaptivity * Order	.33	.57	.00
Aesthetics	.49	.49	.001
Aesthetics * Adaptivity	.09	.77	.00
Aesthetics * Order	6.25	.01	.06
Aesthetics * Adaptivity * Order	.12	.74	.00
Satisfaction	.02	.90	.00
Satisfaction * Adaptivity	.18	.68	.00
Satisfaction * Order	17.74	<.001	.14
Satisfaction * Adaptivity * Order	3.04	.08	.03

first content, $t(52) = 1.57$, $p = .12$; $Cohen's\ d = .45$, or as the second content, $t(55) = 1.25$, $p = .22$; $Cohen's\ d = .34$. Therefore, present results do not support the hypotheses regarding the learning gain due to adaptation. Table 11 shows the means for all the conditions.

9.3.3.2 Adaptation vs. engagement

The UESz score for chemistry, when played as the first content, was significant due to adaptivity, $t(55) = 2.18$, $p = .03$; $Cohen's\ d = .60$. When adaptivity was on, the mean UESz score was 92.95 ($SD = 20.94$), and when it was off, the mean was

80.11 ($SD = 24.03$). However, UESz was not significantly different when chemistry was played as the second content, $t(52) = .24$, $p = .81$; $Cohen's\ d = .07$. For cryptography, it was not significantly different when played as either the first content, $t(52) = .44$, $p = .66$; $Cohen's\ d = .13$, or as the second content, $t(55) = .40$, $p = .69$; $Cohen's\ d = .11$. Therefore, present results only partially support the hypotheses regarding the engagement gain due to adaptation, depending on the content domain. Adaptation helped increase engagement for chemistry content when it was played first, but it did not improve for the cryptography content. It did not improve engagement for either content when they were played second. Table 13 shows the means for all the conditions.

9.3.3.3 Adaptation vs. learning and engagement in CAGE

Mean pre-test score for chemistry content was 10.25 ($SD = 3.80$) and 9.74 ($SD = 4.91$) for post-test. Mean pre-test score for cryptography content was 12.58 ($SD = 5.26$) and for post-test it was 12.68 ($SD = 5.09$) (See Table 10 and 11 for means, standard deviations by the group). The repeated measures ANOVA performed on these data revealed no four-way interaction effect among the four variables, i.e. chemistry score, cryptography score, content order, and adaptivity, $F(1, 107) = 2.65$, $p = .11$, $\eta_p^2 = .02$. The three-way interactions were also insignificant. However, the two-way interaction between the chemistry score and content order was significant, $F(1, 107) = 25.18$, $p < .001$, $\eta_p^2 = .19$. Additionally there was a main effect of the chemistry scores, $F(1, 107) = 41.21$, $p < .001$, $\eta_p^2 = .28$. No other two-way interactions and main effect were found. Analysis results corresponding to the factorial mixed ANOVA are available in Table 12.

A subsequent t-test was used to further analyze the effect of order on chemistry learning, to account for the significant interaction between the chemistry scores and content order. Using percent score change as a dependent variable and content order as an independent variable, no significant effect of the content order was observed, $t(109) = .12$, $p = .91$; $Cohen's\ d = .02$. Therefore, the present results do not support the hypotheses regarding learning gain due to adaptation in CAGE. Further, the paired sample t-test that compared the percent change in score for chemistry and cryptography were insignificant when they were played as first content, $t(109) = .09$, $p = .93$; $Cohen's\ d = .02$, as well as second content, $t(109) = 1.47$, $p = .14$; $Cohen's\ d = .28$.

Mean UESz score for chemistry content was 82.48 ($SD = 22.77$) and for the cryptography content it was 81.17 ($SD = 24.87$) (Table 13). The $2 \times 2 \times 2$ mixed ANOVA did not indicate any significant three-way interactions between the variables, $F(1, 107) = 3.04$, $p < .08$, $\eta_p^2 = .03$. There was no two-way interaction with Adaptivity, $F(1, 107) = 1.23$, $p = .27$, $\eta_p^2 = .01$. However, there was a significant interaction observed between the content and order, $F(1, 107) = 14.89$, $p < .001$, $\eta_p^2 = .12$. Analysis results corresponding to the factorial mixed ANOVA are available in Table 14.

A subsequent t-test was used to further analyze the effect of order and content domain on engagement, to account for the significant interaction between the UESz scores and content order. Using UESz as dependent variable and content order as independent variable, no significant effect of order was observed for either chemistry, $t(109) = 1.13$, $p = .26$; $Cohen's\ d = .21$, or cryptography, $t(109) = 1.76$, $p = .08$; $Cohen's\ d = .34$. Therefore, present results support the hypotheses that adaptation helped sustain engagement in a CAGE game but the reason for this sustenance could not be established.

9.3.3.4 UESz sub-scales

Mean focused attention score for chemistry content was 22.86 ($SD = 9.55$) and for cryptography, it was 23.07 ($SD = 10.42$). The analysis revealed no significant interaction effect between focused attention, content order, and adaptivity. However, there was a significant interaction between the content order and the focused attention (see Table 14).

Mean perceived usability score for chemistry content was 23.75 ($SD = 8.26$) and for cryptography, it was 22.53 ($SD = 7.86$). The analysis unveiled no significant three-way or two-way interaction effect between perceived usability, content order, and adaptivity (see Table 14).

Mean aesthetics score for chemistry content was 17.71 ($SD = 5.35$) and for cryptography, it was 17.50 ($SD = 5.97$). The analysis unveiled no significant interaction effect between aesthetics, content order, and adaptivity. However, there was a significant interaction between the content order and the aesthetics (see Table 14).

Mean satisfaction score for chemistry content was 18.16 ($SD = 7.88$) and for cryptography, it was 18.07 ($SD = 8.50$). The analysis unveiled no significant interaction effect between satisfaction, content order, and adaptivity. However, there was a significant interaction between the content order and the satisfaction (see Table 14).

9.3.4 Parameter Learning

The adaptive group (n=40) played the adaptive version of the game and 71 in the non-adaptive group played a static version of the game. 54 participants played the cryptography content first and 57 played the chemistry content first. Similar to the previous experiment, two equivalent solutions were obtained due to the presence of the latent nodes (Knowledge0 and Knowledge1) and the phenomenon of label switching (Jasra et al., 2005). However, the most interpretable solution is presented in the findings. Although the solutions for the chemistry content were not completely interpretable for all four groups, the solutions obtained for cryptography content were fully interpretable.

9.3.4.1 Chemistry

Conditional probabilities obtained from the parameter learning for chemistry content are summarized in Tables 21, 22, and 23. Conditional probabilities for adaptive and non-adaptive groups are tabulated in Tables 24 and 25. Conditional probabilities for the order groups are depicted in Tables 26 and 27.

Results of the analysis at time $t = 0$ for level 1 were as expected. There is a 39% chance to pick up the first distractor despite having the knowledge, and a 99% chance if the knowledge is missing. Similar to the last experiment, on picking up the first distractor, players get a kickback and a feedback message not to pick them up, and therefore the probability of picking up the second distractor went down to 1%, provided they have the knowledge. However, it remained high (0.84) for the less skilled ones who do not have the knowledge yet. The probability of collecting the third distractor went even further down to almost 0% for knowledgeable ones, corroborating the effectiveness of the feedback system from experiment 1. Slip rate for answering the question incorrectly when players possess the knowledge remained low (8%), while the guess rate for guessing the answer correctly despite no knowledge was high (53%). However, the conditional probabilities of having the knowledge on level 2, given the knowledge on level 1 were unexpected. Ideally, it was assumed that once a player has gained knowledge, they were going to retain it 100% (Pardos & Heffernan, 2010). But the results showed sustenance of about 60% only, a loss of 40%. Results showed a low probability of 35% that a player who was not skilled on level 1 would transition to get the skill on the next level.

Conditional probabilities obtained for $t > 0$ were somewhat unexpected. The probability of picking up the first distractor, given the player has the knowledge, increased to 42%. However, for the players with a low knowledge level, the probabilities of picking it up remained high at almost 100%. The slip rate for answering the question remained at 8%, while the guess rate increased from 53% to 69%. The knowledge retention rate increased from 60% to 77%, while the transition rate decreased from 35% to 12%.

Separate parameter learning for the four groups revealed some differences between them. Conditional probabilities obtained for $t = 0$ were similar for the four groups, as compared to the overall data, the only difference being the Knowledge1 node. While the probability of retaining the knowledge from level 1 to level 2 was comparable, the probability of transition from no knowledge state to knowledge state was much higher in the chemistry second group (77%) compared to the chemistry first (37%), adaptive (60%) and the non-adaptive (62%) groups. However, the probability of picking up the first distractor given no knowledge, at $t > 0$, was higher

(59%) in the chemistry first group compared to the chemistry second (30%), adaptive (52%) and non-adaptive group (39%).

9.3.4.2 Cryptography

Conditional probabilities obtained from the parameter learning for chemistry content are summarized in Tables 28, 29, and 30. Conditional probabilities for adaptive and non-adaptive groups are tabulated in Tables 31 and 32. Conditional probabilities for the order groups are depicted in Tables 33 and 34.

Results of the analysis at time $t = 0$ for level 1 were as expected. There is a 73% chance to pick up the first distractor despite having the knowledge, and almost 100% chance if the knowledge is missing. The probability of picking up the second distractor went down to 22%, provided they have the knowledge. However, it remained high (∼100%) for the less skilled ones who do not have the knowledge yet. The probability of collecting the third distractor went even further down to 1% for knowledgeable ones, corroborating the effectiveness of the feedback system from experiment 1 and chemistry content from experiment 2. Slip rate for answering the question incorrectly when players possess the knowledge remained low (8%), while the guess rate for guessing the answer correctly despite no knowledge was high (63%). However, the conditional probabilities of having the knowledge on level 2, given the knowledge on level 1 were unexpected. Results show sustenance of about 66% only, a loss of 34%. Results show a low probability of 19% that a player who is not skilled on level 1 will transition to get the skill on the next level.

Conditional probabilities obtained for $t > 0$ were somewhat unexpected. The probability of picking up the first distractor, given the player has the knowledge, decreased to 62%. However, for the players with a low knowledge level, the probabilities of picking it up increased to 100%. The slip rate for answering the question decreased from 8 to 3%, while the guess rate increased from 63% to 67%. The knowledge retention rate increased from 66% to 74%, while the transition rate decreased from 19% to 18%.

Separate parameter learning for the four groups revealed some differences between them. Conditional probabilities obtained for $t = 0$ were similar for the four groups, as compared to the overall data, the only difference being the Knowledge1 node. While the probability of transition from no knowledge state to knowledge state was comparable, the probability of retaining the knowledge from level 1 to level 2 was higher in the chemistry first (70%) and adaptive group (72%) compared to the chemistry second (59%) and the non-adaptive (61%) group.

9.3.5 Qualitative analysis

In the non-adaptive group, 81 participants did not give any feedback regarding the game, while 61 did. In the adaptive group, 41 gave feedback and 59 chose not to. In the chemistry second group, 69 participants did not give any feedback regarding the game, while 50 did. In the chemistry first group, 52 gave feedback and 71 chose not to. Results from the thematic analysis are displayed in Table 15. Overall, there were 24 instances of positive, 42 negative, 31 neutral, and 5 mixed feedback. In the feedback data from dropouts, 11 were positive, 12 were negative, and 6 were neutral.

Table 15. Thematic analysis for feedback from experiment 2

Category	Total	Adaptive	Non-adaptive	Chem second	Chem first
Positive	22	12	10	11	11
Negative	43	17	26	18	25
Mixed	8	3	5	1	7
Neutral	29	9	20	20	9
Sub-category	**Total**	**Adaptive**	**Non-adaptive**	**Chem second**	**Chem first**
Fun	12	8	4	5	7
Liked graphics	5	3	2	1	4
Liked music	3	1	2	0	3
Interesting	2	0	2	2	0
Liked game	8	6	2	3	5
Confused	5	3	2	4	1
Frustrated	20	8	12	6	14
Difficult	14	6	8	6	8
Long	7	4	3	3	4
Usability issues	12	7	5	4	8
Bugs	12	4	8	8	4
Suggestion	14	3	11	7	7

9.4 Discussion

Moderate to weak positive correlations were observed between the post-test score and the knowledge level inferred from the DBN. This provides an evidence of validity for the use of Bayesian networks to model learner beliefs in the CAGE based games. However, this does not rule out the possibility that the other network structures are not possible for the current gaming environment.

9.4.1 Learning and Engagement content

The learning for the chemistry content, depicted by the chemistry score was not significant based on the adaptivity. The cryptography learning due to adaptivity was insignificant as well. Therefore, the present results do not support the third research question regarding learning improvement due to adaptation. The results partially supported the fourth research question regarding engagement improvement due to adaptation, as the cryptography engagement did not differ significantly due to adaptivity, but the chemistry engagement was significantly better when it was played as the first content.

Previous study with CAGE led to a reduced engagement for the second content probably due to the fatigue effect (Baron, 2017). However, present results indicated that the player engagement was not significantly different across the two content. Suggesting that the experiment partially supported the fifth research question regarding engagement as the reason behind the sustained player engagement could not be established. Similar results were obtained for learning as the chemistry score as well as cryptography score did not differ significantly by order and adaptivity. Therefore, the fifth research question regarding learning improvement in CAGE was not supported.

The analysis examined adaptivity using affect assessment in a CAGE game to determine its effectiveness for sustaining engagement when playing multiple games that use content agnostic mechanics. There was no significant interaction between the adaptivity and content order, suggesting that the player engagement was maintained when playing multiple contents within a CAGE game. Although the UESz score differed significantly depending on the content order, it was not significantly different when adaptivity came into play. The mean UESz score was better in adaptive condition for both the game contents irrespective of the order but when the adaptivity was on, the UESz mean was lower when played as second content.

A previous study (Baron, 2017) found that engagement was reduced when playing second content in a CAGE game, probably due to fatigue effect or boredom. The current study replicates this finding as the UESz score significantly dropped when second content was played. However, the current study included game adaptation supported by stealth assessment, which possibly helped in preventing this decrease in engagement when playing second content. Therefore, present results may suggest that using adaptation in the current way can help sustain motivation, but not increase it when playing multiple contents within a CAGE game.

A study conducted by Sharek and Wiebe (2015) found that the adaptation in a puzzle-based game led to similar engagement as compared to linear game play or a game play driven by player choices. They used the past and current performance of a player, along with the secondary task and in-game behavior to select the next game level for the player. The current study adapted the game play differently but partially replicated the same results. The overall engagement observed was not significantly different in the adaptive game as compared to the non-adaptive game for cryptography content but it was better for the chemistry game when it was played as first content.

However, the current results should be interpreted with caution as the estimated effect sizes for the analysis were rather low. The ANOVA showed that the means were not significantly different due to adaptivity but the effect size was small. The partial eta squared was just .011, which means that the adaptation by itself accounted for only 1.1% of the overall variance in the scores. Similarly, order explained 12.2% but together order and adaptivity accounted for only 2% of the observed variance in the score. These results suggest that the effect was not present and would not be found even if a larger sample size was used.

A major limitation of the experiment was the online nature of the study. It had to be conducted online due to the pandemic situation. As a result, there was no control over the system in which the game was being run and therefore many participants dropped out of the study. A total of 35% of the participants did not complete the study which could be attributed to potential bugs or issues with the game user interface (UI) as indicated previously. However, the study did not appear to have a problem of attrition concerning any specific condition as the dropouts appeared to be random irrespective of the condition.

9.4.2 Parameter Learning

This experiment did not replicate the findings from the previous experiment regarding adaptation being more beneficial to low domain learners. The transition rate remained low for both the contents. Although the transition rate was high for some groups in the chemistry content, they were not completely interpretable. Therefore, their results were not taken into account for this discussion.

For the cryptography version, playing cryptography as the first content led to a reduction in the probability of picking the first distractor from 80% to 58% from content level one to two and above. While playing it as second led to an increase from 62% to 65%. The adaptive version demonstrated a much lower decrease from 74% to 73% while the non-adaptive had a substantial decrease from 73% to 54%.

Findings from the DBN were not consistent due to adaptation being implemented differently in the two experiments. In the first experiment only affect was used to adapt the game which employed an algorithm that was found to be unreliable. While in the second experiment player log tracking and DBN were used in addition to affect detection which used a different algorithm. Further work is required to determine the way adaptation should be implemented in a way that promotes student learning.

9.4.3 Qualitative analysis

Feedback was more positive in the adaptive group and more negative in the non-adaptive and the chemistry first group. One participant indicated that the game mechanics were entertaining and reinforced learning. Some participants indicated that the graphics were good, but they were lagging because of webcam use, which made the game hard to play and caused frustration. Another said that they did not know that collecting distractors would hurt them until it did, reinforcing the game design which was made for this purpose. Some other participants asked that they really liked the game and if it would be available even after the research study gets over. They said that they are now able to encrypt text using the Caesar cipher.

Participants found the game to be difficult and were more frustrated in the non-adaptive and the chemistry first group. A participant said that the game made them lose their mind and was difficult since they were not familiar with chemistry. Another indicated that the game was frustrating as they could not make a hard jump between two platforms. Some participants suggested that the game was very long and frustrating to play. Another indicated that they had no motivation to play the second content after the first got over, due to the length. Some said that they got frustrated because the game paused when they moved out of the camera's field of view, but it was required for the study.

In the feedback from dropouts, the instances of positives and negatives were approximately the same. Some participants thanked for providing them with a stress reliever game and for letting them participate in the study. They enjoyed the game and were interested in the experimental idea of using the camera to monitor their attention. While some suggested that it was difficult to play as they were not an active gamer. Another indicated that the game got tedious and boring towards the end, although it was fun to play initially.

Chapter 10

CONCLUSION AND FUTURE WORK

This chapter details the current findings with a detailed review of each research question. Conclusion and avenues for future work are also presented.

10.1 General discussion

Through these experiments, the engagement and learning components of a serious game were evaluated using a novel CAGE framework and stealth assessment. Current results indicated that the adaptation was not beneficial for learning irrespective of the adaptation techniques utilized in these experiments. However, it helped in sustaining the engagement, and in some cases even enhance it.

The results supported the first research question regarding the validity of DBN in a CAGE based game. The data collected from the first experiment were used to learn the parameters of the DBN which were subsequently used for the second experiment. The knowledge inferred from these probabilities was then correlated to the post-test score obtained. A weak to positive moderate correlation demonstrated support for the validity of the DBN.

Further, the two experiments showed that the DBN could be implemented in a content agnostic manner. Same network was employed for both the learning contents, which supports the aim for creating content agnostic game based assessment. The current results thus indicate that the DBN can be used in a content-agnostic way.

10.1.1 Adaptation vs. learning

Adaptation has been shown to be effective in some cases (Sampayo-Vargas et al., 2013; van Oostendorp et al., 2014) while ineffective in others (Shute et al., 2020; Vanbecelaere et al., 2020) depending on how it was implemented. The present study did not find significant improvement in learning regardless of the way adaptation was built into the game. In the first experiment, the Chem-o-crypt game was adapted solely using facial emotion tracking to examine its effect on the learning and engagement of a player. The experiment found no significant effect of learning as a result of affective adaptation. However, the facial emotion tracking used for the experiment was not reliable enough to warrant these results. Therefore, the data from this experiment were used to develop another algorithm for affect tracking which used the binomial logistic regression. The second experiment implemented this newly developed affect detection algorithm along with player log tracking and DBN to adapt the game. Despite a different adaptation methodology, it could not provide evidence to support learning gain irrespective of the order in which the game contents were played.

Further, the results from the first experiment indicated adaptation to be beneficial only for low domain learners but not high domain learners that were in agreement with the results obtained in AutoTutor (D'Mello et al., 2010). Although the results from the second experiment do not corroborate the results, it needs to be investigated if adaptation affects learning based on the prior knowledge of the students. Alternately, another study that involves adaptation in the first cage content but not the second content may provide future direction to probe into.

While the current results do not support learning, they indicate possible avenues for future research. Previous studies have shown conflicting results regarding the usefulness of adaptation for learning (Holmes et al., 2009; Sampayo-Vargas et al., 2013; Shute et al., 2020; Vanbecelaere et al., 2020). However, D'Mello et al. (2010) indicated in a non-gaming system that affect-sensitive systems are more useful for low domain learners but not for high domain learners which was replicated for a gaming system with the help of the first experiment. Therefore, it is possible that the adaptive game environment is not required to support the learning for high domain learners and may even be detrimental to their learning. But such an environment may back up and provide the required support for low domain learners.

10.1.2 Adaptive games vs. engagement

A research study by Sharek and Wiebe (2015) found no significant impact of adaptation on engagement. The current dissertation implemented adaptation differently compared to their study and found different results in the first experiment but partially replicated their findings in the second experiment. As stated earlier, the affect detection algorithm used for the first experiment was not reliable enough and it led to a reduction in engagement levels, providing no support for the second research question. The second experiment, however, used multi-modal adaptation which helped in intensifying engagement for chemistry content when it was played as first content. It did not impact engagement for the cryptography content regardless of the order. Chemistry content in these experiments was more complex and difficult than the cryptography content. Therefore, present results may suggest that the multi-modal adaptation could help increase engagement for more complex learning content rather than the simpler ones.

The second experiment also demonstrated some evidence in support of the fifth research question in view of the engagement in CAGE. Previous study has shown adaptive game level sequence to be inconsequential to engagement levels within a puzzle-based game (Sharek & Wiebe, 2015). Similar results were obtained in the second experiment. The multi-modal adaptation probably helped in sustaining engagement across the multiple CAGE contents, although the reason for this sustained engagement could not be confirmed. These results may imply that the adaptation is more useful for a complex content such as chemistry in these experiments, as compared to rather simpler ones. Therefore, to decide whether or not to provide additional adaptive support within a game, it is advised to find out more about the complexity of the learning content.

10.2 Future Work

There are many potential studies that the current experiments could lead to. The first area is the reliability of the affect algorithm that stemmed from the first experiment. The algorithm was not evaluated for its reliability. Further, the algorithm was developed from the data obtained from university undergraduate students. Therefore, an independent experiment involving affect tracking needs to be done for establishing the reliability and validity of the algorithm for a broader range of the population.

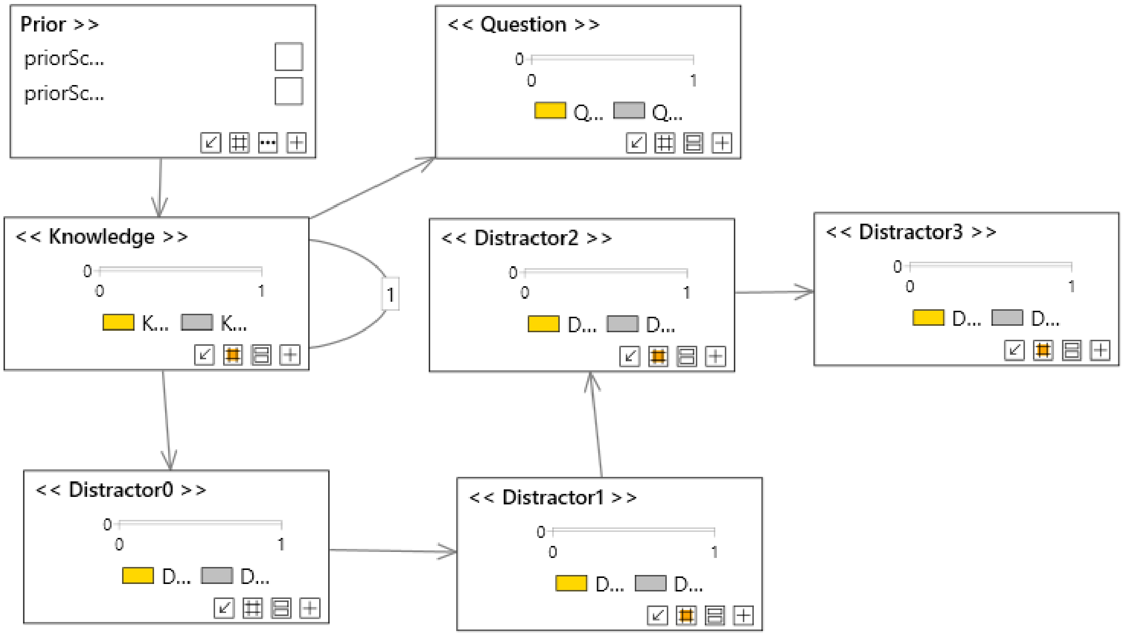
The affect tracking algorithm detected the state of boredom, flow, and frustration. However, there are other affective states such as confusion that may occur during a learning activity (Craig, Graesser, et al., 2004). Therefore, creating an algorithm that could detect these states, and then use them to adapt the game play accordingly, is a possible avenue for future research.

Another possible area of study is the re-evaluation of the second research question with the newer affect tracking algorithm. As indicated previously, the previous affect tracking algorithm was not reliable and therefore the second research question could not be answered conclusively. Once the reliability and validity of the new algorithm are established, it could be used to adapt the game by itself and re-evaluate the second hypothesis.

The DBN used in these experiments could be further optimized. As an example, instead of using three distractors, the game could use an unfixed number of distractors and the DBN be modified accordingly. Further, alternate DBN which has a different structure than the one used in these experiments could also be tested. Figure 17 shows an alternate DBN that has 4 distractors and a different network structure than the one used in the current experiments. In this DBN the probability of picking the first distractor depended on the knowledge level, but the probability of picking further distractors depended only on the previous distractor that was collected, instead of the knowledge node.

The experiments used affect tracking, stealth assessment, and student model created using DBN to adapt the game play. The current literature does not conclude that the game adaptation can boost the learning and engagement of players, and is largely governed by the method which is used to adapt the game. Many methods could be used for stealth assessment as indicated in (Craig, D'Mello, et al., 2004) and therefore a plethora of ways exist that could be used to adapt any game. A potential area of further work is to evaluate different game adaptation techniques and their effect on the situation-specific and state variables identified by Plass et al. (2013).

Figure 17. Alternate DBN with a different network structure.



The results from the first experiment indicated that the adaptation is effective for low domain learners only. Although the same results could not be replicated from the second experiment, it provides for a direction in future studies. A CAGE game could employ adaptation only for a low domain learner and that too until their skill level reaches a predefined threshold. Once this threshold is reached, the game adaptation should be turned off. A game employing such a strategy should be tested and may prove to be more valuable in a CAGE game.

Another potential area of work is a study involving a CAGE game with a wider and diverse range of the population. Most of the CAGE studies that exist involved undergraduate participants from a university and the sample size was rather small. Therefore, the CAGE framework should be evaluated using a large-scale study that involves participants from a broader population.

10.3 Conclusion

This dissertation provides a practical way to implement the Dynamic Bayesian Networks in a serious game and use them for adapting the game play. A key finding from the current dissertation is that the adaptation has the potential to sustain and may even enhance student engagement in a serious game depending on the content domain. Although it may not improve learning for everyone, it could prove to be valuable in promoting it among the low domain learners.

This dissertation is important for three primary reasons: (1) It will help advance the field of educational video games and take a step forward in bringing them into the regular school classrooms (Tüzün, 2007); (2) It will aid in the quick development of multiple educational games with assessment embedded into it, removing the need for explicit examination of the students to gauge their learning (Baron, 2017; Shute et al., 2010); (3) It will help tailor the educational games to the needs of the specific students using the student model (García et al., 2007).

NOTES

1. Angry Birds by Rovio, 2009.

REFERENCES

Adkins, S. (2016). The 2016-2021 global game-based learning market. *Serious play conference.*

Adkins, S. (2017). The 2017-2022 global game-based learning market. *Serious play conference.*

Ali, A., & Sah, M. (2017). Adaptive game-based e-leaming using semantic web technologies. *2017 International Conference on Open Source Systems & Technologies (ICOSST)*, 15–23.

Amresh, A., Clarke, D., & Beckwith, D. (2014). Gamescapes and simapps: New techniques for integrating rich narratives with game mechanics. *Proceedings of the European conference on games based learning*, *1*, 18–25.

Amresh, A., Verma, V., Baron, T., Salla, R., Clarke, D., & Beckwith, D. (2019). Evaluating gamescapes and simapps as effective classroom teaching tools. *European Conference on Games Based Learning*, 22–XII.

Angry Birds. (2021). *Angry birds slingshot stories.* Retrieved February 14, 2021, from https://www.angrybirds.com/

Atmaja, P. W., Muttaqin, F., & Sugiarto, S. (2020). Facilitating educational contents of different subjects with context-agnostic educational game: A pilot case study. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *6*(1), 53–65.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241.

Baker, R. S., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., & Rossi, L. (2012). Sensor-free automated detection of affect in a cognitive tutor for algebra. *Educational Data Mining 2012.*

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM| Journal of Educational Data Mining*, *1*(1), 3–17.

Baron, T. (2017). *An architecture for designing content agnostic game mechanics for educational burst games* (Doctoral dissertation). Arizona State University.

Baron, T., & Amresh, A. (2015). Word towers: Assessing domain knowledge with non-traditional genres. *European Conference on Games Based Learning*, 638.

Baron, T., Heath, C., & Amresh, A. (2016). Towards a context agnostic platform for design and assessment of educational games. *European Conference on Games Based Learning*, 34.

Bartlett, R. C., Collins, S. D. et al. (2011). *Aristotle's nicomachean ethics*. University of Chicago Press.

Bauer, M. I., Williamson, D. M., Mislevy, R. J., & Behrens, J. T. (2003). Using evidence-centered design to develop advanced simulation-based assessment and training. *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 1495–1502.

BayesServer. (2020). Dynamic bayesian networks - an introduction. Retrieved February 4, 2020, from https://www.bayesserver.com/docs/introduction/dynamic-bayesian-networks

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in human-computer interaction*, *2013*.

Biddiss, E., & Irwin, J. (2010). Active video games to promote physical activity in children and youth: A systematic review. *Archives of pediatrics & adolescent medicine*, *164*(7), 664–672.

Bosch, N., Chen, H., D'Mello, S. K., Baker, R., & Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 267–274.

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, and Evaluation*, *8*(1), 9.

Boyle, L., Hancock, F., Seeney, M., & Allen, L. (2009). The implementation of team based assessment in serious games. *2009 Conference in Games and Virtual Worlds for Serious Applications*, 28–35.

Bull, P., & Argyle, M. (2016). *Posture & gesture*. Elsevier Science. https://books.google.com/books?id=7uZdBgAAQBAJ

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, *132*(3), 354.

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31–34.

Chin, J., Dukes, R., & Gamson, W. (2009). Assessment in simulation and gaming: A review of the last 40 years. *Simulation & Gaming*, *40*(4), 553–568.

Chirkov, V., Ryan, R. M., Kim, Y., & Kaplan, U. (2003). Differentiating autonomy from individualism and independence: A self-determination theory perspective on internalization of cultural orientations and well-being. *Journal of personality and social psychology*, *84*(1), 97.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research*, *86*(1), 79–122.

Conrad, S., Clarke-Midura, J., & Klopfer, E. (2014). A framework for structuring learning assessment in a massively multiplayer online educational game: Experiment centered design. *International Journal of Game-Based Learning (IJGBL)*, *4*(1), 37–59.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253–278.

Craig, S. D., D'Mello, S. K., Gholson, B., Witherspoon, A., Sullins, J., & Graesser, A. (2004). Emotions during learning: The first steps toward an affect sensitive intelligent tutoring system. *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 264–268.

Craig, S. D., D'Mello, S. K., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning. *Cognition and Emotion*, *22*(5), 777–788.

Craig, S. D., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, *29*(3), 241–250.

Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using aleks on student's knowledge and behaviors. *Computers & Education*, *68*, 495–504.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction*, *24*(4), 565–591.

Crisp, G. T. (2014). Assessment in next generation learning spaces. *The future of learning and teaching in next generation learning spaces*. Emerald Group Publishing Limited.

Czikszentmihalyi, M. (1990). Flow: The psychology of optimal experience.

DeCharms, R. (1968). Personal causation. *Academic Press, New York.*

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105.

Deci, E. L., & Vansteenkiste, M. (2003). Self-determination theory and basic need satisfaction: Understanding human development in positive psychology.

Del Blanco, Á., Torrente, J., Marchiori, E. J., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2012). A framework for simplifying educator tasks related to the integration of games in the learning flow. *Educational Technology & Society*, *15*(4), 305–318.

D'Mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, *18*(1-2), 45–80.

D'Mello, S. K., & Graesser, A. (2010). Mining bodily patterns of affective experience during learning. *Educational data mining 2010.*

D'Mello, S. K., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, *10*(2), 174–183.

D'Mello, S. K., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., & Graesser, A. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. *International conference on intelligent tutoring systems*, 245–254.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.

Ekman, P., & Friesen, W. (1978). Facial action coding system: A technique for the measurement of facial movement, consulting psychologists press. *Palo Alto*.

El-Nasr, M. S., Drachen, A., & Canossa, A. (2016). *Game analytics*. Springer.

Entertainment Software Association. (2020). *2020 essential facts about the video game industry*. Retrieved February 4, 2020, from https://www.theesa.com/esa-research/2020-essential-facts-about-the-video-game-industry/

Eseryel, D., Ifenthaler, D., & Ge, X. (2011). Alternative assessment strategies for complex problem solving in game-based learning environments. *Multiple perspectives on problem solving and learning in the digital age* (pp. 159–178). Springer.

Esposito, N. (2005). A short and simple definition of what a videogame is.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, *97*(2), 201.

Felder, R. M., & Brent, R. (2005). Understanding student differences. journal of engineering education.

Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Engineering education*, *78*(7), 674–681.

Fernet, C., Guay, F., & Senécal, C. (2004). Adjusting to job demands: The role of work self-determination and job control in predicting burnout. *Journal of vocational behavior*, *65*(1), 39–56.

Fortier, M. S., Sweet, S. N., O'Sullivan, T. L., & Williams, G. C. (2007). A self-determination process model of physical activity adoption in the context of a randomized controlled trial. *Psychology of Sport and Exercise*, *8*(5), 741–757.

Freire, M., Serrano-Laguna, Á., Manero, B., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. *Learning, design, and technology* (pp. 1–29). Springer Nature Switzerland AG.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2-3), 131–163.

Gagne, R. M., Wager, W. W., Golas, K. C., Keller, J. M., & Russell, J. D. (2005). Principles of instructional design, 5th edition. *Performance Improvement*, *44*(2), 44–46. https://doi.org/10.1002/pfi.4140440211

García, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating bayesian networks' precision for detecting students' learning styles. *Computers & Education*, *49*(3), 794–808.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & gaming*, *33*(4), 441–467.

Gilbert, B. (2020). *Video-game industry revenues grew so much during the pandemic that they reportedly exceeded sports and film combined.* Retrieved February 4, 2020, from https://www.msn.com/en-us/entertainment/gaming/video-game-industry-revenues-grew-so-much-during-the-pandemic-that-they-reportedly-exceeded-sports-and-film-combined/ar-BB1cbfN0

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic bulletin & review*, *9*(3), 558–565.

Gronlund, N. E. (1998). *Assessment of student achievement.* ERIC.

Halm, D. S. (2015). The impact of engagement on student learning. *International Journal of Education and Social Science*, *2*(2), 22–33.

Hamel, L., Mislevy, R., & Kennedy, C. A. (2006). *A guide to the padi gradebook.* Retrieved February 4, 2020, from https://padi.sri.com/downloads/TR12_Gradebook.pdf

Harley, J. M. (2016). Measuring emotions: A survey of cutting edge methodologies used in computer-based learning environment research. *Emotions, technology, design, and learning* (pp. 89–114). Elsevier.

Harter, S. (1978). Effectance motivation reconsidered. toward a developmental model. *Human development*, *21*(1), 34–64.

Hines, P. J., Jasny, B. R., & Mervis, J. (2009). Adding a t to the three r's.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental science*, *12*(4), F9–F15.

IJsselsteijn, W., Van Den Hoogen, W., Klimmt, C., De Kort, Y., Lindley, C., Mathiak, K., Poels, K., Ravaja, N., Turpeinen, M., & Vorderer, P. (2008). Measuring the experience of digital game enjoyment. *Proceedings of measuring behavior*, *2008*, 88–89.

iMotions Inc. (2018). Affectiva channel explained [Accessed: 2020-04-17].

Jackman, S. (2017). Pscl: Classes and methods for r developed in the political science computational laboratory. r package version 1.5. 2.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 50–67.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review*, *19*(4), 509–539.

Kelley, D. (1998). *The art of reasoning*. WW Norton.

Kennedy, S., Goggin, K., & Nollen, N. (2004). Adherence to hiv medications: Utility of the theory of self-determination. *Cognitive therapy and research*, *28*(5), 611–628.

Kenny, R. F., & McDaniel, R. (2011). The role teachers' expectations and value assessments of video games play in their adopting and integrating them into their classrooms. *British Journal of Educational Technology*, *42*(2), 197–213.

Kiili, K., & Ketamo, H. (2017). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies*, *11*(2), 255–263.

Kim, Y. J., & Shute, V. (2015). Opportunities and challenges in assessing and supporting creativity in video games. *Video games and creativity* (pp. 99–117). Elsevier.

Kim, Y. J., Almond, R. G., & Shute, V. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, *16*(2), 142–163.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, *43*, 43–56.

Lee, M. J., Ko, A. J., & Kwan, I. (2013). In-game assessments increase novice programmers' engagement and level completion speed. *Proceedings of the ninth annual international ACM conference on International computing education research*, 153–160.

Lepp, A. (2018). Correlating leisure and happiness: The relationship between the leisure experience battery and the satisfaction with life scale. *Annals of Leisure Research*, *21*(2), 246–252. https://doi.org/10.1080/11745398.2017.1325759

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis. *Journal of Personality and social Psychology*, *28*(1), 129.

Litman, D. J., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech communication*, *48*(5), 559–590.

Magdin, M., & Prikler, F. (2018). Real time facial expression recognition using webcam and sdk affectiva. *IJIMAI*, *5*(1), 7–15.

Mayer, I., van Dierendonck, D., Van Ruijven, T., & Wenzler, I. (2013). Stealth assessment of teams in a digital game environment. *International Conference on Games and Learning Alliance*, 224–235.

Mayer, R. E. (2005). Introduction to multimedia learning. *The Cambridge handbook of multimedia learning*, *2*, 1–24.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43–52.

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. e. (2016). Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 3723–3726. https://doi.org/10.1145/2851581.2890247

McFarlane, A., Sparrowhawk, A., & Head, Y. (2002). Report on the educational use of games. teem (teacher evaluating educational multimedia). retrieved february 13, 2010.

Merceron, A., & Yacef, K. (2005). Educational data mining: A case study. *AIED*, 467–474.

Mislevy, R. J. (2011). Evidence-centered design for simulation-based assessment. cresst report 800. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29.

Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., Robinson, D. H., Rutstein, D. W., Shute, V., Stanley, K., et al. (2010). On the roles of external knowledge representations in assessment design.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of educational data mining*, *4*(1), 11–48.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice*, *25*(4), 6–20.

Mislevy, R. J., & Levy, R. (2006). 26 bayesian psychometric modeling from an evidence-centered design perspective. *Handbook of statistics*, *26*, 839–865.

Moreno-Ger, P., Martinez-Ortiz, I., Freire, M., Manero, B., & Fernandez-Manjon, B. (2014). Serious games: A journey from research to application. *2014 IEEE Frontiers in education conference (FIE) proceedings*, 1–4.

Muldner, K., Burleson, W., & VanLehn, K. (2010). “Yes”!: Using tutor and sensor data to predict moments of delight during instructional activities. *International Conference on User Modeling, Adaptation, and Personalization*, 159–170.

Neapolitan, R. E. (2004). *Learning bayesian networks* (Vol. 38). Pearson Prentice Hall Upper Saddle River, NJ.

Nithya, P., Umamaheswari, B., & Umadevi, A. (2016). A survey on educational data mining in field of education. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *5*(1), 69–78.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human behavior*, *24*(5), 2415–2433.

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, *1*(1), 107–128.

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. *International Conference on User Modeling, Adaptation, and Personalization*, 255–266.

Park, C. (2003). Engaging students in the learning process: The learning journal. *Journal of Geography in Higher Education*, *27*(2), 183–199.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105. https://doi.org/10.1207/S15326985EP3702_4

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432–1462.

Plass, J. L., Homer, B. D., Kinzer, C. K., Chang, Y. K., Frye, J., Kaczetow, W., Isbister, K., & Perlin, K. (2013). Metrics in simulations and games for learning. *Game analytics* (pp. 697–729). Springer.

Plass, J. L., Homer, B. D., Pawar, S., Brenner, C., & MacNamara, A. P. (2019). The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cognitive Development*, *49*, 56–67.

R Core Team. (2013). *R: A language and environment for statistical computing* [ISBN 3-900051-07-0]. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Reichenberg, R. (2018). Dynamic bayesian networks in educational measurement: Reviewing and advancing the state of the field. *Applied Measurement in Education*, *31*(4), 335–350.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*(1), 63–96.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, *1*(3), 181–210.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Rosser, J. C., Lynch, P. J., Cuddihy, L., Gentile, D. A., Klonsky, J., & Merrell, R. (2007). The impact of video games on training surgeons in the 21st century. *Archives of surgery*, *142*(2), 181–186.

Roth, G., Assor, A., Kanat-Maymon, Y., & Kaplan, H. (2007). Autonomous motivation for teaching: How self-determined teaching may lead to self-determined learning. *Journal of educational psychology*, *99*(4), 761.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning and Assessment*, *8*(4).

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54–67.

Ryan, R. M., LaGuardia, J. G., & Rawsthorne, L. J. (2005). Self-complexity and the authenticity of self-aspects: Effects on well being and resilience to stressful events. *North American Journal of Psychology*, *7*(3).

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, *30*(4), 344–360.

Sabourin, J., Mott, B., & Lester, J. C. (2011). Modeling learner affect with theoretically grounded dynamic bayesian networks. *International conference on affective computing and intelligent interaction*, 286–295.

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, *69*, 452–462.

Scheuer, O., & McLaren, B. M. (2012). Educational data mining. *Encyclopedia of the Sciences of Learning*, 1075–1079.

Schmierbach, M. (2017). Immersion in games exemplifies why digital media create complex responses to ads. *Digital advertising: Theory and research, third edition* (pp. 427–430). Taylor; Francis.

Sharek, D., & Wiebe, E. (2015). Investigating real-time predictors of engagement: Implications for adaptive videogames and online training. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, *7*(1), 20–37.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2014). Student engagement in high school classrooms from the perspective of flow theory. *Applications of flow in human development and education* (pp. 475–494). Springer.

Shute, V. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503–524.

Shute, V., & Ke, F. (2012). Games, learning, and assessment. *Assessment in game-based learning* (pp. 43–58). Springer.

Shute, V., & Kim, Y. J. (2011). Does playing the world of goo facilitate learning. *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning*, 359–387.

Shute, V., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., & Wang, C.-Y. (2010). Modeling, assessing, and supporting key competencies within game environments. *Computer-based diagnostics and systematic analysis of knowledge* (pp. 281–309). Springer.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*.

Shute, V., & Spector, J. M. (2008). Scorm 2.0 white paper: Stealth assessment in virtual worlds. *Unpublished manuscript*.

Shute, V., & Torres, R. (2012). Where streams converge: Using evidence-centered design to assess quest to learn. *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*, *91124*.

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. The mit press.

Shute, V., Ventura, M., & Ke, F. (2015). The power of play: The effects of portal 2 and lumosity on cognitive and noncognitive skills. *Computers & education*, *80*, 58–67.

Shute, V., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in newton's playground. *The Journal of Educational Research*, *106*(6), 423–430.

Shute, V., Ventura, M., Small, M., & Goldberg, B. (2013). Modeling student competencies in video games using stealth assessment. *Design recommendations for intelligent tutoring systems*, *1*, 141–152.

Shute, V., & Wang, L. (2015). Measuring problem solving skills in portal 2. *E-learning systems, environments and approaches* (pp. 11–24). Springer.

Sicart, M. (2008). Defining game mechanics. *Game Studies*, *8*(2).

Skinner, B. F. (2019). *The behavior of organisms: An experimental analysis*. BF Skinner Foundation.

Soenens, B., Vansteenkiste, M., Lens, W., Luyckx, K., Goossens, L., Beyers, W., & Ryan, R. M. (2007). Conceptualizing parental autonomy support: Adolescent perceptions of promotion of independence versus promotion of volitional functioning. *Developmental psychology*, *43*(3), 633.

Soloman, B. A., & Felder, R. M. (2005). Index of learning styles questionnaire. *NC State University. Available online at: https://www.engr.ncsu.edu/stem-resources/legacy-site/learning-styles/ (last visited on 02.14.2021)*, *70*.

Sørebø, Ø., & Hæhre, R. (2012). Investigating students' perceived discipline relevance subsequent to playing educational computer games: A personal interest and self-determination theory approach. *Scandinavian Journal of Educational Research*, *56*(4), 345–362.

Squire, K. (2003). Video games in education. *Int. J. Intell. Games & Simulation*, *2*(1), 49–62.

Steinberg, S. (2012). *The modern parent's guide to kids and video games*. Lulu.com.

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating rasch/irt parameters with dichotomous items. *Journal of applied measurement*.

Subani, H. (2009). *In-game advertising; t-mobile sidekick in need for speed undercover*. Retrieved February 14, 2021, from https://www.techtangerine.com/2009/07/30/in-game-advertising-t-mobile-sidekick-in-need-for-speed-undercover

Tadayon, R., Amresh, A., McDaniel, T., & Panchanathan, S. (2018). Real-time stealth intervention for motor learning using player flow-state. *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)*, 1–8.

Tang, S., Hanneghan, M., & El Rhalibi, A. (2009). Introduction to games-based learning. *Games-based learning advancements for multi-sensory human computer interfaces: Techniques and effective practices* (pp. 1–17). IGI Global.

Taylor, T. (2012). Ceiling effect. encyclopedia of research design.

Tüzün, H. (2007). Blending video games with learning: Issues and challenges with classroom implementations in the turkish context. *British Journal of Educational Technology*, *38*(3), 465–477.

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE review*, *41*(2), 16.

Vanbecelaere, S., Van den Berghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., & Depaepe, F. (2020). The effectiveness of adaptive versus non-adaptive learning with digital educational games. *Journal of Computer Assisted Learning*, *36*(4), 502–513.

van Oostendorp, H., Van der Spek, E. D., & Linssen, J. (2014). Adapting the complexity level of a serious game to the proficiency of players. *EAI Endorsed Trans. Serious Games*, *1*(2), e5.

Ventura, M., Shute, V., & Small, M. (2014). Assessing persistence in educational games. *Design recommendations for adaptive intelligent tutoring systems: Learner modeling*, *2*(2014), 93–101.

Verma, V., Baron, T., Bansal, A., & Amresh, A. (2019). Emerging practices in game-based assessment. *Game-based assessment revisited* (pp. 327–346). Springer.

Vygotsky, L. (1978). Zone of proximal development: A new approach. *Mind in society: The development of higher psychological processes*, 84–91.

Walker, A. A., & Engelhard Jr, G. (2014). Game-based assessments: A promising way to create idiographic perspectives. *Measurement: Interdisciplinary Research & Perspectives*, *12*(1-2), 57–61.

West, D. M., & Bleiberg, J. (2013). Issues in governance studies.

Wiebe, E. N., Lamb, A., Hardy, M., & Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the user engagement scale. *Computers in Human Behavior*, *32*, 123–132.

Williams, D., Yee, N., & Caplan, S. E. (2008). Who plays, how much, and why? debunking the stereotypical gamer profile. *Journal of computer-mediated communication*, *13*(4), 993–1018.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in r. *Journal of statistical software*, *27*(8), 1–25.

Zhou, L., Li, F., Wu, S., & Zhou, M. (2020). “School’s Out, But Class’s On”, the largest online education in the world today: Taking china’s practical exploration during the covid-19 epidemic prevention and control as an example. *Best Evidence in Chinese Education*, *4*(2), 501–519. https://doi.org/10.15354/bece.20.ar023

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79–87.

APPENDIX A

CHEMISTRY FUN FACTS

1. A catalyst is a substance that changes the rate of a chemical reaction, but is chemically unchanged at the end of the reaction. An inhibitor does the opposite - it slows down chemical reactions.
2. During World War I, Haber's process provided Germany with a source of ammonia for the production of explosives, compensating for the Allied trade blockade on Chilean saltpeter.
3. Enzymes are proteins that act as catalysts in biochemical reactions.
4. Common types of catalysts include enzymes, acid-base catalysts, and heterogeneous (or surface) catalysts.
5. A single chlorine atom is able to react with an average of 100,000 ozone molecules before it is removed from the catalytic cycle and thus cause a lot of damage to the ozone layer protecting us.
6. The Antarctic ozone hole is an area of the Antarctic stratosphere in which the recent ozone levels have dropped to as low as 33% of their pre-1975 values.
7. Many therapeutic drugs are enzyme inhibitors. Important examples are penicillin, which inhibits an enzyme necessary for bacterial cell wall synthesis, and aspirin, an inhibitor of the synthesis of molecules that mediate pain and swelling.
8. In 1995 Stuart Kauffman proposed that life initially arose as auto-catalytic chemical networks.
9. Catalysts break down paper pulp to produce the smooth paper in your magazine.
10. At its heart, a catalyst is a way to save energy.
11. In the absence of catalysis, it takes several weeks for starch to hydrolyze to glucose; a trace of the enzyme ptyalin, found in human saliva, accelerates the reaction so that starches can be digested.
12. More than 90 percent of the chemical products are made using catalyst.
13. The deliberate application of catalysts to industrial processes was undertaken in the 19th century. P. Phillips, an English chemist, patented the use of platinum to oxidize sulfur dioxide to sulfur trioxide with air.

APPENDIX B

CRYPTOGRAPHY FUN FACTS

1. Mechanical Ciphers are those that were developed around the second World War, which rely on sophisticated gearing mechanisms to encipher text.
2. Atbash cipher is used in the Bible. The Old Testament, in Jeremiah 25:26 and 51:41, uses the name 'Sheshach' in place of 'Babel'.
3. The Atbash cipher is trivial to break since there is no key, as soon as you know it is an Atbash cipher you can simply decrypt it.
4. The Enigma machines were a series of electro-mechanical rotor cipher machines developed and used in the early-to mid-20th century to protect commercial, diplomatic and military communication.
5. The film, The Imitation Game (2014) tells the story of Alan Turing and his attempts to crack the Enigma machine code during World War II.
6. Alan Turing cracked the Germany's Enigma code shortening the World War II by two to four years and saving an estimated 14 million to 21 million lives, historians claim.
7. Chinese writing is not conducive to cryptography, but they did make use of steganography. They would write a message on silk ribbon and ball it up, cover it in wax and have a courier swallow it or insert it in his rectum.
8. Julius Caesar (100BC-44BC) is credited as the first person to use a cipher in military affairs.
9. Cryptanalysis is the art of breaking codes and ciphers. The Caesar cipher is probably the easiest of all ciphers to break.
10. Natural English text has a very distinct frequency distribution of letters that can be used to help crack codes.
11. The Enigma cipher machine was invented by a German engineer, Arthur Scherbius, who applied for his patent on February 23, 1918.
12. Enigma machine used by the Nazis had the key space of 10^{23}, which means 100,000 operators, each checking one key setting every second would take twice the age of the universe to break the code. Despite these overwhelming odds, the Allies did just that.
13. The theoretical key space of Enigma machine is 3×10^{114}, which is far larger than the number of atoms in the universe.
14. The battle of wits between codemakers and codebreakers has been the driving force for innovation in cipher technology for centuries.
15. The knowledge of the Allies breaking the Nazi Enigma code in WW2 was kept secret for 29 years, despite over 15,000 people working to break that code.
16. Although most people claim they're not familar with cryptography, they are often familar with the concept of ciphers, whether or not they are actually conscious of it. Are you?
17. The ROT13 cipher is trivial to break since there is no key, as soon as you know it is an ROT13 cipher you can simply decrypt it.
18. Encryption is a term that comes from the science of cryptography. It includes the coding and decoding of messages in order to protect their contents.

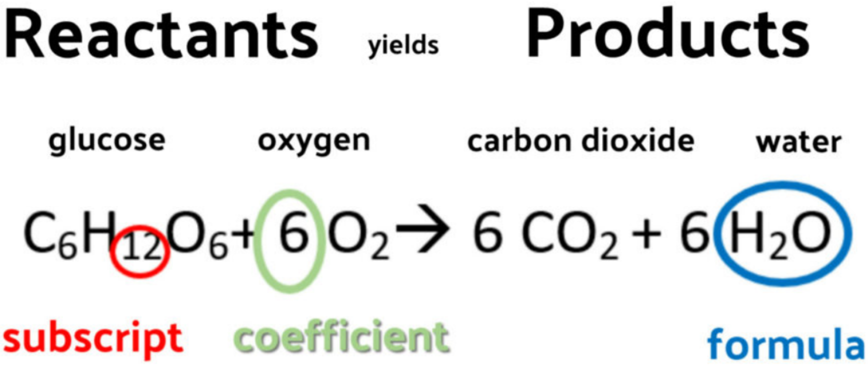
19. The oldest encryption attempt known to mankind dates back to the kingdom of Egypt, around two thousand years before Christ. The ciphers are found on the tomb of Khnumhotep II. They may have been, however, a joke or an attempt to create a mystic atmosphere.
20. Cryptography comes from the Greek words kryptos and graphein, which mean hidden and writing, respectively (Pawlan, 1998).
21. Wonder why all those websites you sign up for require a password? This is your access to the public key, and gives the company the ability to use the private key on your private information.
22. In India around 400 BCE to 200 CE, Mlecchita vikalpa or the art of understanding writing in cypher, and the writing of words in a peculiar way was documented in the Kama Sutra for the purpose of communication between lovers.

APPENDIX C

CHEMISTRY READING MATERIAL

C.1 Reading chunk 1

A chemical equation is a written description of what happens in a chemical reaction. The starting materials, called reactants, are listed on the lefthand side of the equation. Next comes an arrow that indicates the direction of the reaction. The righthand side of the reaction lists the substances that are made, called products.



C.2 Reading chunk 2

A balanced chemical equation tells you the amounts of reactants and products needed to satisfy the Law of Conservation of Mass. Basically, this means there are the same numbers of each type of atoms on the left side of the equation as there are on the right side of the equation. It sounds like it should be simple to balance equations, but it's a skill that takes practice. Here's the process you follow, step by step, to balance equations. You can apply these same steps to balance any unbalanced chemical equation...

The first step is to write down the unbalanced chemical equation. Let's practice using a reaction from real life, the burning of propane C_3H_8 in the presence of oxygen to produce water and carbon dioxide. To write the reaction, you need to identify the reactants (propane and oxygen) and the products (water and carbon dioxide). Unbalanced chemical equation for the example is:

$$C_3H_8 + O_2 \longrightarrow H_2O + CO_2$$

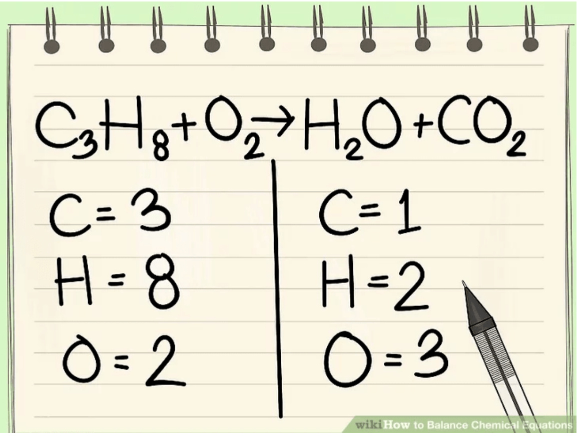
Note the reactants always go on the left side of the arrow. A 'plus' sign separates them. Next there is an arrow indicating the direction of the reaction (reactants become products). The products are always on the right side of the arrow. The order in which you write the reactants and products is not important.

C.3 Reading chunk 3

The next step for balancing the chemical equation is to determine how many atoms of each element are present on each side of the arrow. To do this, keep in mind a subscript indicates the number of atoms. For example, O_2 has 2 atoms of oxygen. There are 3 atoms of carbon and 8 atoms of hydrogen in C_3H_8. When there is no subscript, it means there is 1 atom.

On the reactant side: 3 C, 8 H and 2 O

On the product side: 1 C, 2 H and 3

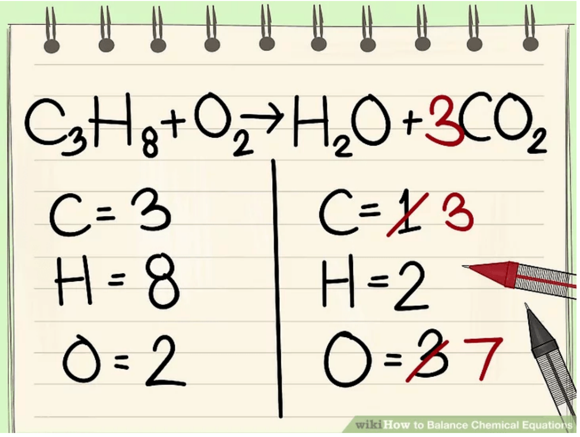

$C_3H_8 + O_2 \rightarrow H_2O + CO_2$
C=3
H=8
O=2
C=1
H=2
O=3
wikiHow to Balance Chemical Equations

C.4 Reading chunk 4

How do you know the equation isn't already balanced? Because the number of atoms on each side isn't the same! Conservation of Mass states mass isn't created or destroyed in a chemical reaction, so you need to add coefficients in front of the chemical formulas to adjust the number of atoms so they will be the same on both sides. When balancing equations, you never change subscripts. You add coefficients. Coefficients are whole number multipliers. If, for example, you write 2 H_2O, that means you have 2 times the number of atoms in each water molecule, which would be 4 hydrogen atoms and 2 oxygen atoms. As with subscripts, you don't write the coefficient of '1', so if you don't see a coefficient, it means there is one molecule.

C.5 Reading chunk 5

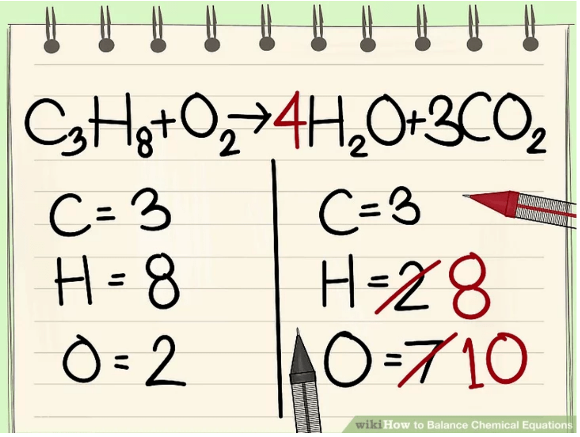
There is a strategy that will help you balance equations more quickly. It is called balancing by inspection. Basically, you look at how many atoms you have on each side of the equation and add coefficients to the molecules to balance out the number of atoms. Balance atoms present in a single molecule of reactant and product first. Balance any oxygen or hydrogen atoms last. The reason is because they usually appear in multiple reactants and products, so if you tackle them first you're usually making extra work for yourself. In the example, carbon is present in one reactant and one product, so balance its atoms first. There are three atoms of carbon on the left and one on the right, so put a coefficient of 3 on the right as shown below.


$C_3H_8+O_2 \rightarrow H_2O+3CO_2$
C=3
H=8
O=2
C=~~1~~ 3
H=2
O=~~3~~ 7
wiki How to Balance Chemical Equations

C.6 Reading chunk 6

While that would balance carbon, you already know you're going to have to adjust oxygen, too, because it isn't balanced. Since you have balanced all atoms besides the hydrogen and oxygen, you can address the hydrogen atoms. You have 8 on the left side. So you'll need 8 on the right side. Use a coefficient to achieve this as shown below.

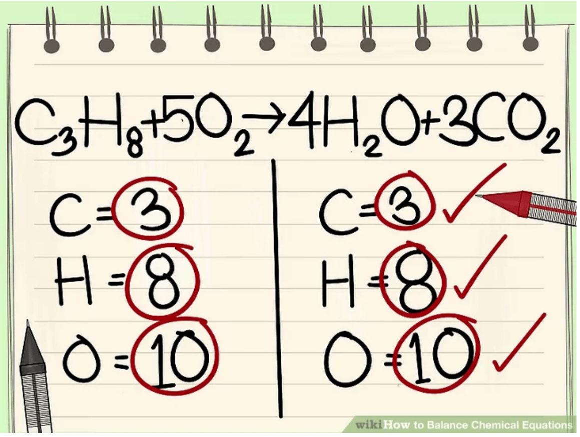
On the right side, you now added a 4 as the coefficient because the subscript showed that you already had 2 hydrogen atoms. When you multiply the coefficient 4 times by the subscript 2, you end up with 8. The other 6 atoms of oxygen come from $3\,CO_2$ ($3 \times 2 = 6$ atoms of oxygen + the other $4 = 10$). Remember to account for the


$C_3H_8+O_2 \rightarrow 4H_2O+3CO_2$
C=3
H=8
O=2
C=3
H=~~2~~ 8
O=~~7~~ 10
wiki How to Balance Chemical Equations

coefficients that you've used to balance out the other atoms. Because you've added coefficients to the molecules on the right side of the equation, the number of oxygen atoms has changed.

C.7 Reading chunk 7

You now have 4 oxygen atoms in the water molecules and 6 oxygen atoms in the carbon dioxide molecules. That makes a total of 10 oxygen atoms. Add a coefficient of 5 to the oxygen molecule on the left side of the equation. You now have 10 oxygen atoms on each side. The carbon, hydrogen, and oxygen atoms are balanced. Your equation is complete.


$C_3H_8+5O_2 \rightarrow 4H_2O+3CO_2$
C=3
H=8
O=10
C=3 ✓
H=8 ✓
O=10 ✓
wiki How to Balance Chemical Equations

C.8 Reading chunk 8

Note: You could have written a balanced equation using multiples of the coefficients. For example, if you double all of the coefficients, you still have a balanced equation:

$$2\,C_3H_8 + 10\,O_2 \longrightarrow 8\,H_2O + 6\,CO_2$$

However, chemists always write the simplest equation, so check your work to make sure you can't reduce your coefficients.
List of chemical symbols of elements used in the game:

1. O - Oxygen
2. N - Nitrogen
3. H - Hydrogen
4. Zn - Zinc
5. S - Sulfur
6. Cl - Chlorine
7. Al - Aluminium
8. C - Carbon
9. Na - Sodium
10. Fe - Iron
11. P - Phosphorous
12. Ba - Barium
13. Ca - Calcium
14. Ag - Silver
15. K - Potassium
16. Cu - Copper
17. Mn - Manganese

APPENDIX D

CRYPTOGRAPHY READING MATERIAL

D.1 Reading chunk 1

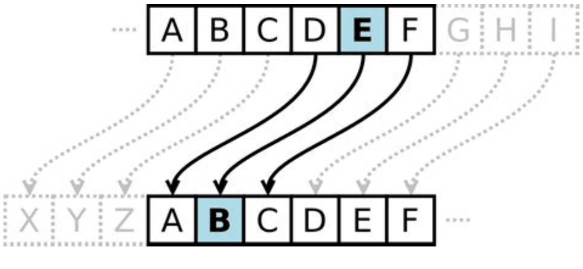
Virtually anyone who can read will have come across codes or ciphers in some form. Even an occasional attempt at solving crosswords, for example, will ensure that the reader is acquainted with anagrams, which are a form of cipher known as *transpositions*. Enciphered messages also appear in children's toys such as secret decoder rings, children's comics, the personal columns of newspapers and stories by numerous authors from at least as far back as Conan Doyle and Edgar Allan Poe.

The Caesar cipher is one of the earliest known and simplest ciphers. It is a type of substitution cipher in which each letter in the plaintext is 'shifted' a certain number of places down the alphabet. For example, with a shift of 1, A would be replaced by B, B would become C, and so on. The method is named after Julius Caesar, who used it in his private correspondence. More complex encryption schemes employ the Caesar cipher as one element of the encryption process. To pass an encrypted message from one person to another, it is first necessary that both parties have the 'key' for the cipher, so that the sender may encrypt it and the receiver may decrypt it. For the Caesar cipher, the key is the number of characters to shift the cipher alphabet. Simon Singh's 'The Code Book' is an excellent introduction to ciphers and codes, and includes a section on Caesar ciphers. Encryption of a letter x by a shift n can be described mathematically as,

$$E_n(x) = (x + n) \mod 26$$

D.2 Reading chunk 2

Following is an example where key is 23 (or -3), thus each occurrence of ‘E’ in the plaintext becomes ‘B’ in the ciphertext.



The replacement remains the same throughout the message, so the cipher is classed as a type of monoalphabetic substitution, as opposed to polyalphabetic substitution.

Kahn (1967) describes instances of lovers engaging in secret communications enciphered using the Caesar cipher in The Times. Caesar ciphers can be found today in children's toys such as secret decoder rings.

Cryptanalysis is the art of breaking codes and ciphers. The Caesar cipher is probably the easiest of all ciphers to break. If you happen to know what a piece of the ciphertext is, or you can guess a piece, then this will allow you to immediately find the key. The method used is to take the ciphertext, try decrypting it with each key, then see which decryption looks like English text. This simplistic method of cryptanalysis only works on very simple ciphers such as the Caesar cipher and the rail fence cipher, even slightly more complex ciphers can have far too many keys to check all of them.

D.3 Reading chunk 3

The Caesar cipher can be broken using the same techniques as for a general simple substitution cipher, such as frequency analysis or pattern words. For example, in the English language the plaintext frequencies of the letters E, T, (usually most frequent), and Q, Z (typically least frequent) are particularly distinctive. Since there are only a limited number of possible shifts (26 in English), they can each be tested in turn in a brute force attack.

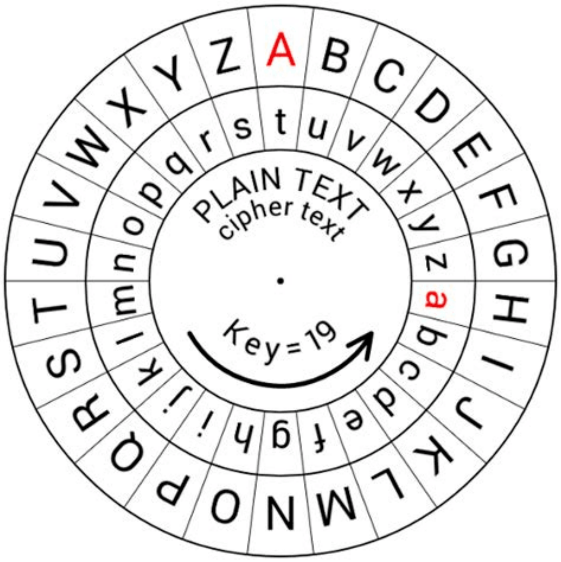
For natural language plaintext, there will typically be only one plausible decryption, although for extremely short plaintexts, multiple candidates are possible. For example, the ciphertext MPQY could, plausibly, decrypt to either 'aden' or 'know' (assuming the plaintext is in English); similarly, 'ALIIP' to 'dolls' or 'wheel'. Decryption of a letter x by a shift n can be described mathematically as,

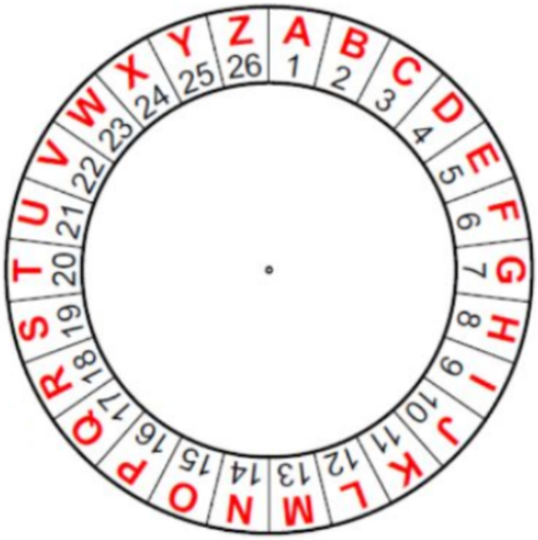
$$D_n(x) = (x - n) \mod 26$$

With the Caesar cipher, encrypting a text multiple times provides no additional security. This is because two encryptions of, say, shift A and shift B, will be equivalent to a single encryption with shift $A + B$. A construction of 2 rotating disks with a Caesar cipher can be used to encrypt or decrypt the code.

D.4 Reading chunk 4

Following picture illustrate the rotating disks corresponding to $key = 19$.


A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
t u v w x y z a b c d e f g h i j k l m n o p q r s
PLAIN TEXT
cipher text
Key = 19



D.5 Reading chunk 5

In the 19th century, the personal advertisements section in newspapers would sometimes be used to exchange messages encrypted using simple cipher schemes.

In April 2006, fugitive Mafia boss Bernardo Provenzano was captured in Sicily partly because some of his messages, clumsily written in a variation of the Caesar cipher, were broken. In 2011, Rajib Karim was convicted in the United Kingdom of 'terrorism offences' after using the Caesar cipher to communicate with Bangladeshi Islamic activists discussing plots to blow up British Airways planes or disrupt their IT networks.

APPENDIX E

EXPERIMENT 1: PRE AND POST TEST FOR CHEMISTRY

1. Which of the following depicts a chemical equation?

   a) $a = b \mod c$ b) $a^2 + b^2 = c^2$

   c) $\mathbf{H_2 + O_2 \longrightarrow H_2O}$ d) $F = m \times a$

2. Which of the following is a reactant in this equation?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) $\mathbf{O_2}$ b) H_4

   c) H_2O d) H_2

3. Which of the following is a product in this equation?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) O_2 b) H_4

   c) $\mathbf{H_2O}$ d) H_2

4. What is the subscript of Oxygen on the reactant side?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 4 b) **2**

   c) 1 d) 0

5. What is the subscript of Hydrogen on the product side?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 4 b) **2**

   c) 1 d) 0

6. What is the coefficient of O on the reactant side?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 4 b) **1**

   c) 2 d) 0

7. What is the coefficient of H_2O on the reactant side?
   $CH_4 + O_2 \longrightarrow CO_2 + 3\,H_2O$

   a) 1 b) **3**

   c) 2 d) 4

8. Is the following chemical equation balanced?
   $H_2 + O_2 \longrightarrow H_2O$

   a) **No, adding coefficient 2 to $\mathbf{H_2O}$ and $\mathbf{H_2}$ will balance it**

   b) No, adding coefficient 4 to H_2O will balance it

   c) No, adding coefficient 2 to H_2O will balance it

   d) Yes

9. Which of the following represents a balanced chemical equation?

   a) $\mathbf{Ca + S \longrightarrow CaS}$ b) $Ca + S \longrightarrow 2\,CaS$

   c) $Ca + 2\,S \longrightarrow CaS$ d) $2\,Ca + S \longrightarrow CaS$

10. Which of the following represents a balanced chemical equation?

   a) $\mathbf{2\,H_2 + O_2 \longrightarrow 2\,H_2O}$ b) $H_2 + O_2 \longrightarrow H_2O$

   c) $H_2 + O_2 \longrightarrow 2\,H_2O$ d) $2\,H_2 + 2\,O_2 \longrightarrow 2\,H_2O$

11. If you change the coefficient of H_2O on the product side to 2, what should be the coefficient of O on the reactant side to balance the equation?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 1 b) **2**

   c) 4 d) 1/2

12. If you change the coefficient of CH_4 on the reactant side to 3, what should be the coefficient of CO_2 on the product side to balance the number of carbon atoms in this equation?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 1 b) **3**

   c) 4 d) 2

13. What is the number of O atoms in this equation on the product side?
   $CH_4 + O_2 \longrightarrow 2\,CO_2 + 3\,H_2O$

   a) 5 b) **7**

   c) 3 d) 2

14. What is the number of O atoms in the following equation on the products side?
   $CaCl_2 + 2\,AgNO_3 \longrightarrow Ca(NO_3)_2 + 2\,AgCl$

   a) 3 b) **6**

   c) 4 d) 2

15. Are the number of Hydrogen (H) atoms balanced in this equation?
   $C_3H_8 + O_2 \longrightarrow CO_2 + H_2O$

   a) **No, there are 6 less on the product side**

   b) No, there are 2 more on the product side

   c) No, there are 8 more on the reactant side

   d) Yes

16. If you change the coefficient of H on the reactant side to 3, will it balance the following equation?
   $H_2 + O_2 \longrightarrow 2\,H_2O$

a) No, it should be changed to 4

b) **No, it should be changed to 2**

c) No, it should be removed

d) Yes

17. What are the coefficients of the following equation when it's balanced?
$N_2O_4 \longrightarrow NO_2$

a) 2,3

b) **1,2**

c) 3,4

d) 2,1

18. In the reaction, $xCu + yHNO_3 \longrightarrow Cu(NO_3)_2 + 2\,NO_2 + 2\,H_2O$, the coefficients x and y are:

a) 2,3

b) **1,4**

c) 1,3

d) 3,8

19. Which of the following represents a balanced chemical equation?

a) $\mathbf{2\,PO_4^{3-} + 3\,Ca^{2+} \longrightarrow Ca_3(PO_4)_2}$

b) $2\,PO_4^{3-} + Ca^{2+} \longrightarrow 2\,Ca_3(PO_4)_2$

c) $PO_4^{3-} + {}_3Ca^{2+} \longrightarrow Ca_3(PO_4)_2$

d) $2\,PO_4^{3-} + Ca^{2+} \longrightarrow Ca_3(PO_4)_2$

20. For the reaction, $MnO_4^- + C_2O_4^{2-} + H^+ \longrightarrow Mn^{2+} + CO_2 + H_2O$, the correct coefficients of the reactants in the balanced reaction are:

a) **2,5,16**

b) 16,5,2

c) 5,16,2

d) 2,16,5

APPENDIX F

EXPERIMENT 2: PRE-TEST FOR CHEMISTRY

1. Which of the following depicts a chemical equation?

   a) $a = b \mod c$ b) $a^2 + b^2 = c^2$

   c) $\mathbf{H_2 + O_2 \longrightarrow H_2O}$ d) $F = m \times a$

2. Which of the following is a reactant in this equation?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) $\mathbf{O_2}$ b) H_4

   c) H_2O d) H_2

3. Which of the following is a product in this equation?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) O_2 b) H_4

   c) $\mathbf{H_2O}$ d) H_2

4. What is the subscript of Oxygen on the reactant side?
   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 4 b) **2**

   c) 1 d) 0

5. What is the coefficient of H_2O on the product side?
   $CH_4 + O_2 \longrightarrow CO_2 + 3\,H_2O$

   a) 6 b) **3**

   c) 2 d) 4

6. The value of x + y is —— and the value of z + w is —— in the following equation:
   $xC_4H_{10} + yO_2 \longrightarrow zCO_2 + wH_2O$

   a) 15,15 b) **15,18**

   c) 18,18 d) 14,18

7. The sum of all the coefficients of the reactants and the products when the following equation is balanced is:
   $NaCl + SO_2 + H_2 + O_2 \longrightarrow Na_2SO_4 + HCl$

   a) 16 b) **15**

   c) 17 d) 14

8. Is the following chemical equation balanced?
   $2\,Fe(NO_3)_3 + 3\,(NH_4)_2CO_3 \longrightarrow Fe_2(CO_3)_3 + NH_4NO_3$

a) **No, adding coefficient 6 to NH4NO3 will balance it**

b) No, adding coefficient 3 to NH4NO3 will balance it

c) No, adding coefficient 5 to NH4NO3 will balance it

d) Yes

9. Which of the following represents a balanced chemical equation? (Note: This question had to be removed from analysis, as none of the answer choices were correct, due to typographical error in CO_3 which was supposed to be CO_2)

   a) $Al_2(CO_3)_3 + 2\,H_3PO_4 \longrightarrow 2\,AlPO_4 + CO_3 + 3\,H_2O$

   b) $Al_2(CO_3)_3 + 2\,H_3PO_4 \longrightarrow 2\,AlPO_4 + 3\,CO_3 + H_2O$

   c) $Al_2(CO_3)_3 + H_3PO_4 \longrightarrow 2\,AlPO_4 + 3\,CO_3 + 3\,H_2O$

   d) $Al_2(CO_3)_3 + 2\,H_3PO_4 \longrightarrow 2\,AlPO_4 + 3\,CO_3 + 3\,H_2O$

10. Which of the following represents a balanced chemical equation?

   a) $3\,Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 3\,H_2O + Ba_3(PO_4)_2$

   b) $Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 6\,H_2O + Ba_3(PO_4)_2$

   c) $3\,Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 6\,H_2O + 2\,Ba_3(PO_4)_2$

   d) $\mathbf{3\,Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 6\,H_2O + Ba_3(PO_4)_2}$

11. If you change the coefficient of $Fe(C_2H_3O_2)$ on the product side to 2, what should be the coefficient of $HC_2H_3O_2$ on reactant side to balance the number of Hydrogen atoms?

   $Fe + HC_2H_3O_2 \longrightarrow Fe(C_2H_3O_2)_3 + H_2$

   a) 2 b) **5**

   c) 4 d) 1

12. If you change the coefficient of CH_4 on the reactant side to 3, what should be the coefficient of CO_2 on product side to balance the number of carbon atoms in this equation?

   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 1 b) **3**

   c) 2 d) 4

13. What is the number of O atoms in the following equation on the products side?

   $CH_4 + O_2 \longrightarrow 2\,CO_2 + 3\,H_2O$

   a) 2 b) **7**

   c) 3 d) 5

14. What is the number of O atoms in the following equation on the products side?

$CaCl_2 + 2\,AgNO_3 \longrightarrow Ca(NO_3)_2 + 2\,AgCl$

a) 3
b) **6**
c) 4
d) 2

15. Are the number of Hydrogen (H) atoms balanced in the following equation?

$K_4FeCN_6 + H_2SO_4 + H_2O \longrightarrow K_2SO_4 + FeSO_4 + (NH_4)_2SO_4 + CO$

a) No, there are 8 more on the reactant side
b) **No, there are 4 more on the product side**
c) No, there are 6 less on the product side
d) Yes

16. If you change the coefficient of H_2O on the product side to 2, will it balance the number of Hydrogen atoms?

$C_6H_5COOH + O_2 \longrightarrow CO_2 + H_2O$

a) No, it should be removed
b) **No, it should be changed to 3**
c) No, it should be changed to 4
d) Yes

17. What are the coefficients of the following equation when it's balanced?

$N_2O_4 \longrightarrow NO_2$

a) 2,3
b) **1,2**
c) 3,4
d) 2,1

18. In the reaction, $xCu + yHNO_3 \longrightarrow Cu(NO_3)_2 + 2\,NO_2 + 2\,H_2O$, the coefficients x and y are:

a) 2,3
b) **1,4**
c) 1,3
d) 3,8

19. Which of the following represents a balanced chemical equation?

a) **$\mathbf{2\,PO_4^{3-} + 3\,Ca^{2+} \longrightarrow Ca_3(PO_4)_2}$**
b) $2\,PO_4^{3-} + Ca^{2+} \longrightarrow 2\,Ca_3(PO_4)_2$
c) $PO_4^{3-} + {}_3Ca^{2+} \longrightarrow Ca_3(PO_4)_2$
d) $2\,PO_4^{3-} + Ca^{2+} \longrightarrow Ca_3(PO_4)_2$

20. For the reaction, $MnO_4^- + C_2O_4^{2-} + H^+ \longrightarrow Mn^{2+} + CO_2 + H_2O$, the correct coefficients of the reactants in the balanced reaction are:

a) **2,5,16**
b) 16,5,2
c) 5,16,2
d) 2,16,5

APPENDIX G

EXPERIMENT 2: POST-TEST FOR CHEMISTRY

1. Which of the following depicts a chemical equation?

   a) $x = y \mod a$ b) $x^2 + y^2 = z^2$

   c) $\mathbf{C + O_2 \longrightarrow CO_2}$ d) $F = m \times a$

2. Which of the following is a reactant in this equation?

   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) $\mathbf{CH_4}$ b) H_4

   c) H_2O d) H_2

3. Which of the following is a product in this equation?

   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) H_2O b) H_4

   c) $\mathbf{CH_4}$ d) H_2

4. What is the subscript of Hydrogen on the reactant side?

   $CH_4 + O_2 \longrightarrow CO_2 + H_2O$

   a) 2 b) **4**

   c) 1 d) 0

5. What is the coefficient of CO_2 on the product side?

   $CH_4 + O_2 \longrightarrow CO_2 + 3\,H_2O$

   a) 6 b) **2**

   c) 3 d) 4

6. The value of x is —— and the value of y + z + w is —— in the following equation:

   $xNaHCO_3 \longrightarrow yNa_2CO_3 + zCO_2 + wH_2O$

   a) 1,3 b) **2,3**

   c) 3,2 d) 3,3

7. The sum of all the coefficients of the reactants and the products when the following equation is balanced is:

   $C_7H_{16} + 11\,O_2 \longrightarrow 7\,CO_2 + 8\,H_2O$

   a) 28 b) **27**

   c) 15 d) 16

8. Is the following chemical equation balanced?

   $2\,Fe(NO_3)_3 + 3\,(NH_4)_2CO_3 \longrightarrow Fe_2(CO_3)_3 + NH_4NO_3$

a) **No, adding coefficient 6 to NH4NO3 will balance it**

b) No, adding coefficient 3 to NH4NO3 will balance it

c) No, adding coefficient 5 to NH4NO3 will balance it

d) Yes

9. Which of the following represents a balanced chemical equation?

a) $Al_2(CO_3)_3 + 2\,H_3PO_4 \longrightarrow 2\,AlPO_4 + CO_2 + 3\,H_2O$

b) $Al_2(CO_3)_3 + 2\,H_3PO_4 \longrightarrow 2\,AlPO_4 + 3\,CO_2 + H_2O$

c) $Al_2(CO_3)_3 + H_3PO_4 \longrightarrow 2\,AlPO_4 + 3\,CO_2 + 3\,H_2O$

d) $\mathbf{Al_2(CO_3)_3 + 2\,H_3PO_4 \longrightarrow 2\,AlPO_4 + 3\,CO_2 + 3\,H_2O}$

10. Which of the following represents a balanced chemical equation?

a) $3\,Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 3\,H_2O + Ba_3(PO_4)_2$

b) $Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 6\,H_2O + Ba_3(PO_4)_2$

c) $3\,Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 6\,H_2O + 2\,Ba_3(PO_4)_2$

d) $\mathbf{3\,Ba(OH)_2 + 2\,H_3PO_4 \longrightarrow 6\,H_2O + Ba_3(PO_4)_2}$

11. If you change the coefficient of $Fe(C_2H_3O_2)$ on the product side to 6, what should be the coefficient of $HC_2H_3O_2$ on reactant side to balance the number of Hydrogen atoms?
$Fe + HC_2H_3O_2 \longrightarrow Fe(C_2H_3O_2)_3 + H_2$

a) 4 b) **14**

c) 5 d) 15

12. If you change the coefficient of CH_4 on the reactant side to 2, what should be the coefficient of CO_2 on product side to balance the number of carbon atoms in this equation?
$CH_4 + O_2 \longrightarrow CO_2 + H_2O$

a) 1 b) **2**

c) 3 d) 4

13. What is the number of O atoms in the following equation on the products side?
$CH_4 + O_2 \longrightarrow 2\,CO_2 + 3\,H_2O$

a) 2 b) **7**

c) 3 d) 5

14. What is the number of O atoms in the following equation on the products side?
$CaCl_2 + 2\,AgNO_3 \longrightarrow Ca(NO_3)_2 + 2\,AgCl$

a) 3

b) **6**

c) 4

d) 2

15. Are the number of Oxygen (O) atoms balanced in the following equation?
$K_4FeCN_6 + H_2SO_4 + H_2O \longrightarrow K_2SO_4 + FeSO_4 + (NH_4)_2SO_4 + CO$

a) No, there are 8 more on the reactant side

b) **No, there are 4 more on the product side**

c) No, there are 6 less on the product side

d) Yes

16. If you change the coefficient of H on the product side to 3, will it balance the number of Hydrogen atoms?
$C_6H_5COOH + O_2 \longrightarrow CO_2 + H_2O$

a) No, it should be removed

b) No, it should be changed to 2

c) No, it should be changed to 4

d) **Yes**

17. What are the coefficients of the following equation when it's balanced?
$KCl \longrightarrow K + Cl_2$

a) 2,2,3

b) **2,2,1**

c) 1,3,4

d) 1,2,1

18. In the reaction, $Cu + 4\,HNO_3 \longrightarrow Cu(NO_3)_2 + xNO_2 + yH_2O$, the coefficients x and y are:

a) 2,3

b) **2,2**

c) 1,3

d) 3,8

19. Which of the following represents a balanced chemical equation?

a) $\mathbf{2\,PO_4^{3-} + 3\,Ca^{2+} \longrightarrow Ca_3(PO_4)_2}$

b) $2\,PO_4^{3-} + Ca^{2+} \longrightarrow 2\,Ca_3(PO_4)_2$

c) $PO_4^{3-} + {}_3Ca^{2+} \longrightarrow Ca_3(PO_4)_2$

d) $2\,PO_4^{3-} + Ca^{2+} \longrightarrow Ca_3(PO_4)_2$

20. For the reaction, $CuCl_2 + K^+ + PO_4^{3-} \longrightarrow KCl + Cu_3(PO_4)_2$, the correct coefficients of the reactants in the balanced reaction are:

a) **3,6,2**

b) 3,2,6

c) 2,3,6

d) 6,2,3

APPENDIX H

EXPERIMENT 2: PRE-TEST FOR CRYPTOGRAPHY

1. What will be the value of A after a shift of 1?

   a) A b) **B**

   c) C d) D

2. What will be the value of D after a shift of 2?

   a) E b) **F**

   c) B d) C

3. What will be the value of K after a shift of 25?

   a) L b) **J**

   c) M d) I

4. What will be the value of B after a shift of 24?

   a) Y b) **Z**

   c) A d) D

5. What will be the value of K after a shift of 26?

   a) J b) **K**

   c) L d) I

6. What will be the value of P after a shift of 0 (zero)?

   a) O b) **P**

   c) Q d) R

7. Which of the following could be a possible encryption of the phrase 'AB'?

   a) YX b) **CD**

   c) KM d) AZ

8. Which of the following could be a possible decryption of the word 'EBE'?

   a) BAD b) **DAD**

   c) BOB d) ROR

9. Which of the following represents a encryption of the word 'MSG' using the key 2?

   a) YES b) **OUI**

   c) NTH d) KQE

10. Which of the following represents a decryption of the word 'MSG' using the key 2?

a) YES b) **KQE**

c) OUI d) NTH

11. Encrypting the word 'PEACE' using the key 1 will lead to the cipher text ——

a) QFCEF b) **QFBDF**

c) ODZBD d) ODACD

12. Encrypting the phrase 'COME AT ONCE' using 2 as encryption key will lead to the cipher text ——.

a) AMKC YR MLAC b) **EQOG CV QPEG**

c) EQOG CC QQEG d) ADDC YR MLBC

13. Which of the following CAN NOT be the possible decryption for the cipher text 'DSP' when decrypted using a given key?

a) FUR b) **WHY**

c) LAX d) SHE

14. Which of the following CAN NOT be a possible decryption of the phrase 'ALIIP'?

a) DOLLS b) **ZKHHP**

c) WHEEL d) XIFFM

15. Which of the following could be a possible decryption of the phrase 'LIFE IS BEAUTIFUL'?

a) KYEZ HM AXZUSHFTL b) **MJGF JT CFBVUJGVM**

c) NZIG LU DGCWWKHWO d) MJHF KT CFDVUGZZ

16. If you encrypted G using the key 1 and got the cipher text as I, then what went wrong?

a) used the key as 4 instead of 1

b) **used the key as 2 instead of 1**

c) used the key as 25 instead of 1

d) used the key as 24 instead of 1

17. If you encrypted M using the key 25 and got the cipher text as N, then what went wrong?

a) used the key as 26 instead of 25

b) **decrypted instead of encrypting**

c) used the key as 4 instead of 25

d) used the key as 10 instead of 25

18. If you encrypted 'BOB' using the key 2 and got the cipher text as 'DAD', then what went wrong?

    a) used the key as 1 instead of 2

    b) **wrongly encrypted O to A, instead of Q**

    c) used the key as 25 instead of 1

    d) wrongly encrypted B to D, instead of Z

19. If you encrypted 'APPLE' using the key 1 and got the cipher text as 'BQQMG', then what went wrong?

    a) used the key as 2 instead of 1

    b) **wrongly encrypted E to G, instead of F**

    c) used the key as 25 instead of 1

    d) wrongly encrypted L to M, instead of K

20. If you encrypted 'MARLEY' using the key 25 and got the cipher text as 'NBSMFZ', then what went wrong?

    a) used the key as 26 instead of 25

    b) **used the key as 1 instead of 25**

    c) wrongly encrypted Y to Z, instead of X

    d) wrongly encrypted M to N, instead of L

APPENDIX I

EXPERIMENT 2: POST-TEST FOR CRYPTOGRAPHY

1. What will be the value of Z after a shift of 1?
   a) B b) **A**
   c) C d) D
2. What will be the value of E after a shift of 2?
   a) E b) **G**
   c) B d) C
3. What will be the value of L after a shift of 25?
   a) L b) **K**
   c) M d) J
4. What will be the value of C after a shift of 23?
   a) Y b) **Z**
   c) A d) D
5. What will be the value of I after a shift of 26?
   a) J b) **I**
   c) L d) K
6. What will be the value of O after a shift of 0 (zero)?
   a) P b) **O**
   c) Q d) R
7. Which of the following could be a possible encryption of the phrase 'EF'?
   a) YX b) **CD**
   c) KM d) AZ
8. Which of the following could be a possible decryption of the word 'DAD'?
   a) BAD b) **FCF**
   c) BOB d) ROR
9. Which of the following represents a encryption of the word 'MSG' using the key 3?
   a) YES b) **PVJ**
   c) NTH d) JPD
10. Which of the following represents a decryption of the word 'MSG' using the key 3?

a) YES b) **JPD**

c) PVJ d) KQE

11. Encrypting the word 'PEACE' using the key 2 will lead to the cipher text ——.

    a) RFCEG b) **RGCEG**

    c) RGDEG d) RGCFG

12. Encrypting the phrase 'COME AT ONCE' using 1 as encryption key will lead to the cipher text ——.

    a) EQOG CC QQEG b) **DPNF BU PODF**

    c) EQOG CC QQEG d) DPNF BU PODE

13. Which of the following CAN NOT be the possible decryption for the cipher text 'DSP' when decrypted using a given key?

    a) FUR b) **WHY**

    c) LAX d) SHE

14. Which of the following CAN NOT be a possible decryption of the phrase 'ALIIP'?

    a) DOLLS b) **ZKHHP**

    c) WHEEL d) XIFFM

15. Which of the following could be a possible decryption of the phrase 'LIFE IS BEAUTIFUL'?

    a) KHED HR ADZTSHETL b) **KHED HR ADZTSHETK**

    c) MJGF JT CFBVUJGVN d) NKHG KU DGCWVKHWM

16. If you encrypted G using the key 1 and got the cipher text as J, then what went wrong?

    a) used the key as 4 instead of 1

    b) **used the key as 3 instead of 1**

    c) used the key as 25 instead of 1

    d) decrypted instead of encrypting

17. If you encrypted M using the key 25 and got the cipher text as N, then what went wrong?

    a) used the key as 26 instead of 25

    b) **decrypted instead of encrypting**

    c) used the key as 4 instead of 25

    d) used the key as 10 instead of 25

18. If you encrypted 'BOB' using the key 2 and got the cipher text as 'DAD', then what went wrong?

    a) used the key as 1 instead of 2

    b) **wrongly encrypted O to A, instead of Q**

    c) used the key as 25 instead of 1

    d) wrongly encrypted O to A, instead of Q

19. If you encrypted 'APPLE' using the key 1 and got the cipher text as 'BQQNF', then what went wrong?

    a) used the key as 2 instead of 1

    b) **wrongly encrypted L to N, instead of M**

    c) used the key as 25 instead of 1

    d) wrongly encrypted E to F, instead of G

20. If you encrypted 'MARLEY' using the key 25 and got the cipher text as 'LZQKDX', then what went wrong?

    a) used the key as 26 instead of 25

    b) **used the key as 1 instead of 25**

    c) wrongly encrypted Y to Z, instead of X

    d) wrongly encrypted M to N, instead of L

APPENDIX J

PROMPTS USED TO INFORM GAME MECHANICS

Following is an exhaustive list of prompts that were used in the game to make the player aware of the game mechanics. The list follows the order in which they were encountered or learned in the game.

1. Move around using the arrow keys.
2. Collect 100 coins for a 1-up.
3. Use up and down arrow keys to climb the ladder!
4. Use space-bar to jump and avoid the enemies.
5. Avoid the Spikes!
6. Some platforms start moving when you try to reach them!
7. Jumping across three blocks is hard, I hope you can make it!
8. Avoid the water, you are hydrophobic!
9. Avoid flying bats.
10. Pick hearts for a 1-up.
11. Finish the level.
12. Hey! You need to balance the chemical equation shown above by collecting the exact number of molecules required that participate in the chemical reaction.
13. Hey! You need to encode/decode text shown above by collecting the letters that appear in the transformed text.
14. Ouch! I don't need anymore molecules of this type.
15. Ouch! This letter does not exist in the resultant text.
16. Ouch! You do not need anymore of this letter.
17. Looks like you have collected everything. Let's "GO" ahead!

APPENDIX K

CHEM-O-CRYPT CHUNK LAYOUTS

Figure 18. Chunk 1 layout for difficulty setting 1

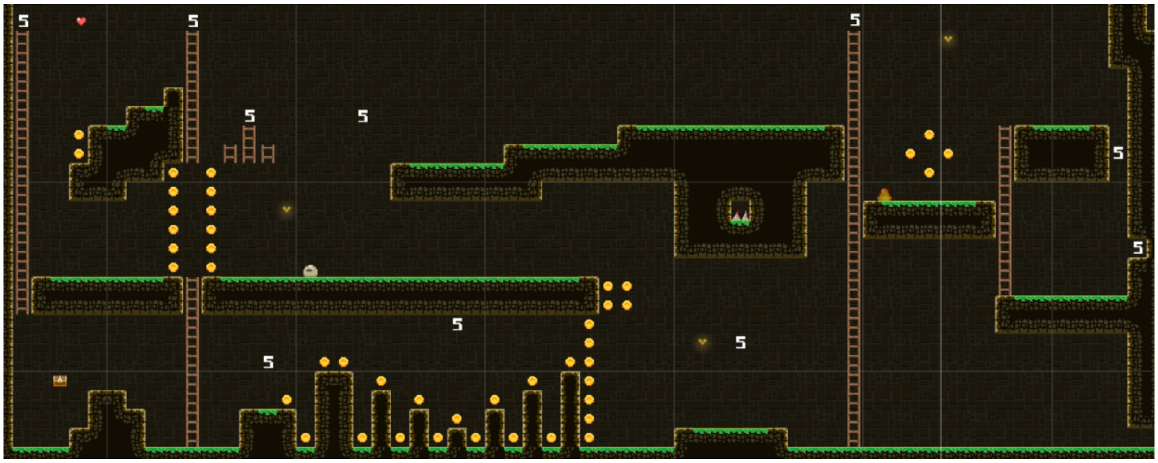


Figure 19. Chunk 1 layout for difficulty setting 2

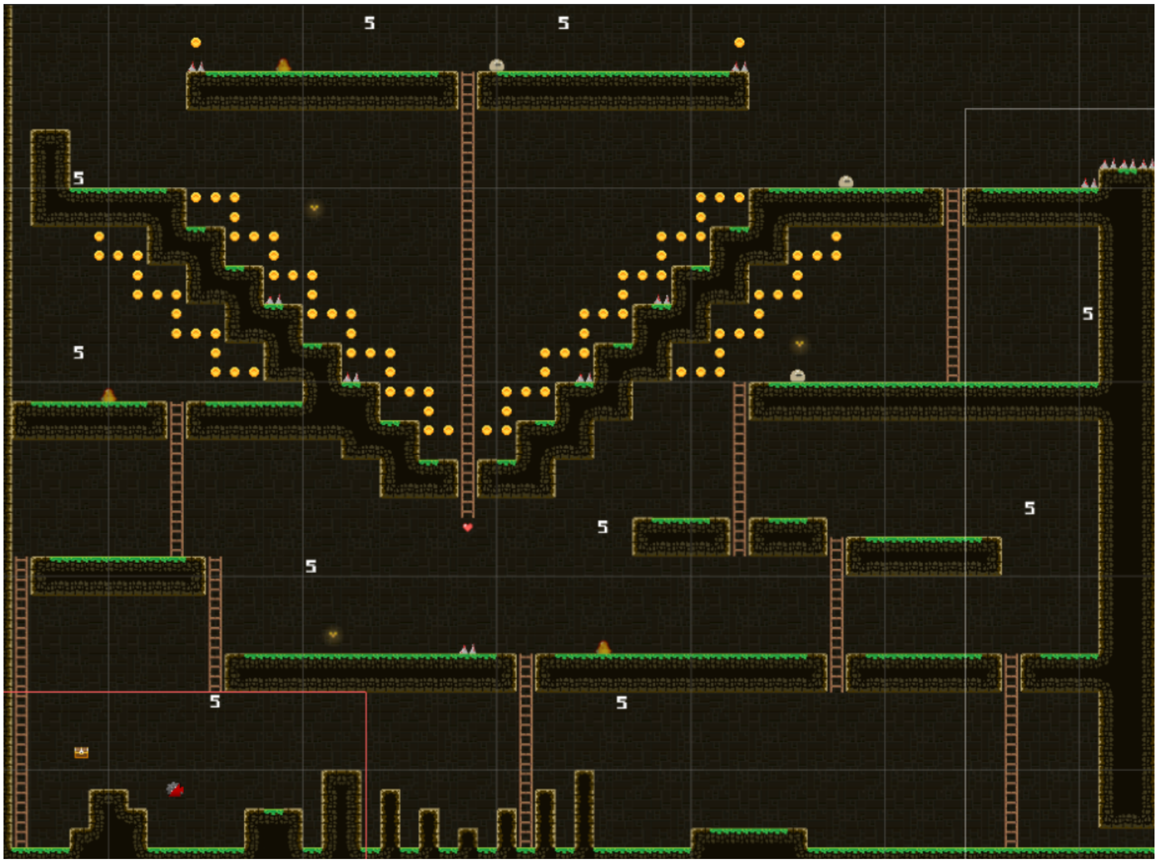


Figure 20. Chunk 1 layout for difficulty setting 3

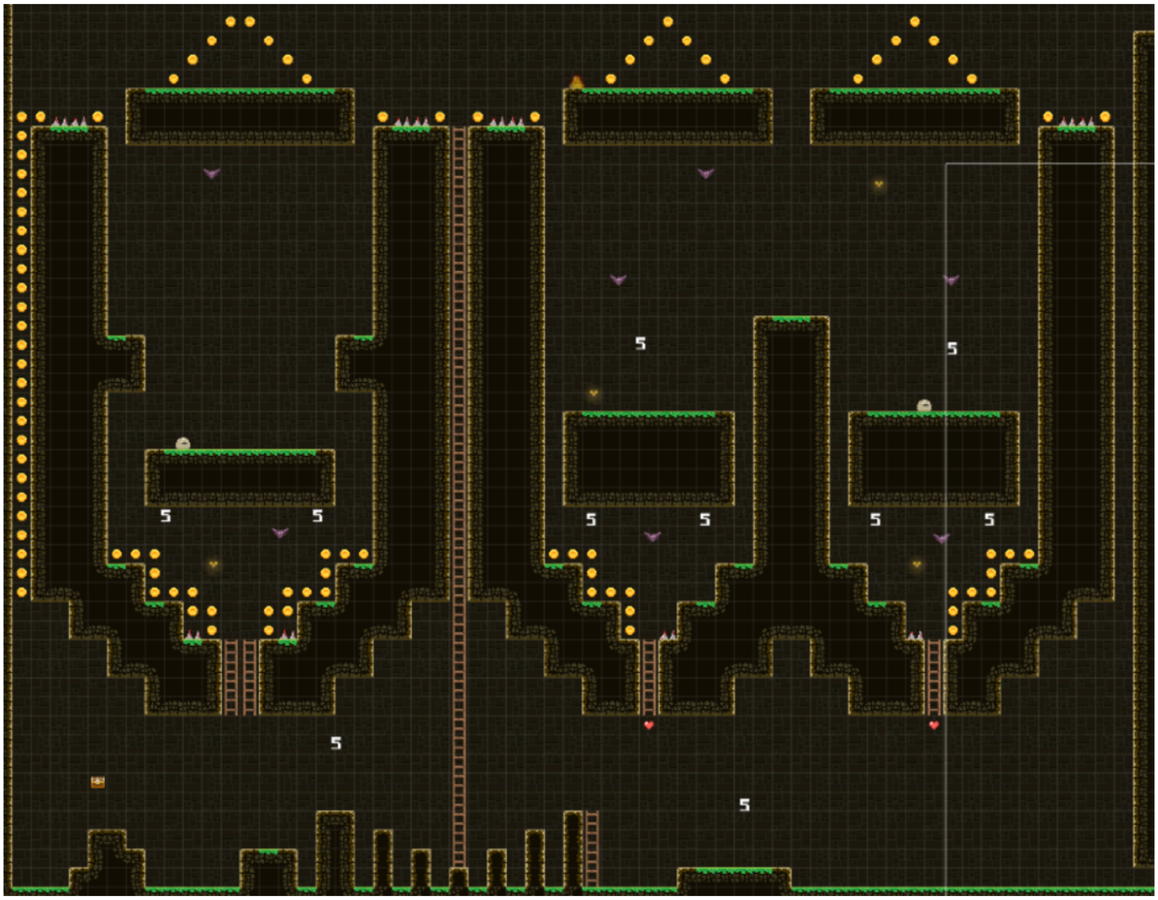


Figure 21. Chunk 1 layout for difficulty setting 4

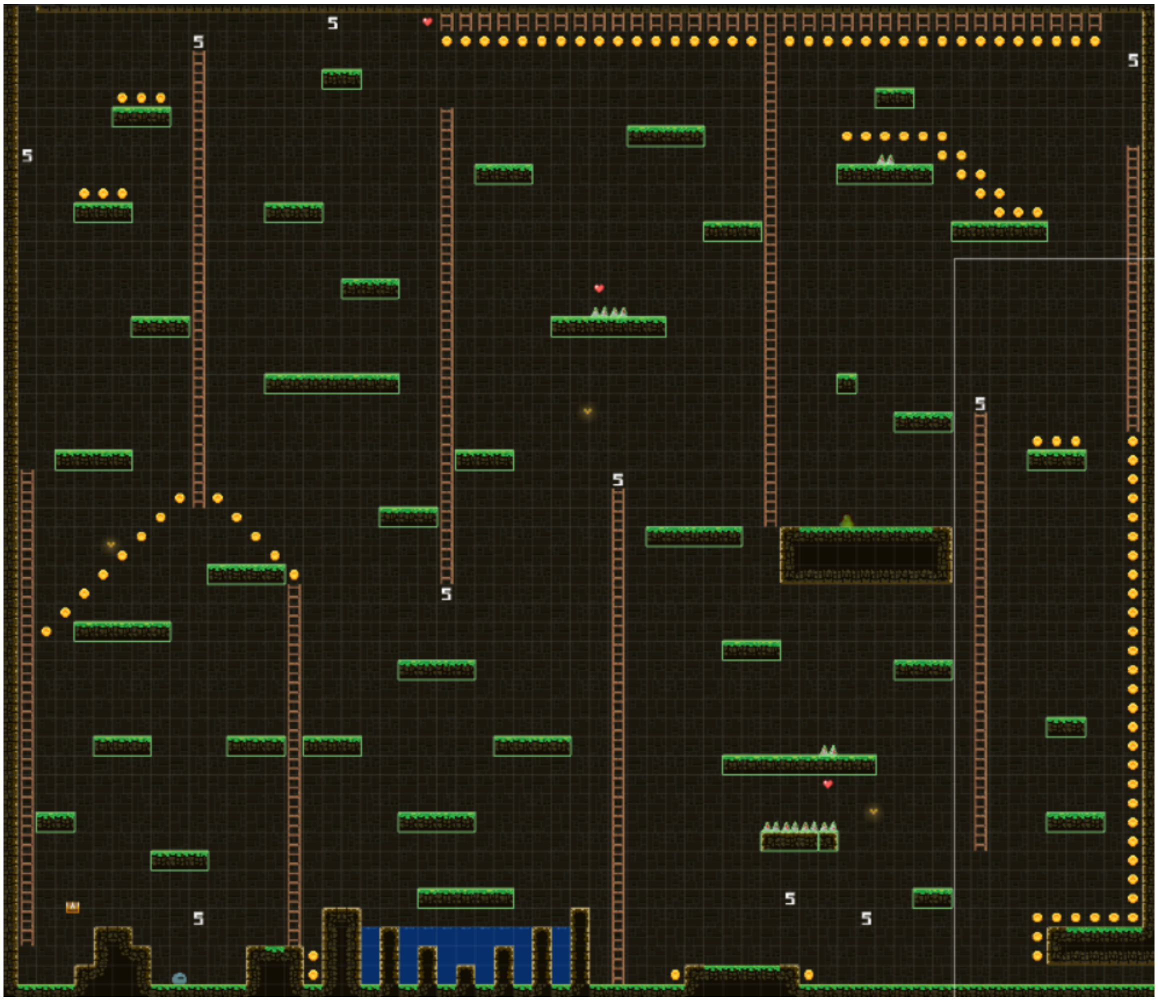


Figure 22. Chunk 2 layout for difficulty setting 1

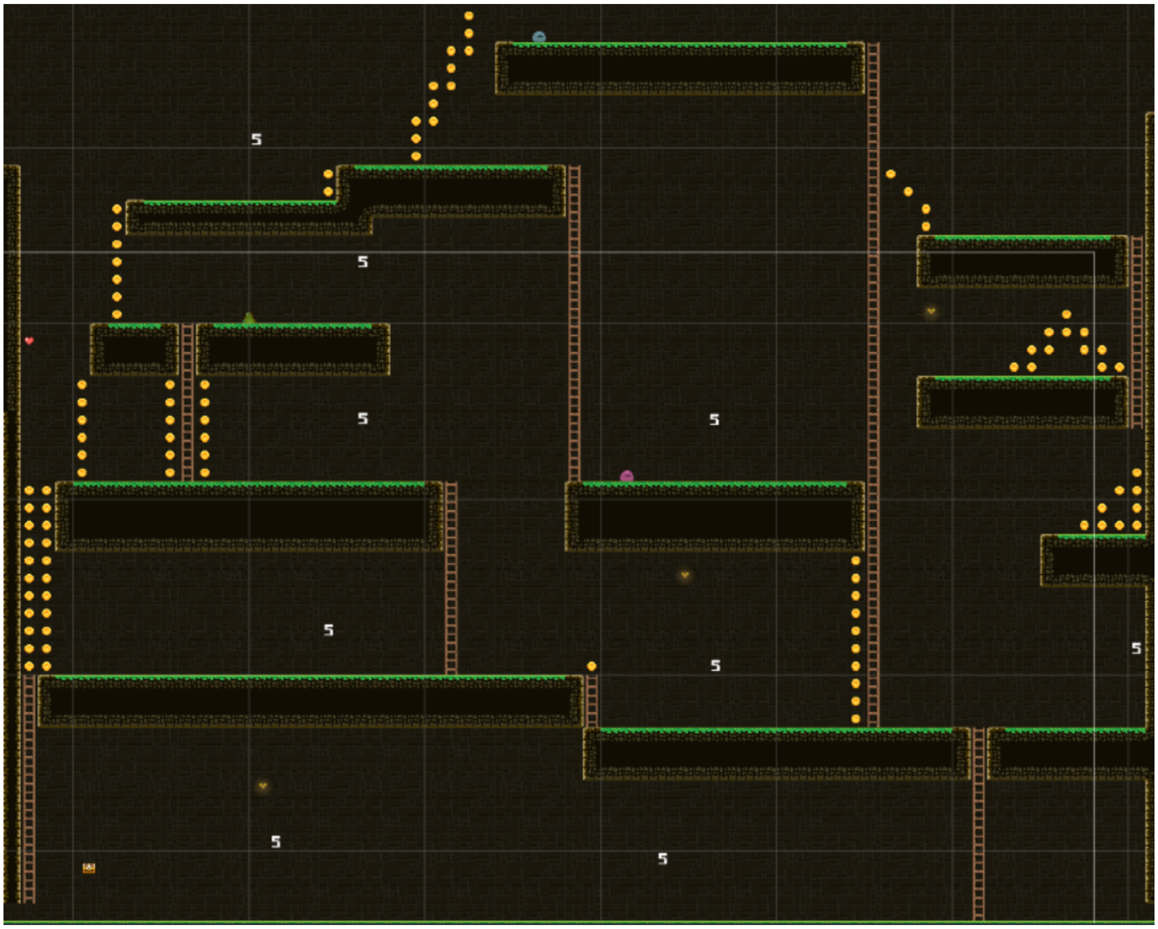


Figure 23. Chunk 2 layout for difficulty setting 2

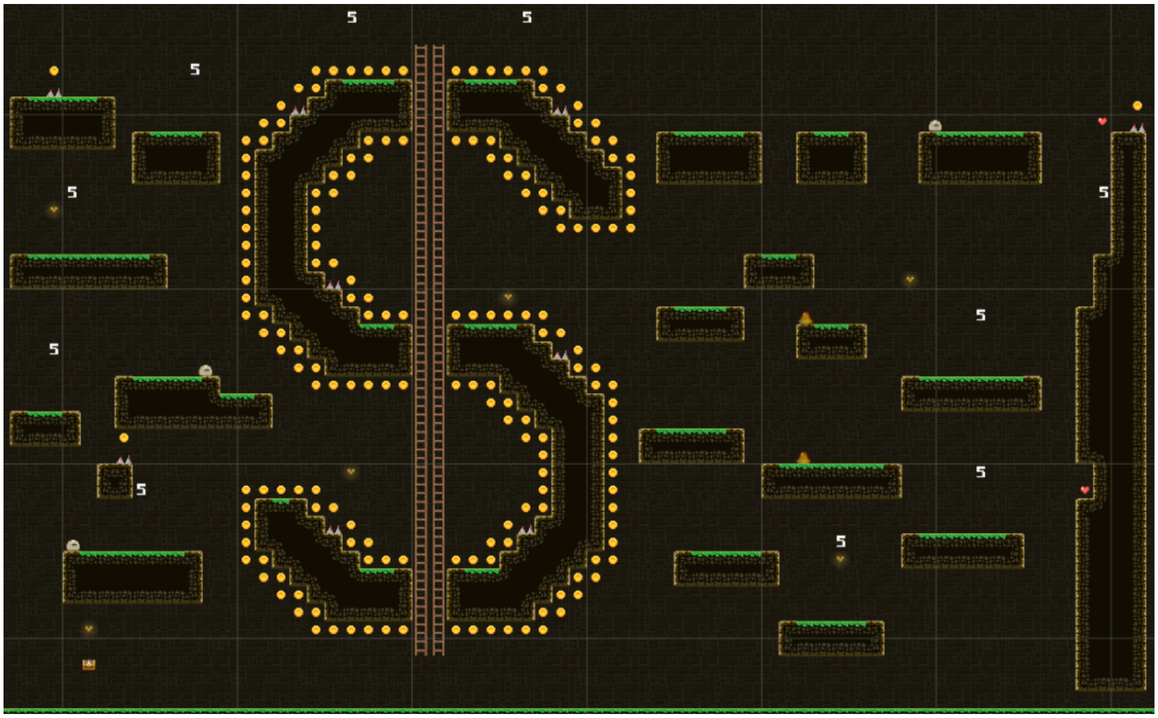


Figure 24. Chunk 2 layout for difficulty setting 3

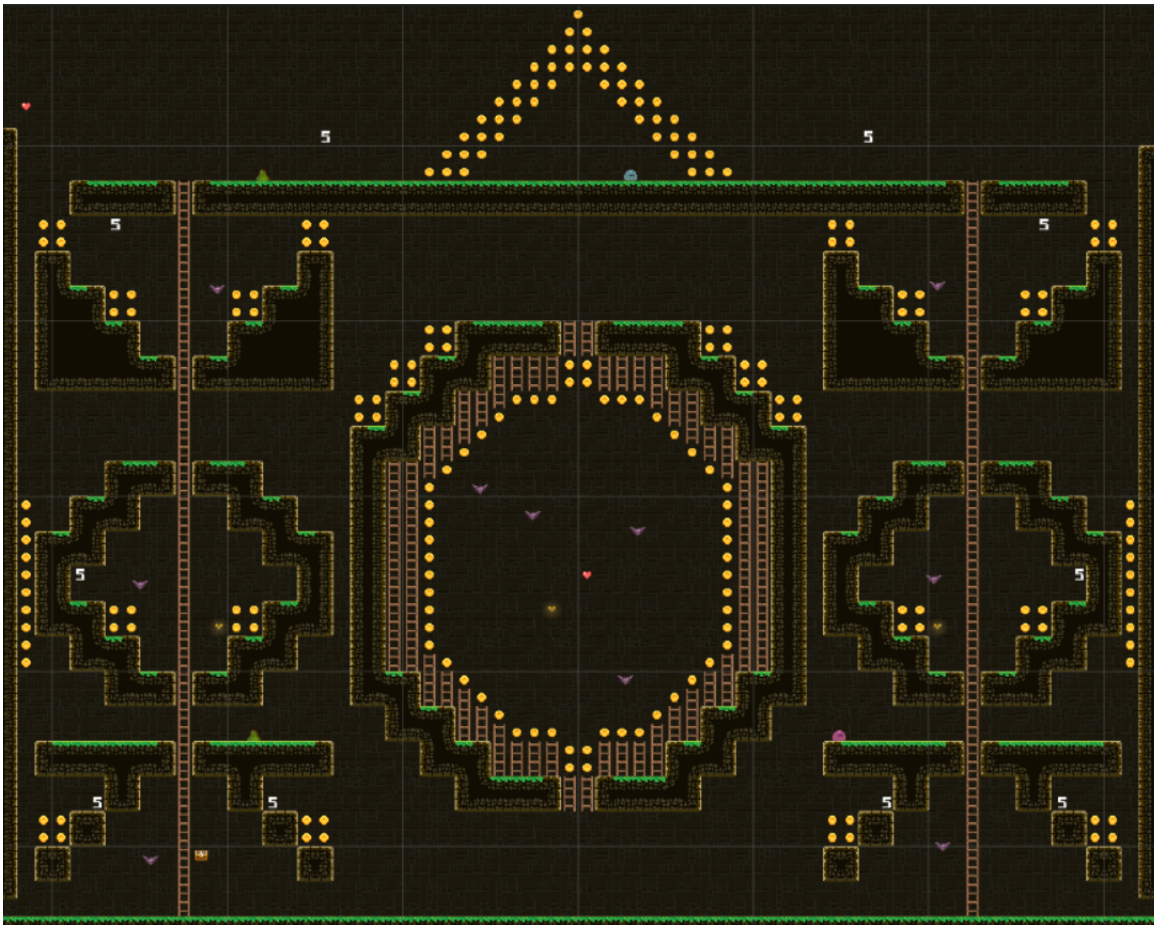


Figure 25. Chunk 2 layout for difficulty setting 4

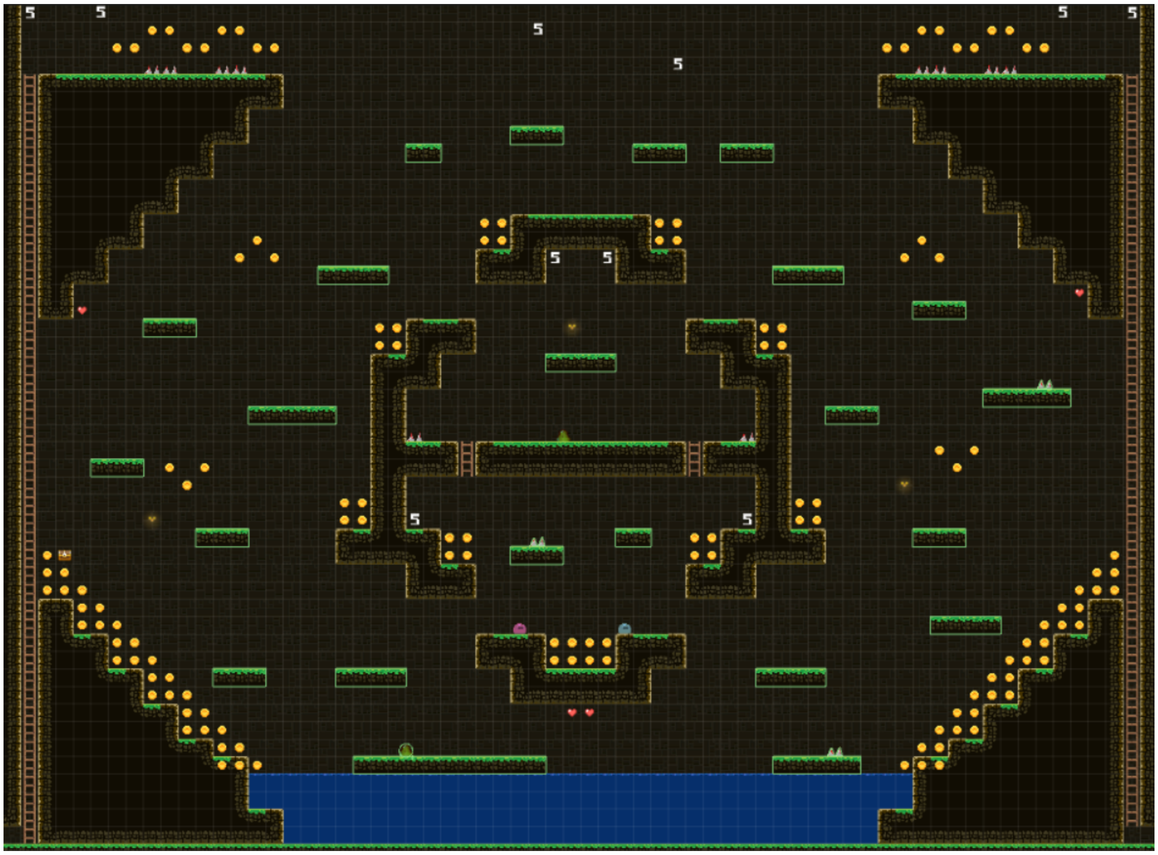


Figure 26. Chunk 3 layout for difficulty setting 1

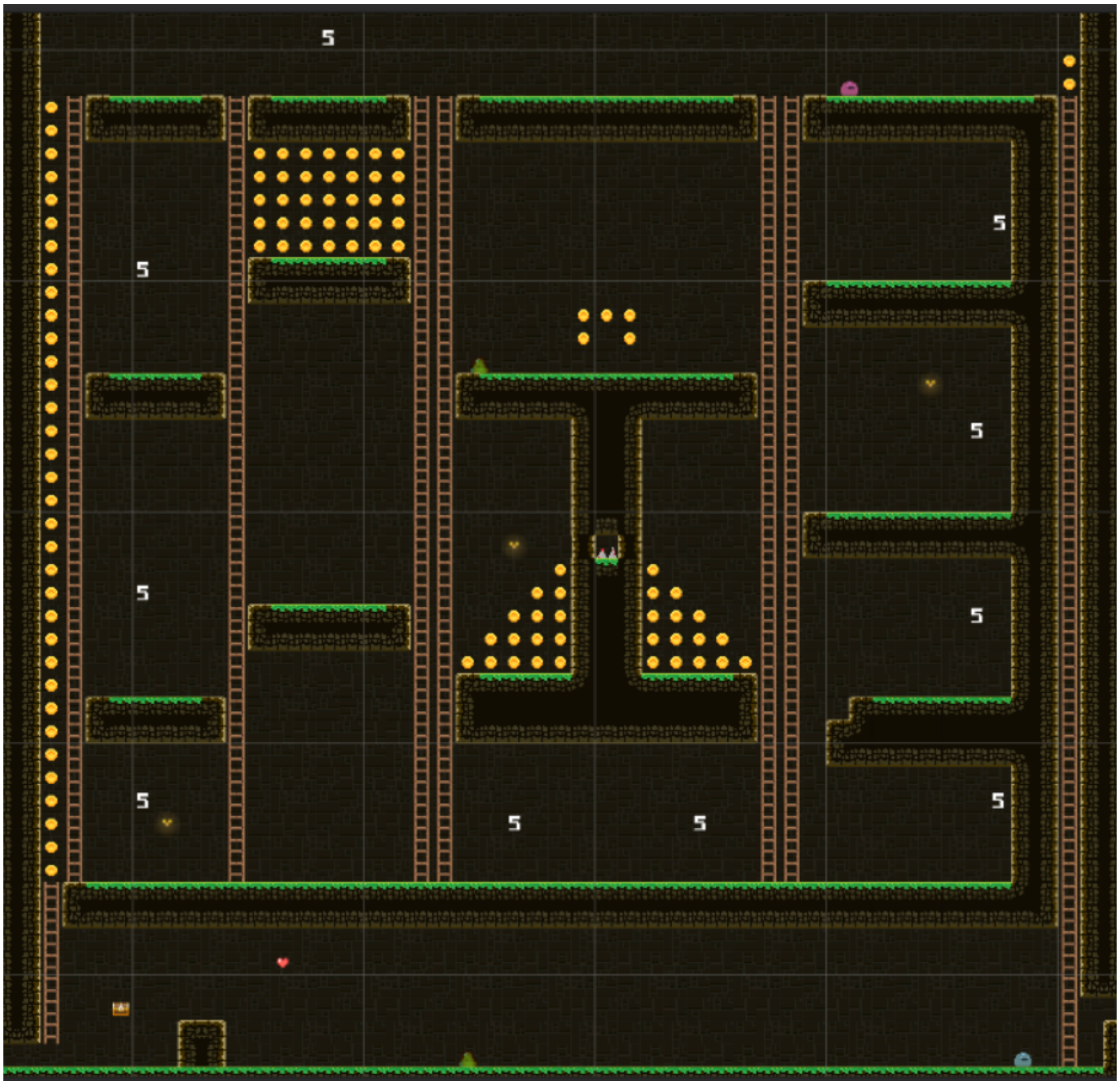


Figure 27. Chunk 3 layout for difficulty setting 2

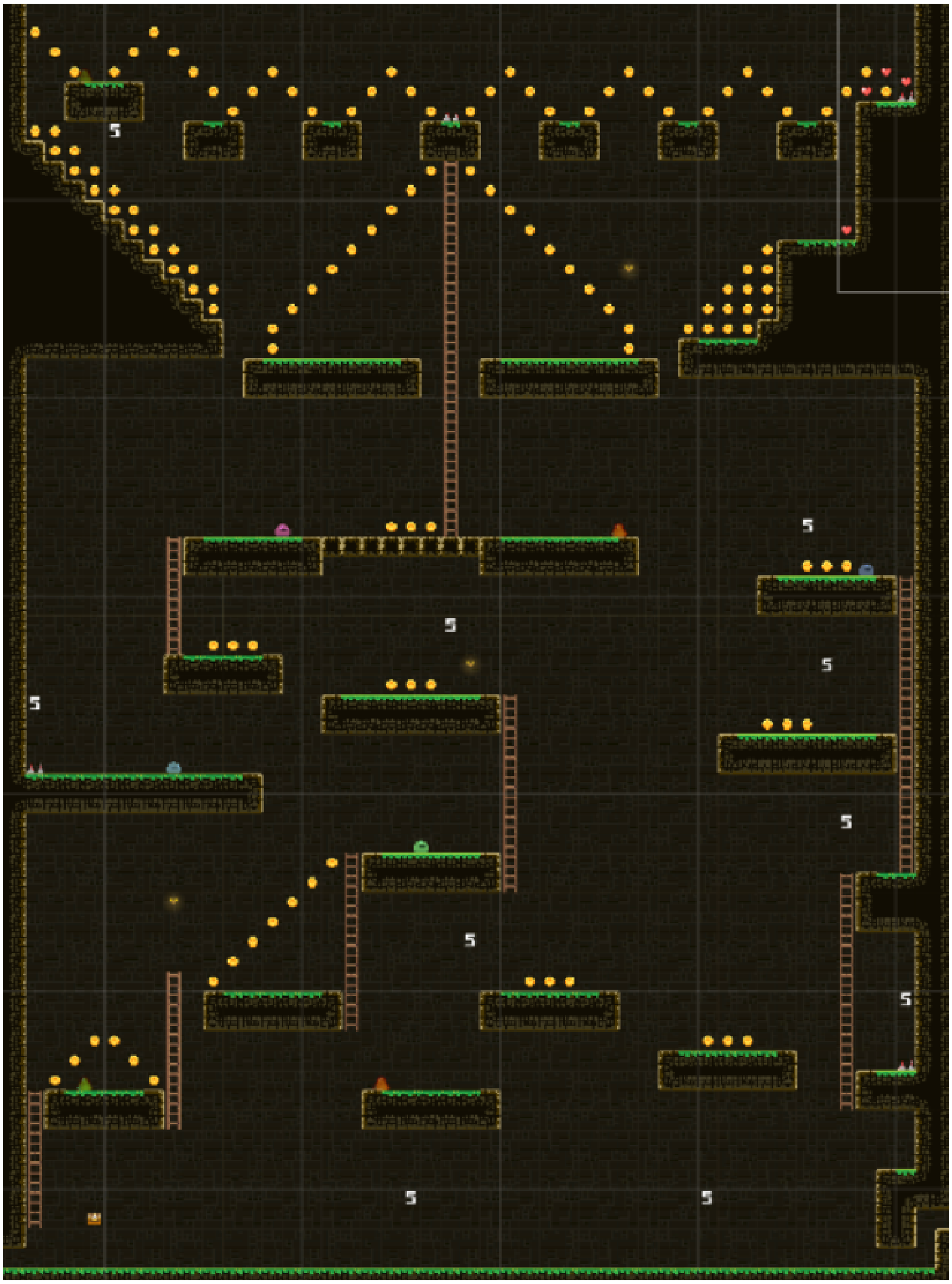


Figure 28. Chunk 3 layout for difficulty setting 3

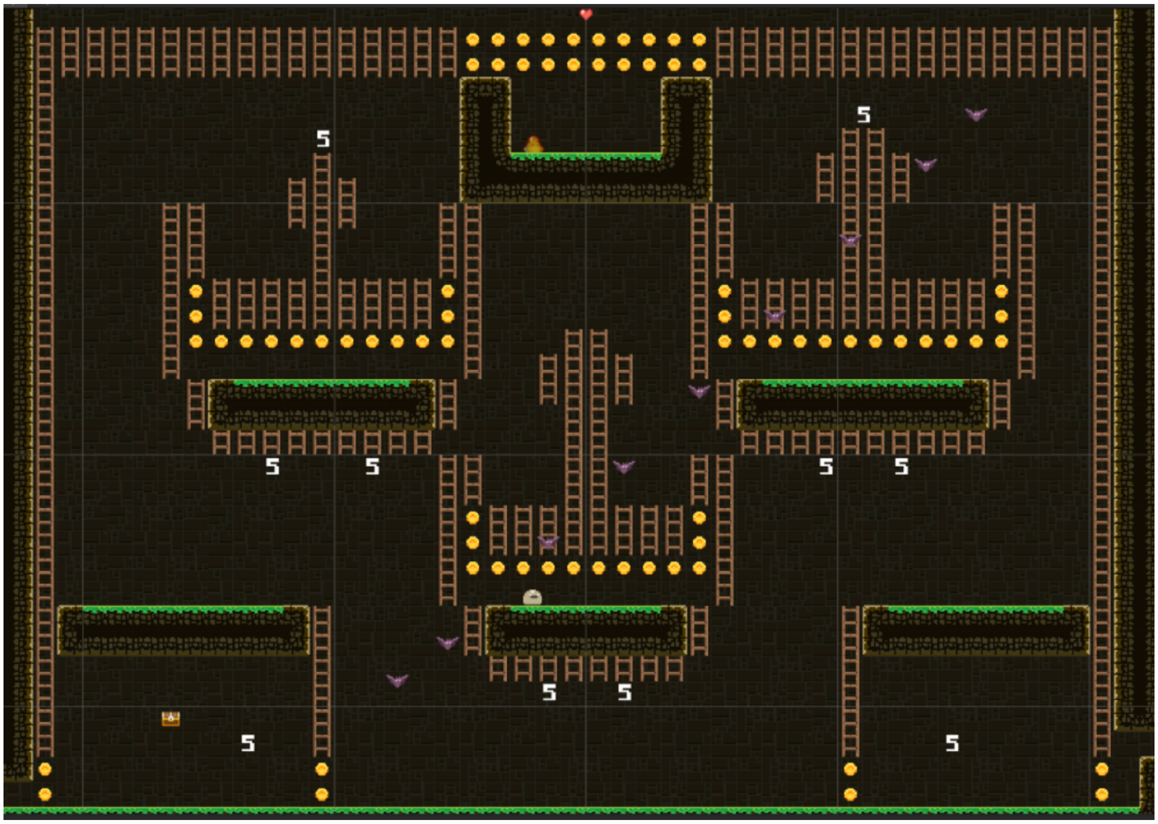


Figure 29. Chunk 3 layout for difficulty setting 4

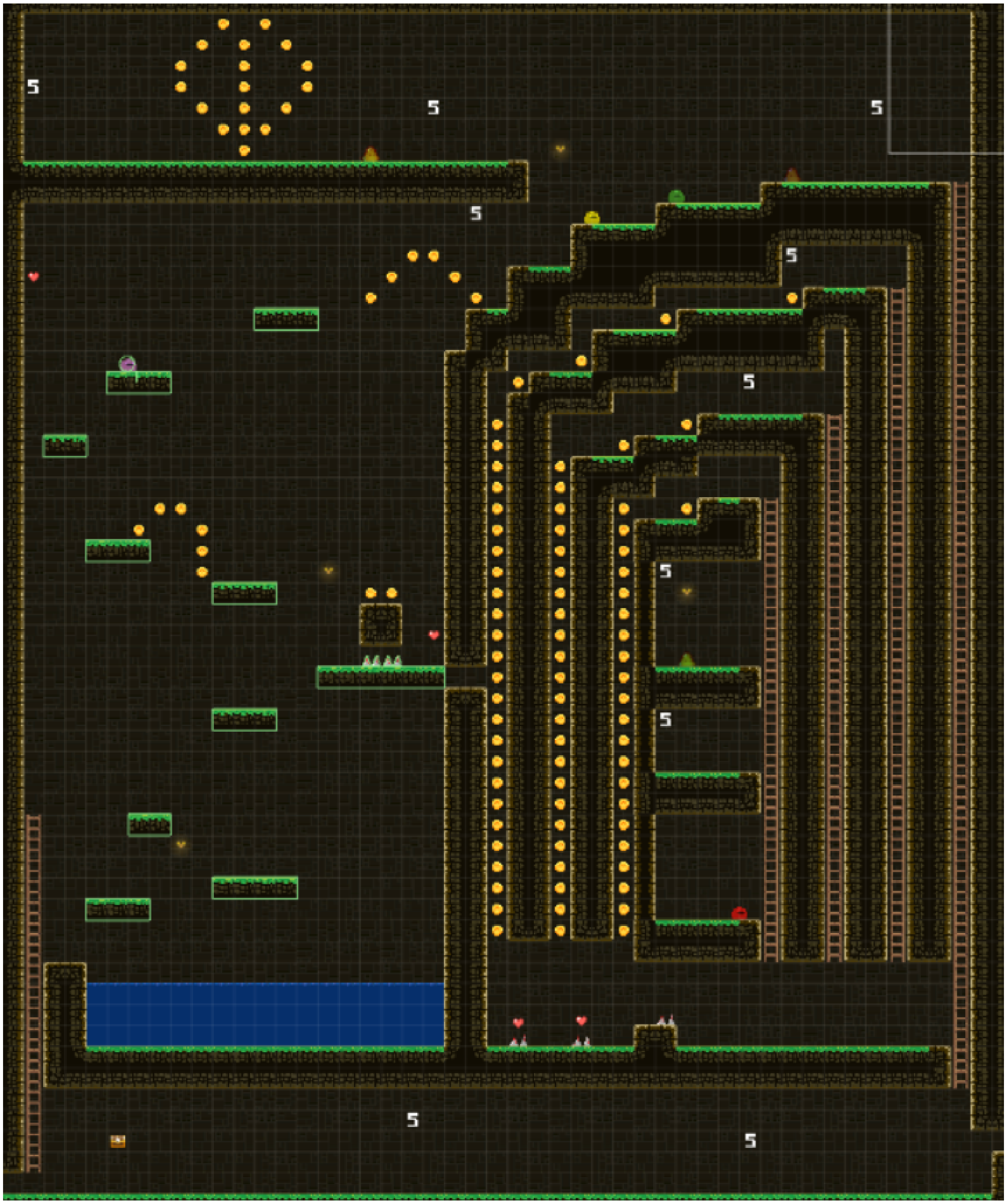


Figure 30. Chunk 4 layout for difficulty setting 1

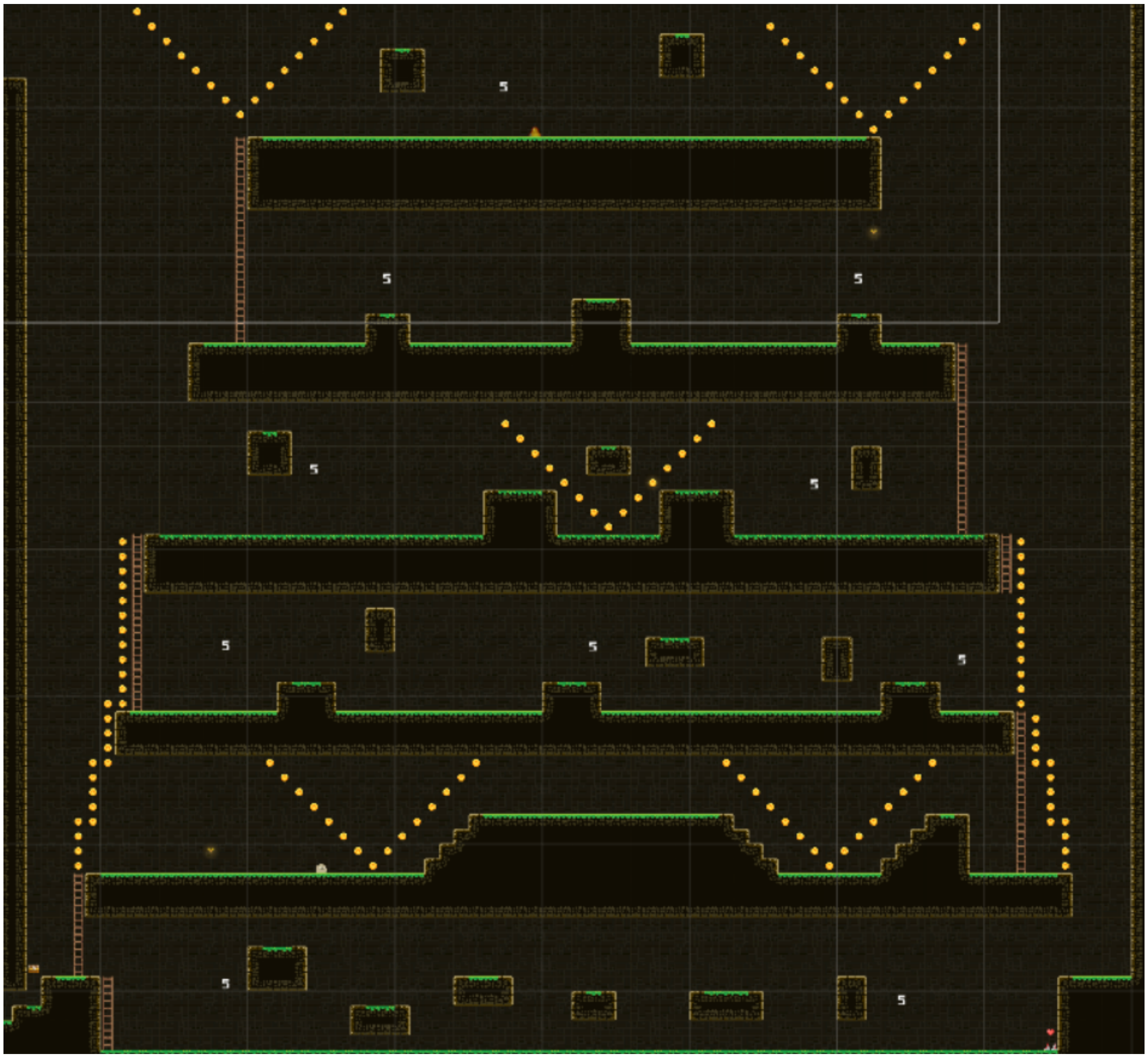


Figure 31. Chunk 4 layout for difficulty setting 2

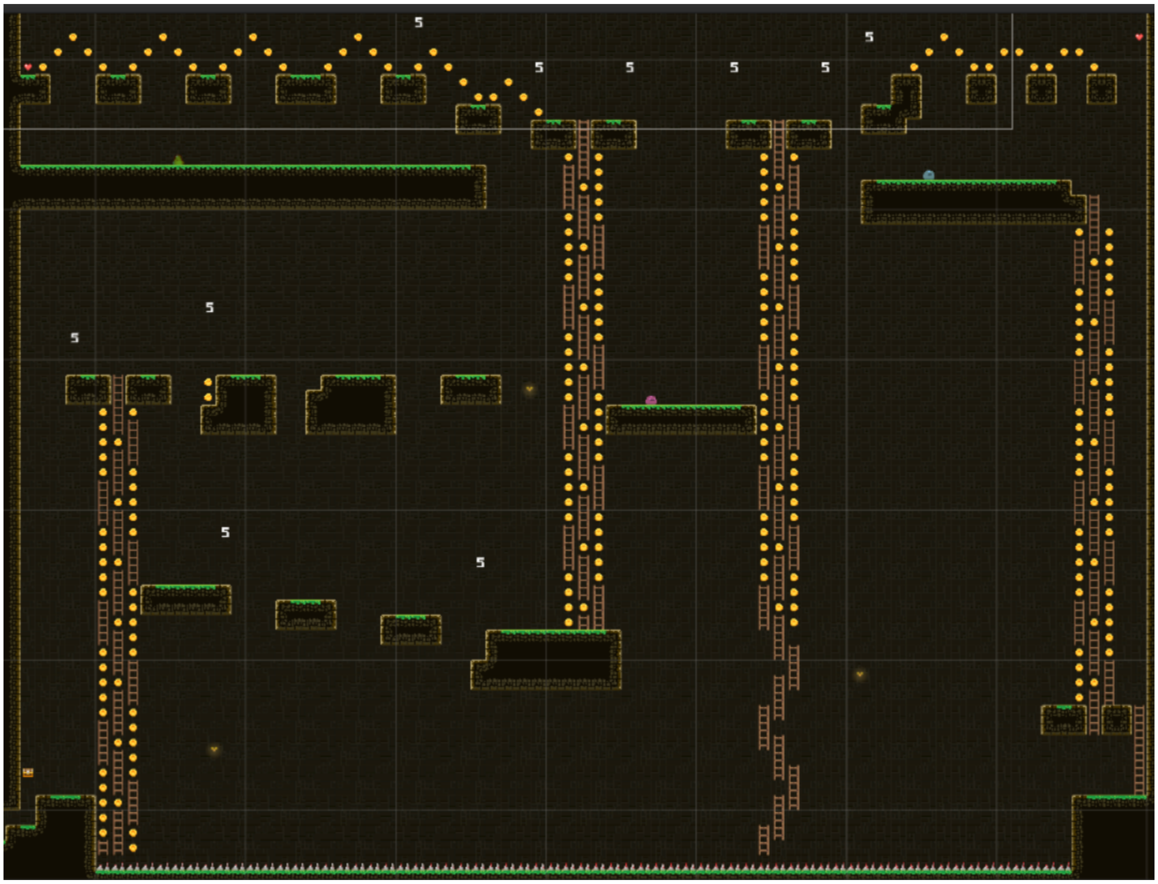


Figure 32. Chunk 4 layout for difficulty setting 3

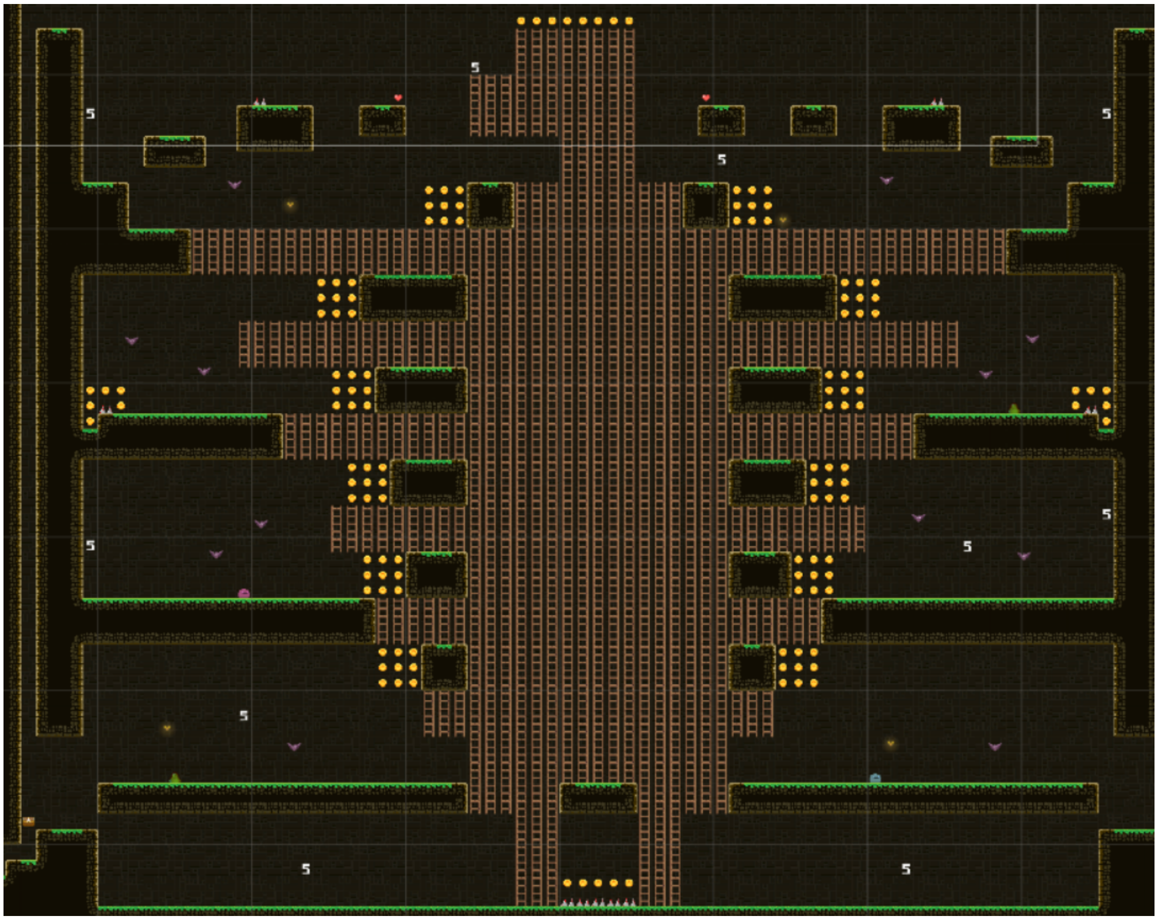
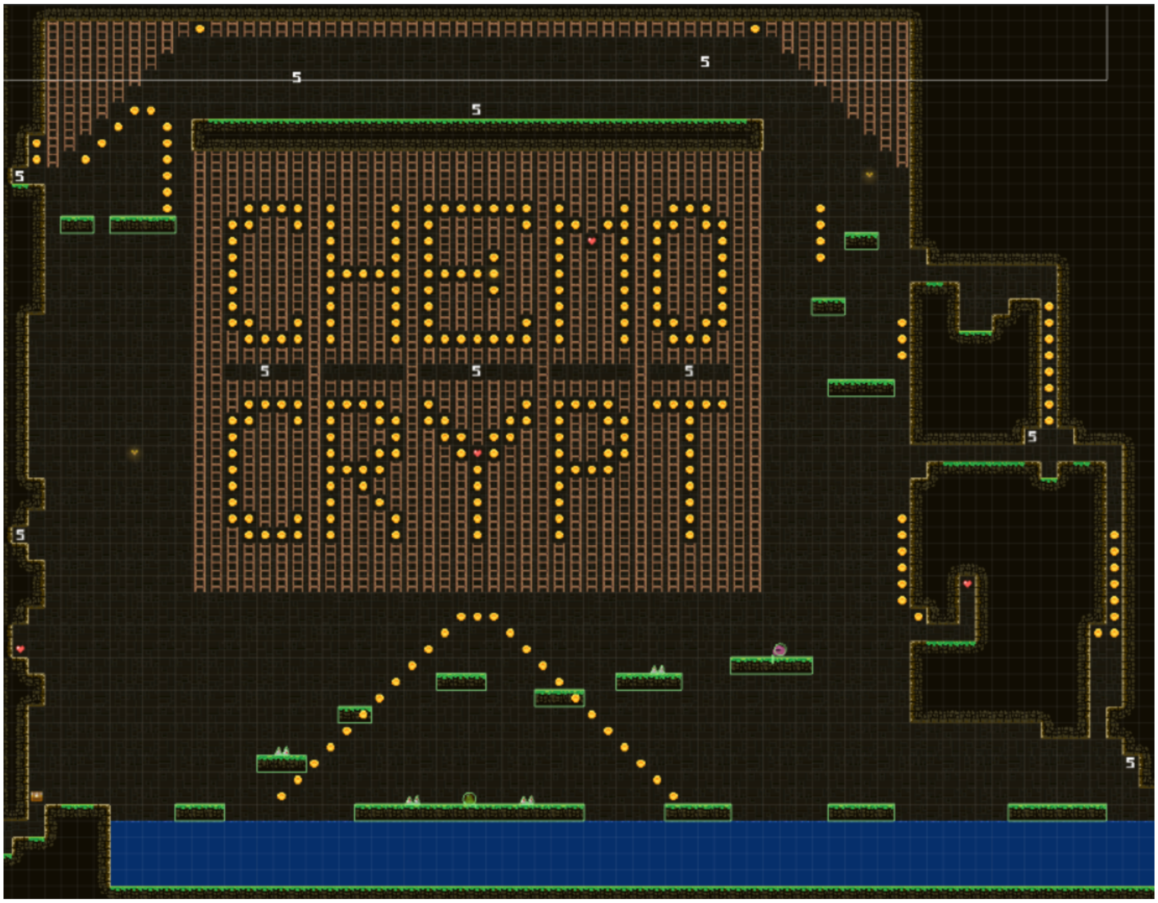
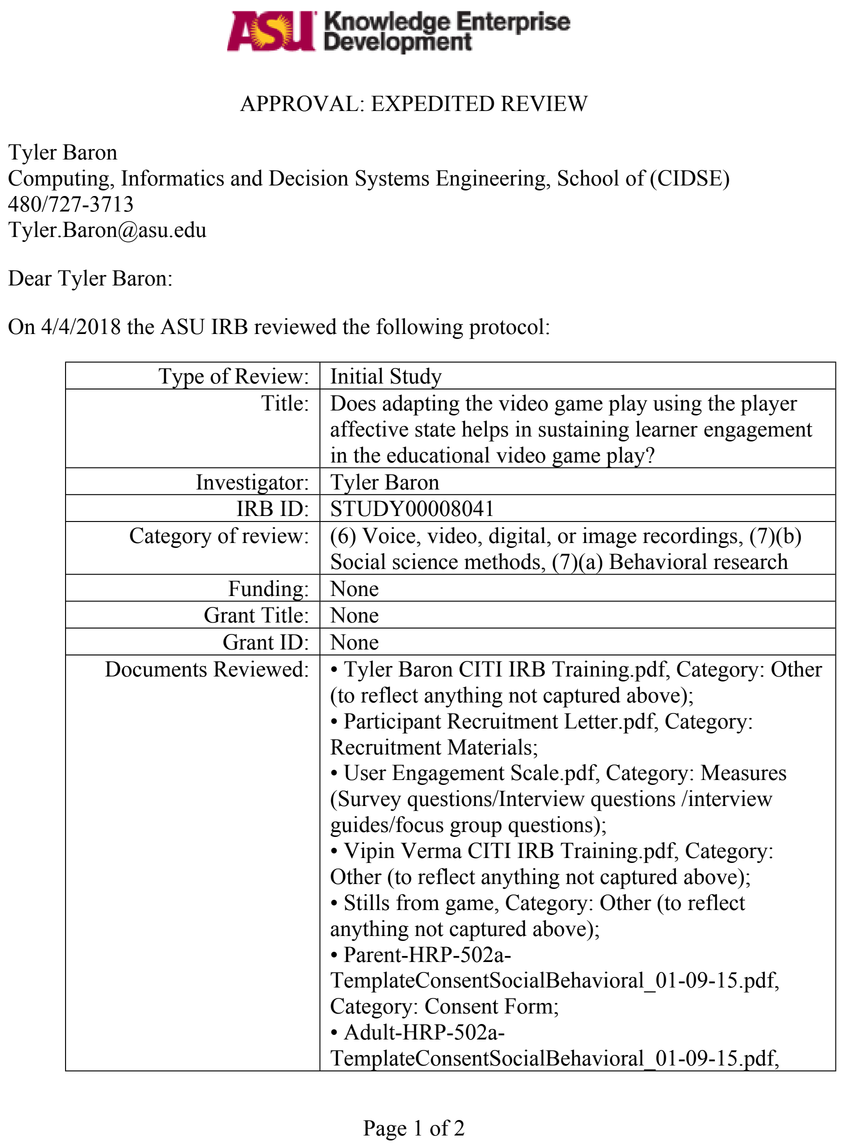


Figure 33. Chunk 4 layout for difficulty setting 4



APPENDIX L

IRB PERMISSION FOR HUMAN SUBJECT TESTING



APPROVAL: EXPEDITED REVIEW

Tyler Baron
Computing, Informatics and Decision Systems Engineering, School of (CIDSE)
480/727-3713
Tyler.Baron@asu.edu

Dear Tyler Baron:

On 4/4/2018 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Does adapting the video game play using the player affective state helps in sustaining learner engagement in the educational video game play?
Investigator:	Tyler Baron
IRB ID:	STUDY00008041
Category of review:	(6) Voice, video, digital, or image recordings, (7)(b) Social science methods, (7)(a) Behavioral research
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• Tyler Baron CITI IRB Training.pdf, Category: Other (to reflect anything not captured above); • Participant Recruitment Letter.pdf, Category: Recruitment Materials; • User Engagement Scale.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Vipin Verma CITI IRB Training.pdf, Category: Other (to reflect anything not captured above); • Stills from game, Category: Other (to reflect anything not captured above); • Parent-HRP-502a-TemplateConsentSocialBehavioral_01-09-15.pdf, Category: Consent Form; • Adult-HRP-502a-TemplateConsentSocialBehavioral_01-09-15.pdf,

Page 1 of 2

ASU Knowledge Enterprise Development

EXEMPTION GRANTED

Scotty Craig
IAFSE-PS: Human Systems Engineering (HSE)
480/727-1006
Scotty.Craig@asu.edu

Dear Scotty Craig:

On 11/7/2019 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Content Agnostic Game Based Stealth Assessment
Investigator:	Scotty Craig
IRB ID:	STUDY00010832
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• In-game reading material and pre/post test content, Category: Resource list; • User Engagement Scale.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • AjayBansal_IRB_CompletionCertificate7634258.pdf, Category: Other; • Vipin Verma CITI IRB Training.pdf, Category: Other; • gameScreenshots.pdf, Category: Technical materials/diagrams; • Adult-HRP-502a-TemplateConsentSocialBehavioral_01-09-15.pdf, Category: Consent Form; • HRP-503a-TEMPLATE_PROTOCOL_SocialBehavioralV02-10-15.docx, Category: IRB Protocol; • Recruitment Posting on SONA, Category: Recruitment materials/advertisements /verbal

APPENDIX M

PLOTS CORRESPONDING TO THE ITEM RESPONSE ANALYSIS FOR CHEMISTRY AND CRYPTOGRAPHY TESTS USED IN EXPERIMENTS 1 AND 2.

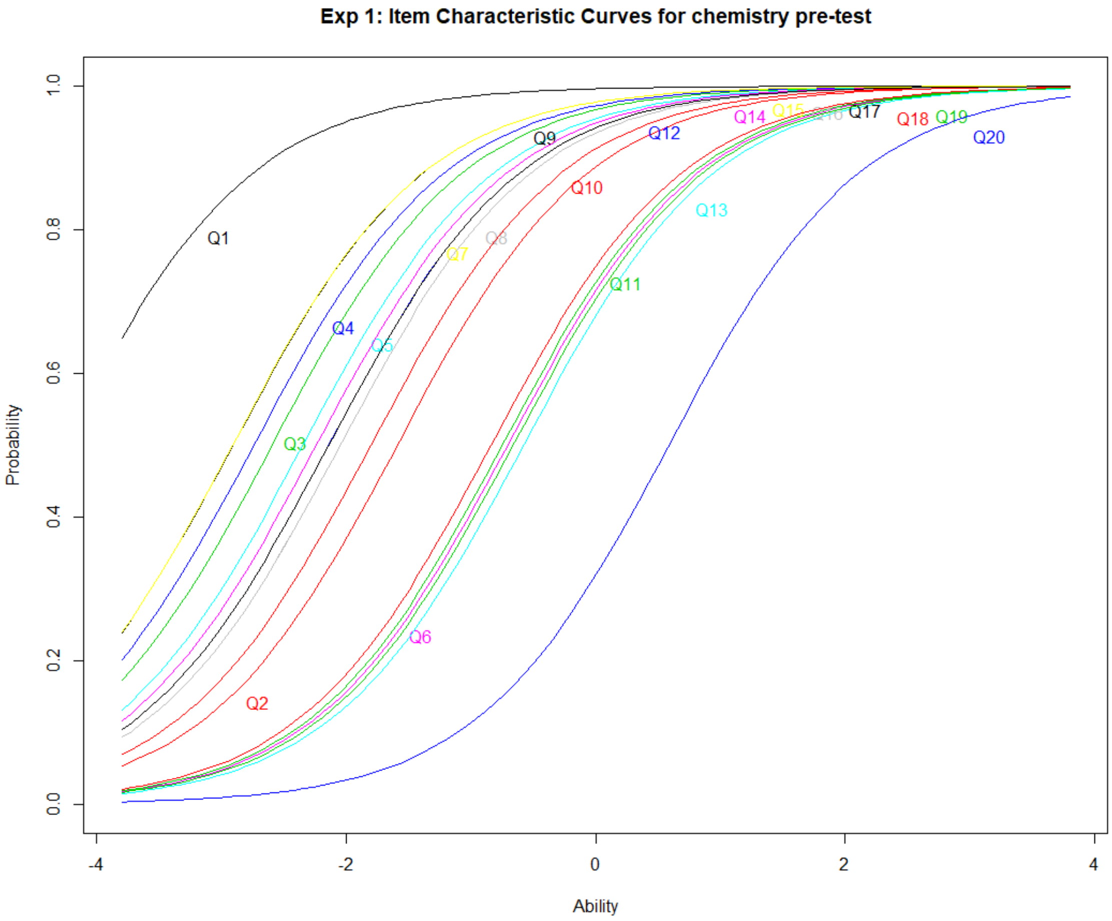


Figure 34. Item characteristic curves for pre-test used in Experiment 1.

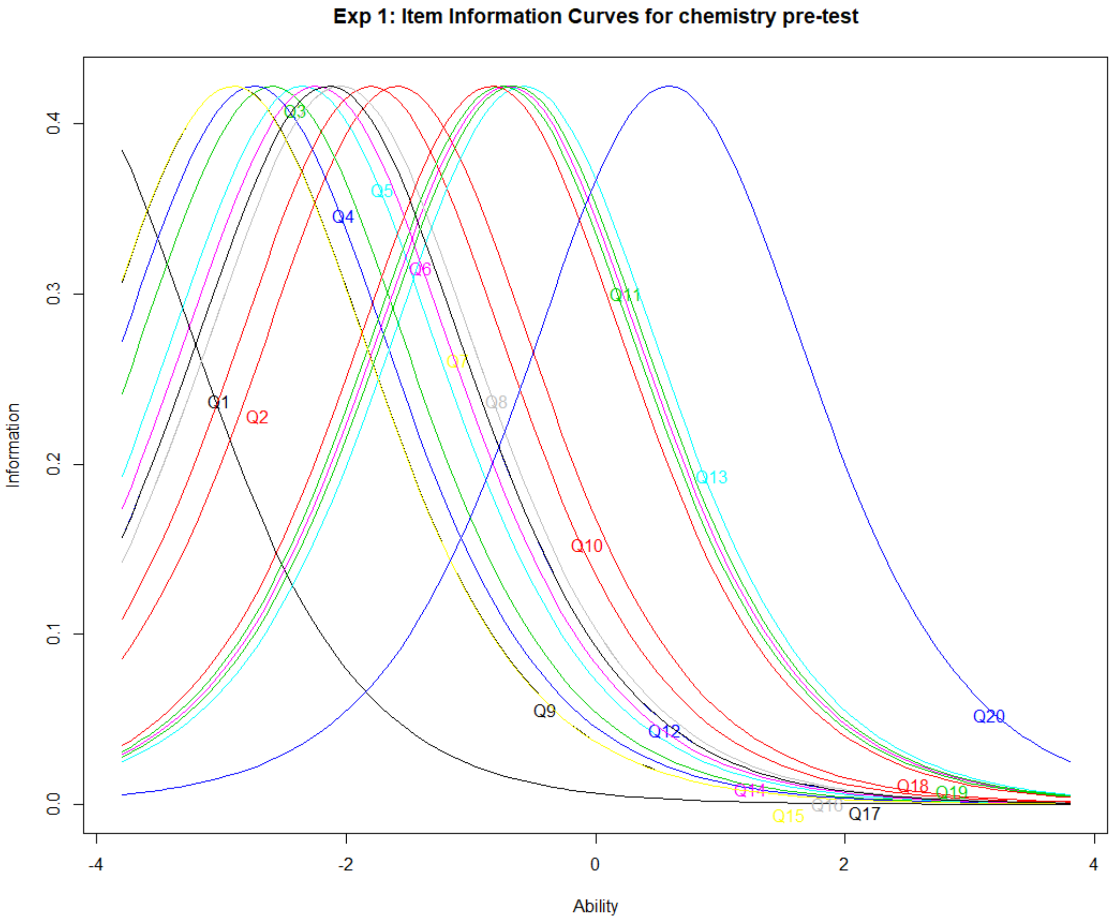


Figure 35. Item information curves for pre-test used in Experiment 1.

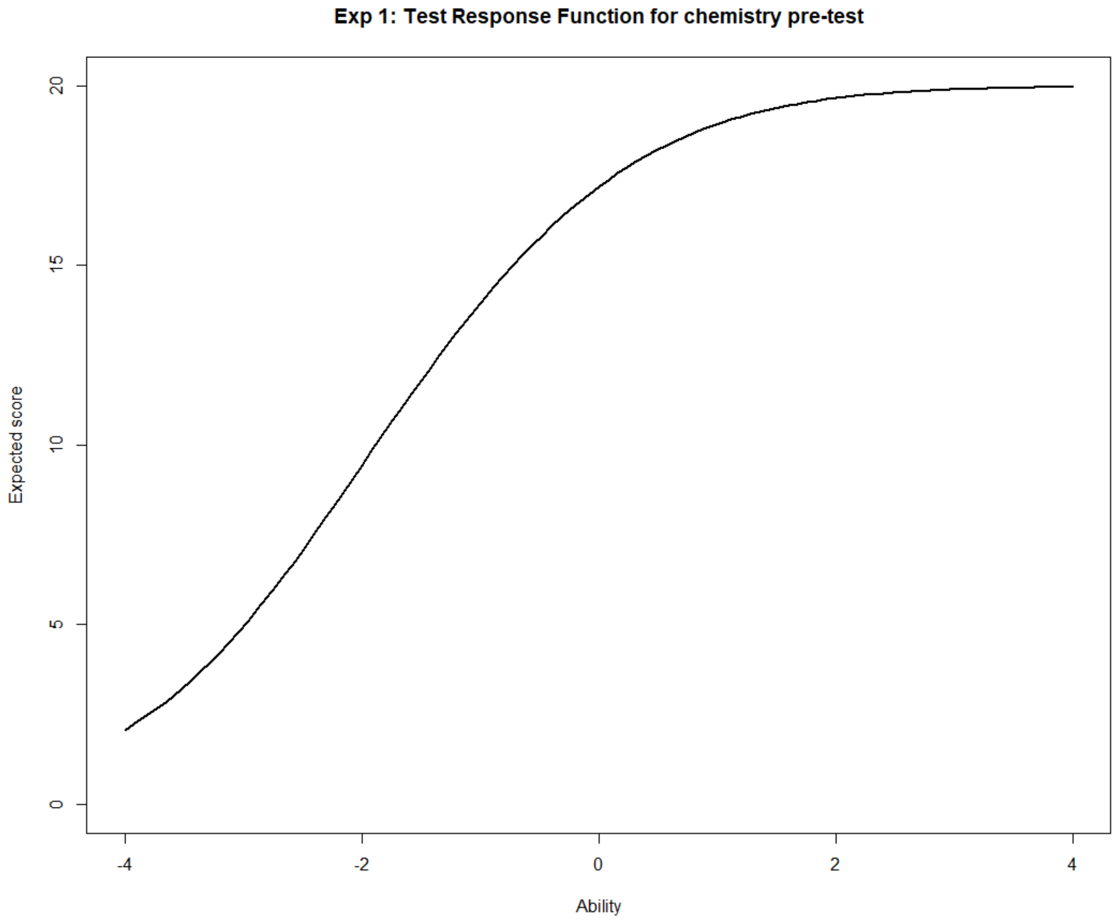


Figure 36. Test response function for pre-test used in Experiment 1.

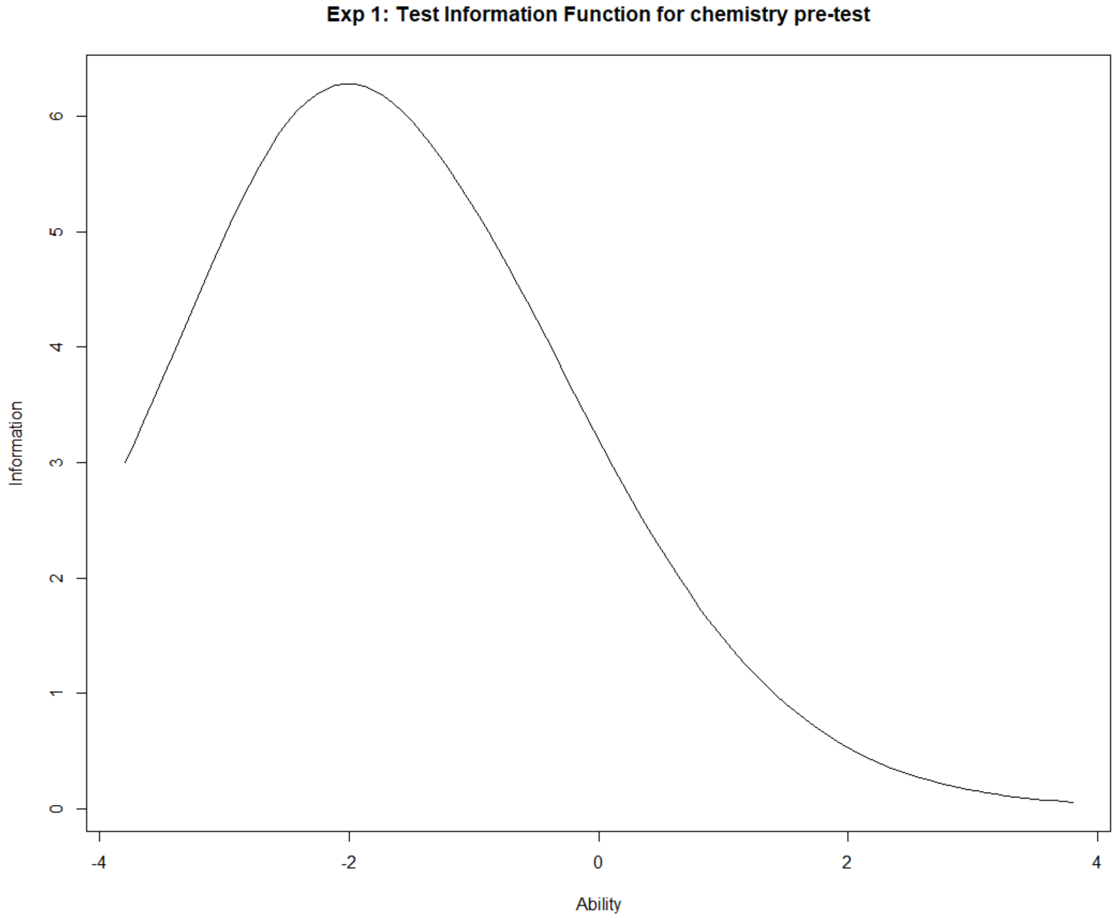


Figure 37. Test information function for pre-test used in Experiment 1.

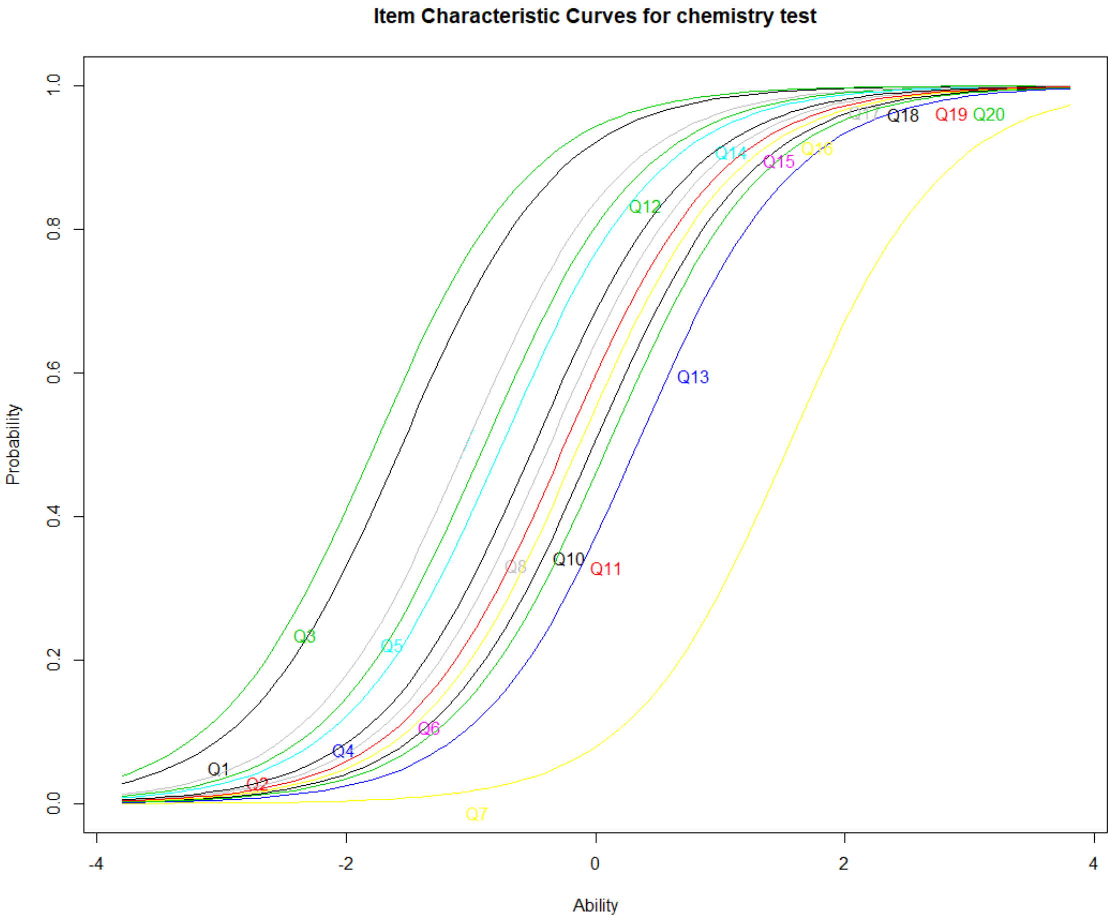


Figure 38. Experiment 2: Item characteristic curves for the test used to validate chemistry items.

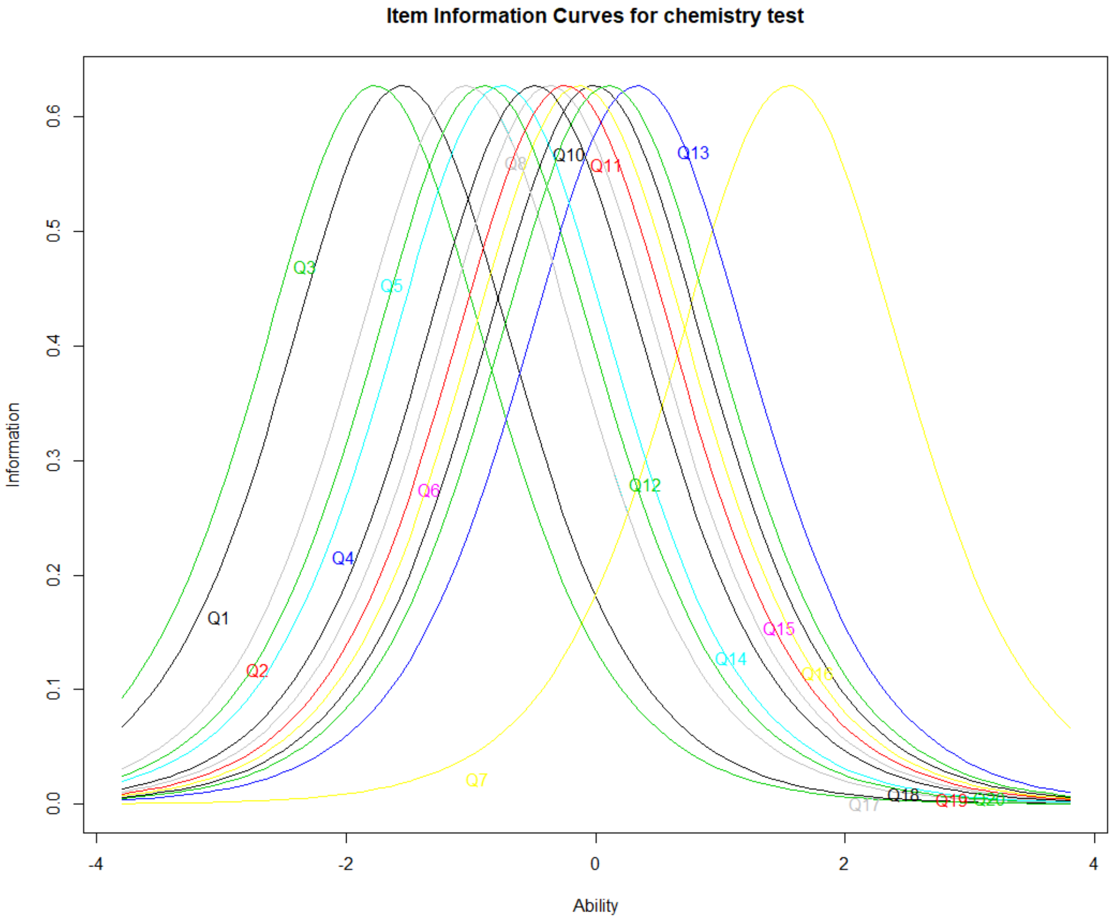


Figure 39. Experiment 2: Item information curves for the test used to validate chemistry items.

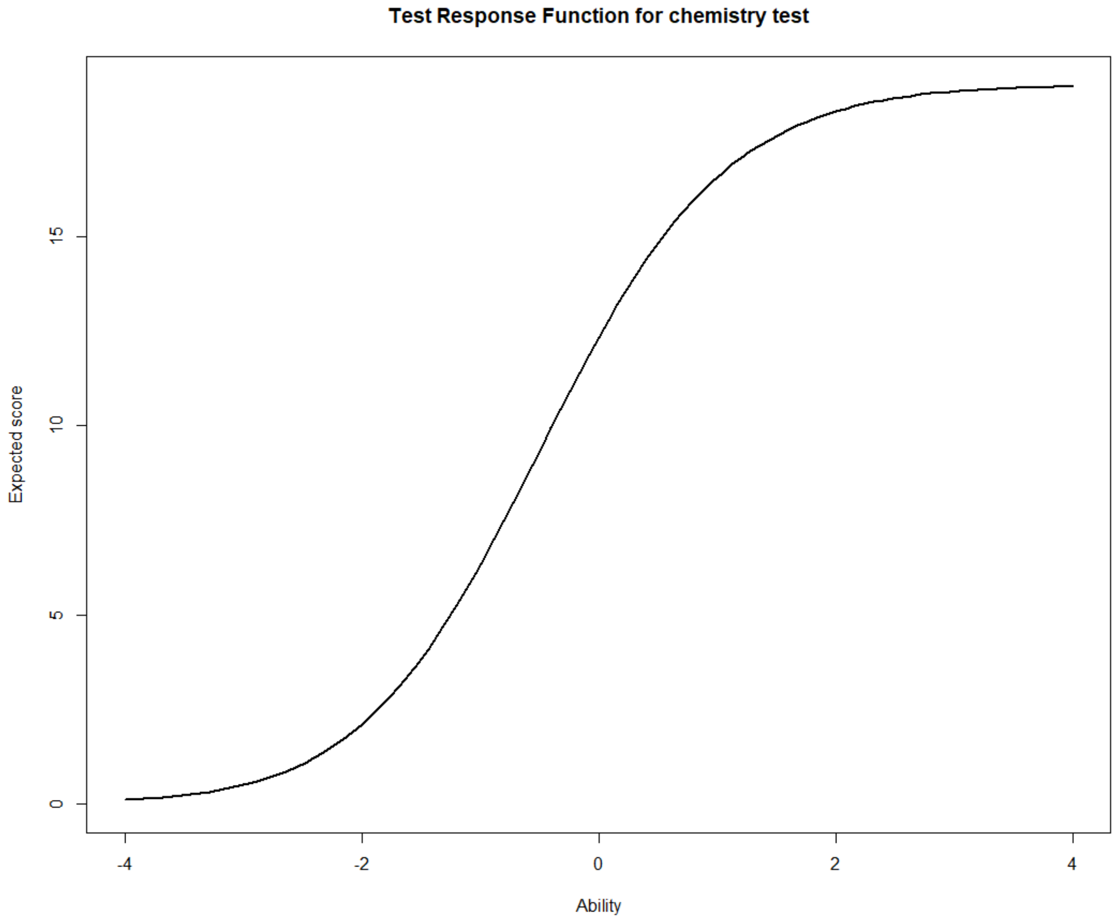


Figure 40. Experiment 2: Test response function for the test used to validate chemistry items.

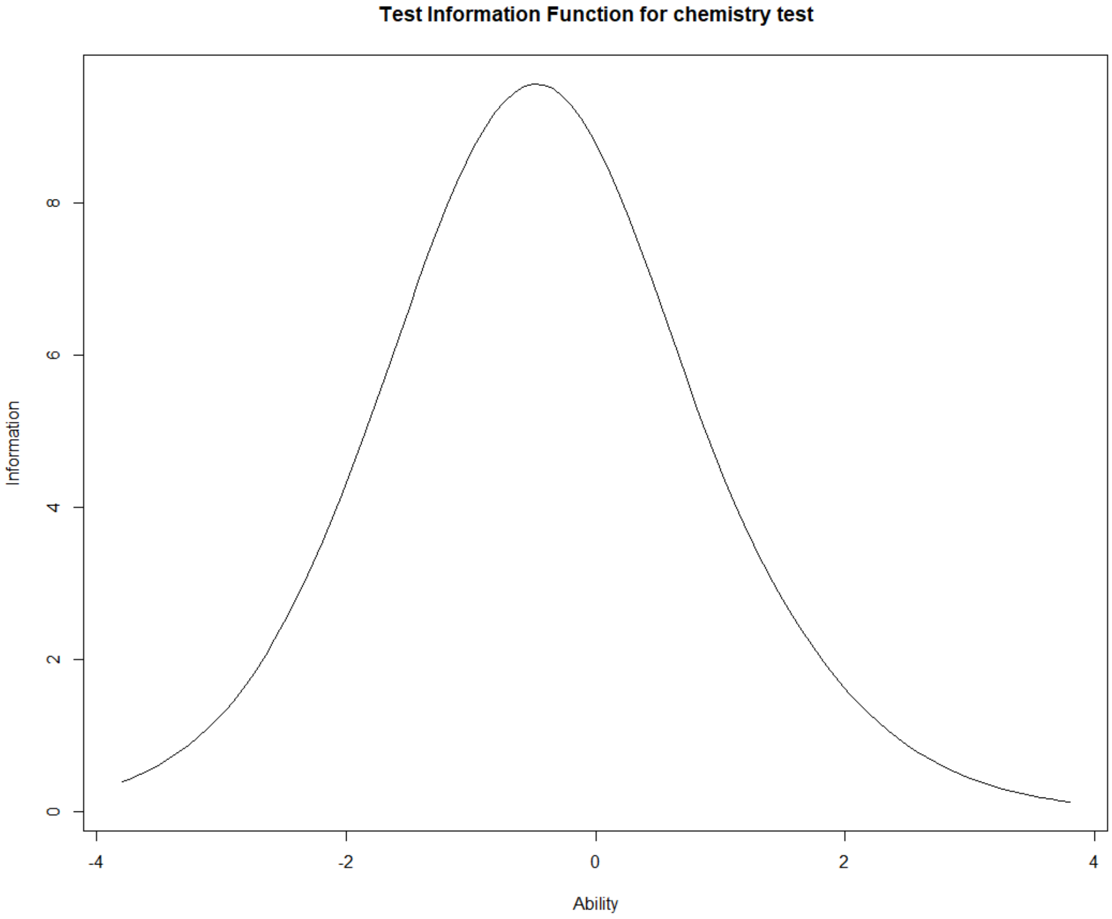


Figure 41. Experiment 2: Test information function for the test used to validate chemistry items.

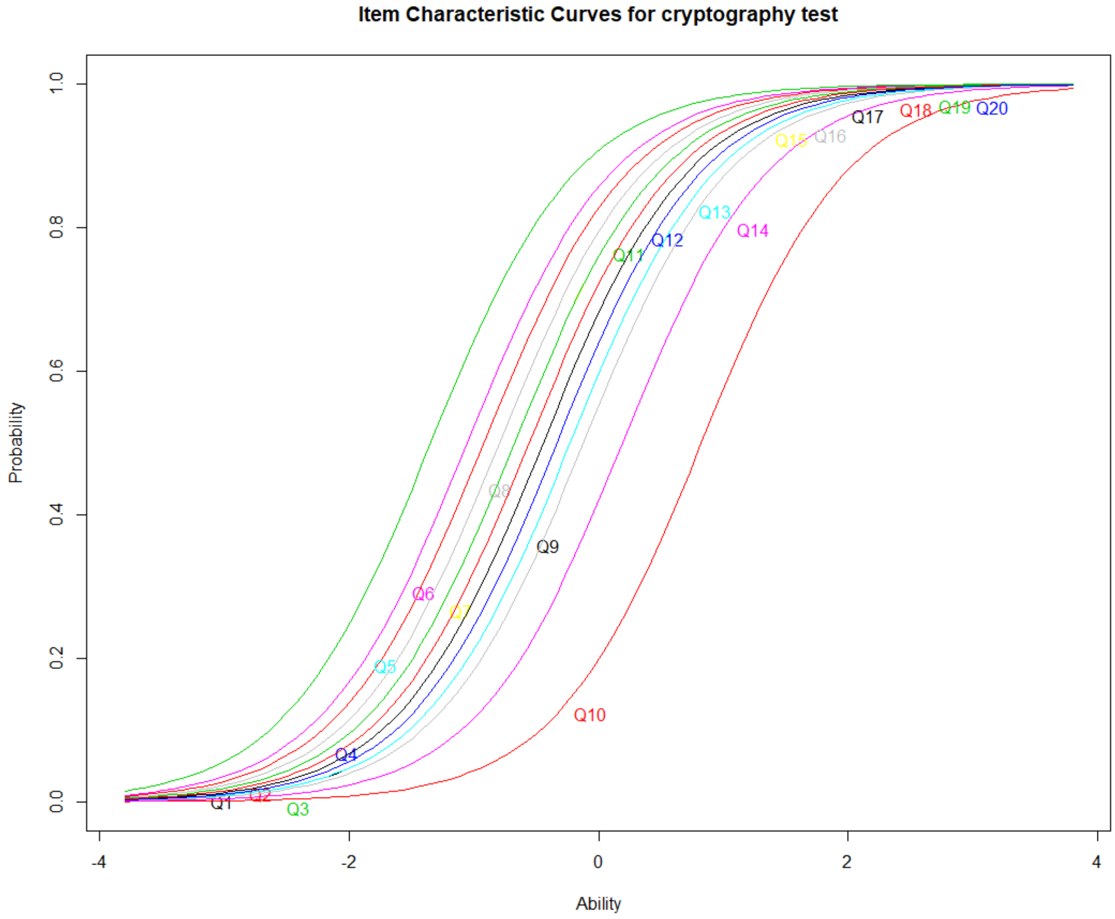


Figure 42. Experiment 2: Item characteristic curves for the test used to validate cryptography items.

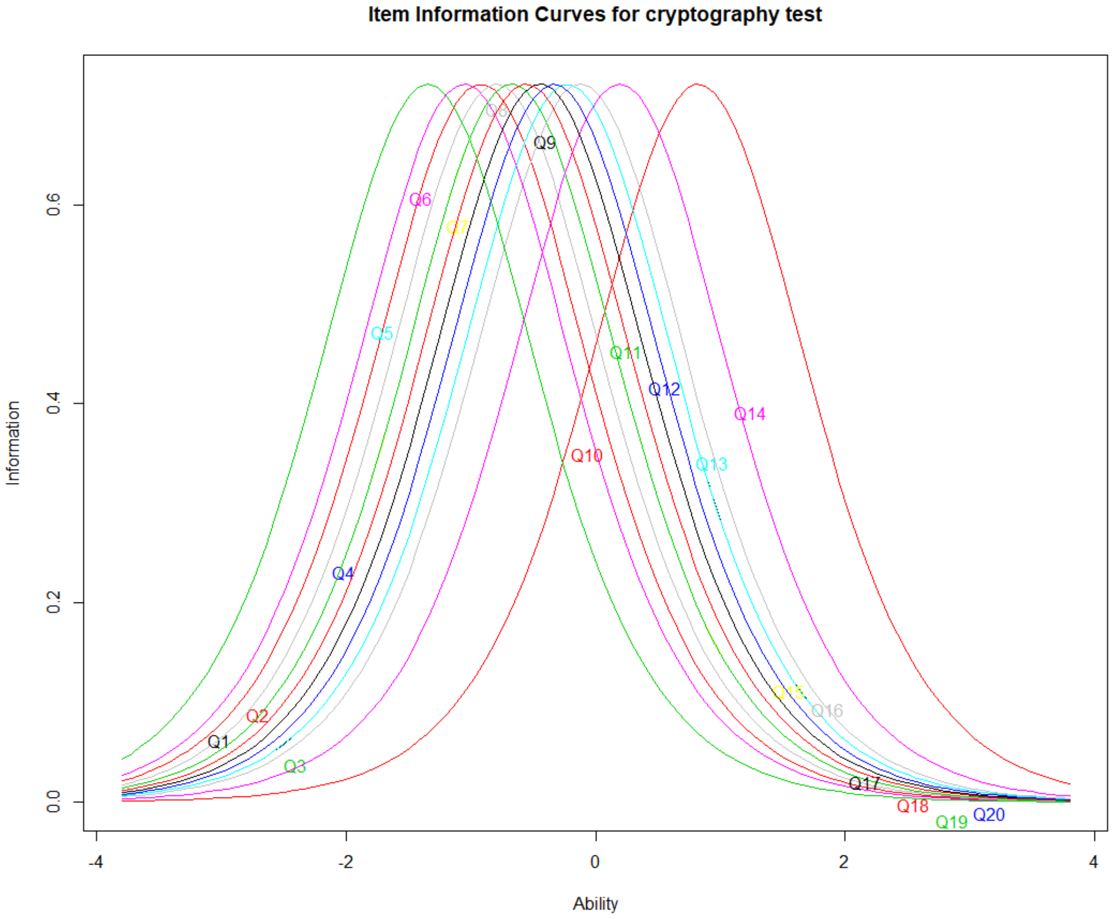


Figure 43. Experiment 2: Item information curves for the test used to validate cryptography items.

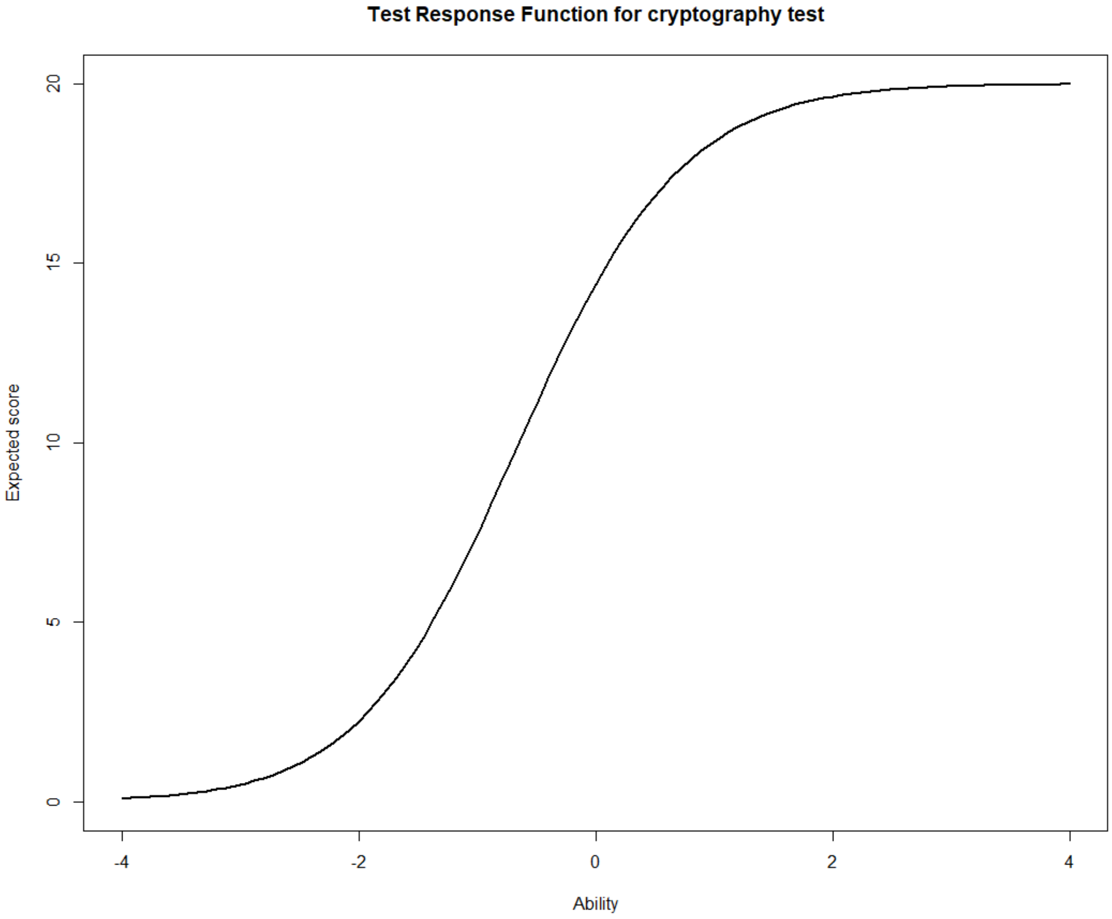


Figure 44. Experiment 2: Test response function for the test used to validate cryptography items.

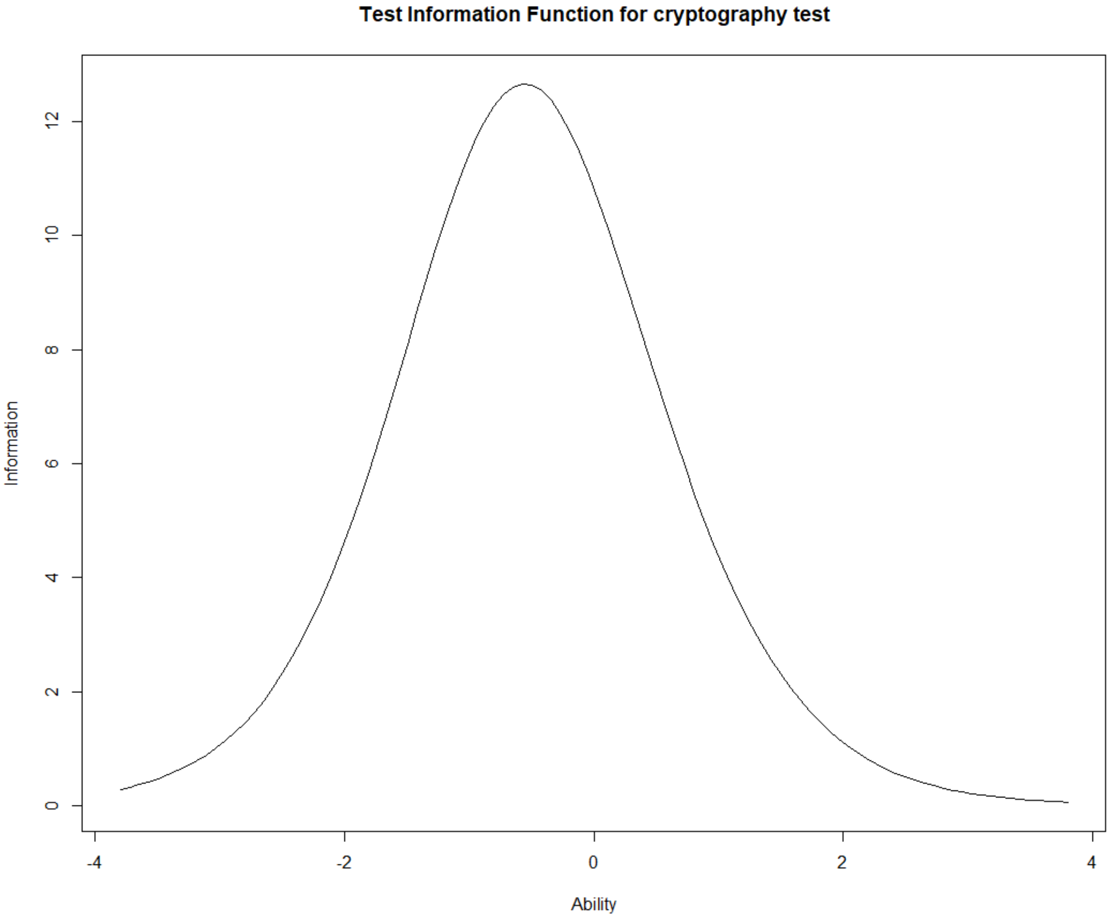


Figure 45. Experiment 2: Test information function for the test used to validate cryptography items.

APPENDIX N

TABLES FOR CONDITIONAL PROBABILITIES OBTAINED FROM PARAMETER LEARNING

Table 16. Exp 1: Learned conditional probabilities for the Prior node.

Data source	Prior Score (pre-test score)										
	0	1	2	3	4	5	6	7	8	9	10
Overall data	.00	.00	.00	.03	.04	.05	.04	.07	.27	.44	.07
Test group	.00	.00	.00	.02	.02	.02	.06	.07	.27	.45	.09
Control group	.00	.00	.00	.04	.06	.08	.02	.08	.27	.42	.04

Table 17. Exp 1: Learned conditional probabilities for Knowledge node at $t = 0$.

Prior	Knowledge0					
	Overall data		Test group		Control group	
	True	False	True	False	True	False
0	.50	.50	.50	.50	.50	.50
1	.50	.50	.50	.50	.50	.50
2	.50	.50	.50	.50	.50	.50
3	.02	.98	.04	.96	.02	.98
4	.73	.27	.96	.04	.60	.40
5	.58	.42	.96	.04	.45	.55
6	.01	.99	.01	.99	.04	.96
7	.62	.38	.50	.50	.74	.26
8	.69	.31	.74	.26	.65	.35
9	.78	.22	.75	.25	.83	.17
10	.86	.14	.82	.18	.98	.02

Table 18. Exp 1: Learned conditional probabilities for overall data.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.52	.48	.01	.99	.00	1.00	.97	.03	.53	.47
False	.99	.01	.94	.06	.28	.72	.58	.42	.34	.66
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.34	.66	.00	1.00	.00	1.00	.92	.08	.80	.20
False	1.00	.00	.64	.36	.20	.80	.75	.25	.11	.89

Table 19. Exp 1: Learned conditional probabilities for the test group.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.53	.47	.02	.98	.01	.99	.97	.03	.62	.38
False	.99	.01	.98	.02	.31	.69	.57	.43	.71	.29
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.62	.38	.01	.99	.00	1.00	.82	.18	.76	.24
False	.99	.01	.98	.02	.47	.53	.94	.06	.35	.65

Table 20. Exp 1: Learned conditional probabilities for the control group.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.53	.47	.03	.97	.01	.99	.97	.03	.61	.39
False	.98	.02	.88	.12	.26	.74	.56	.44	.32	.68
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.30	.70	.00	1.00	.00	1.00	.93	.07	.78	.22
False	.99	.01	.67	.33	.14	.86	.70	.30	.14	.86

Table 21. Exp 2 chemistry: Learned conditional probabilities for the Prior node.

Data source	Prior Score (pre-test score)										
	0	1	2	3	4	5	6	7	8	9	10
Overall data	0.00	0.02	0.10	0.16	0.19	0.13	0.15	0.14	0.11	0.00	0.00
Stealth	0.00	0.00	0.12	0.12	0.22	0.15	0.10	0.17	0.10	0.00	0.00
Non-stealth	0.00	0.03	0.08	0.18	0.17	0.11	0.18	0.13	0.11	0.00	0.00
Chem second	0.00	0.04	0.17	0.18	0.18	0.09	0.17	0.09	0.07	0.00	0.00
Chem first	0.00	0.00	0.04	0.14	0.19	0.16	0.14	0.19	0.14	0.00	0.00

Table 22. Exp 2 chemistry: Learned conditional probabilities for Knowledge node at $t = 0$.

Prior	Knowledge0									
	Overall data		Stealth		Non-stealth		Chem second		Chem first	
	True	False	True	False	True	False	True	False	True	False
0	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
1	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
2	.46	.54	.24	.76	.62	.38	.52	.48	.02	.98
3	.42	.58	.42	.58	.37	.63	.46	.54	.25	.75
4	.69	.31	.59	.41	.72	.28	.64	.36	.70	.30
5	.74	.26	.53	.47	.86	.14	.99	.01	.64	.36
6	.66	.34	.39	.61	.68	.32	.65	.35	.60	.40
7	.73	.27	.51	.49	.78	.22	.65	.35	.73	.27
8	1.00	.00	.99	.01	.99	.01	.99	.01	.99	.01
9	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
10	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Table 23. Exp 2 chemistry: Learned conditional probabilities for overall data.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.39	.61	.01	.99	.00	1.00	.92	.08	.60	.40
False	.99	.01	.84	.16	.30	.70	.53	.47	.35	.65
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.42	.58	.00	1.00	.00	1.00	.92	.08	.77	.23
False	1.00	.00	.80	.20	.39	.61	.69	.31	.12	.88

Table 24. Exp 2 chemistry: Learned conditional probabilities for the adaptive group.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.11	.89	.01	.99	.01	.99	.94	.06	.33	.67
False	.98	.02	.57	.43	.22	.78	.73	.27	.60	.40
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
1.00	.00	.88	.12	.41	.59	.67	.33	.73	.27	
.52	.48	.02	.98	.00	1.00	.93	.07	.28	.72	

Table 25. Exp 2 chemistry: Learned conditional probabilities for the non-adaptive group.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.47	.53	.01	.99	.01	.99	.89	.11	.39	.61
False	.99	.01	.94	.06	.32	.68	.47	.53	.62	.38
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	1.00	.00	.21	.79	.40	.60	.68	.32	.93	.07
False	.39	.61	1.00	.00	.00	1.00	.91	.09	.22	.78

Table 26. Exp 2 chemistry: Learned conditional probabilities when chemistry was played second.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.40	.60	.01	.99	.01	.99	.96	.04	.38	.62
False	.98	.02	.62	.38	.16	.84	.74	.26	.77	.23
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	1.00	.00	.83	.17	.37	.63	.80	.20	.94	.06
False	.30	.70	.01	.99	.00	1.00	.96	.04	.22	.78

Table 27. Exp 2 chemistry: Learned conditional probabilities when chemistry was played first.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.34	.66	.01	.99	.01	.99	.85	.15	.34	.66
False	.99	.01	.96	.04	.39	.61	.44	.56	.37	.63
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	1.00	.00	.98	.02	.52	.48	.61	.39	.54	.46
False	.59	.41	.01	.99	.00	1.00	.81	.19	.28	.72

Table 28. Exp 2 cryptography: Learned conditional probabilities for the Prior node.

Data source	Prior Score (pre-test score)										
	0	1	2	3	4	5	6	7	8	9	10
Overall data	.00	.06	.09	.11	.08	.06	.12	.13	.13	.23	.00
Stealth	.00	.08	.03	.17	.08	.15	.08	.10	.12	.20	.00
Non-stealth	.00	.04	.13	.07	.09	.00	.14	.14	.13	.24	.00
Chem second	.00	.06	.07	.04	.09	.07	.06	.18	.13	.29	.00
Chem first	.00	.06	.11	.18	.07	.04	.18	.07	.13	.16	.00

Table 29. Exp 2 cryptography: Learned conditional probabilities for Knowledge node at $t = 0$.

Prior	Knowledge0									
	Overall data		Stealth		Non-stealth		Chem second		Chem first	
	True	False	True	False	True	False	True	False	True	False
0	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
1	.50	.50	.66	.34	.34	.66	.66	.34	.34	.66
2	.01	.99	.04	.96	.01	.99	.01	.99	.01	.99
3	.14	.86	.01	.99	.31	.69	.02	.98	.10	.90
4	.12	.88	.34	.66	.01	.99	.01	.99	.26	.74
5	.01	.99	.01	.99	.50	.50	.01	.99	.02	.98
6	.34	.66	.41	.59	.30	.70	.01	.99	.40	.60
7	.35	.65	.44	.56	.30	.70	.40	.60	.01	.99
8	.51	.49	.03	.97	.75	.25	.85	.15	.27	.73
9	.65	.35	.75	.25	.63	.37	.56	.44	.80	.20
10	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Table 30. Exp 2 cryptography: Learned conditional probabilities for overall data.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.73	.27	.22	.78	.01	.99	.92	.08	.66	.34
False	1.00	.00	1.00	.00	.90	.10	.63	.37	.19	.81
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.62	.38	.17	.83	.00	1.00	.97	.03	.74	.26
False	1.00	.00	1.00	.00	.81	.19	.67	.33	.18	.82

Table 31. Exp 2 cryptography: Learned conditional probabilities for the adaptive group.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.74	.26	.19	.81	.03	.97	.98	.02	.72	.28
False	.99	.01	.99	.01	.82	.18	.64	.36	.22	.78
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.73	.27	.26	.74	.01	.99	.99	.01	.82	.18
False	1.00	.00	.98	.02	.83	.17	.72	.28	.12	.88

Table 32. Exp 2 cryptography: Learned conditional probabilities for the non-adaptive group.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.73	.27	.24	.76	.01	.99	.88	.12	.61	.39
False	.99	.01	.99	.01	.95	.05	.62	.38	.17	.83
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.54	.46	.10	.90	.00	1.00	.97	.03	.70	.30
False	1.00	.00	1.00	.00	.79	.21	.64	.36	.20	.80

Table 33. Exp 2 cryptography: Learned conditional probabilities when cryptography was played first.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.80	.20	.34	.66	.02	.98	.94	.06	.59	.41
False	.99	.01	.99	.01	.99	.01	.66	.34	.18	.82
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.58	.42	.17	.83	.01	.99	.98	.02	.71	.29
False	1.00	.00	1.00	.00	.79	.21	.74	.26	.18	.82

Table 34. Exp 2 cryptography: Learned conditional probabilities when cryptography was played second.

Knowledge0	Distractor00		Distractor01		Distractor02		Question0		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.62	.38	.03	.97	.02	.98	.92	.08	.70	.30
False	.99	.01	.99	.01	.79	.21	.59	.41	.23	.77
Knowledge1	Distractor10		Distractor11		Distractor12		Question1		Knowledge1	
	True	False	True	False	True	False	True	False	True	False
True	.65	.35	.17	.83	.01	.99	.97	.03	.76	.24
False	1.00	.00	.99	.01	.82	.18	.60	.40	.18	.82