Designing Mnemonics Serious Games to Promote Knowledge Retention in Memory-Intensive Courses

Kingson Fung

A Thesis Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
for the Degree of Master of Science (MSc) in Computer Science

Graduate Program in Electrical Engineering and Computer Science

York University
Toronto, Ontario

December 2025

Abstract

The increased difficulty of memory-intensive courses due to many factors necessitates using technological tools to promote retrieval practice and long-term learning. Hence, a mnemonics game, based on the RADAR framework proposed by Oyibo, was implemented to foster knowledge retention. Fifty-two students, comprising an experimental group ($n = 30$) and a control group ($n = 22$), were recruited to undertake a study, which involved watching a 10-minute Biology lecture on Biology Organization, Cranial Nerves, and Krebs Cycle and taking repeated tests immediately after a 45-minute preparation, one-week gap pre- and post-15-minute revision. In all three tests, students who used the RADAR game performed better in recall than those who did not. Coupled with the study participants stating they found the game easy to use, enjoyable, useful, and trustworthy, and their willingness to adopt it, the experimental group's better performance highlights the need to incorporate mnemonics-based games in memory-intensive courses to promote long-term learning.

Publications

Academic Journals

1. **K. Fung** and K. Oyibo, “Examining the design characteristics of mnemonics serious games on the app stores: A systematic heuristic review,” *Applied Sciences*, vol. 15, no. 14, p. 7772, 2025.
2. **K. Fung** and K. Oyibo, “Examining the effectiveness of mnemonics serious games in enhancing memory and learning: A scoping review,” *Applied Sciences*, vol. 14, no. 23, p. 11379, 2024.
3. F. Noorbehbahani, **K. Fung**, and K. Oyibo, ”Current state of serious games in human trafficking: Evaluation, gaps, and future research directions,” *Multimodal Technologies and Interaction*, vol. 8, no. 11, p. 95, 2024.

Conference Papers

1. **K. Fung**, F. Noorbehbahani, and K. Oyibo, ”An empirical analysis of user experience and educational utility of serious games aimed at creating public awareness about human trafficking,” in *ICERI2024 Proceedings*, pp. 310-319, IATED, 2024.

Acknowledgments

First, I would like to thank Dr. Kiemute Oyibo, my supervisor and a member of my examining committee, for providing all the ideas of the SANKOFA toolkit and the Design Thinking framework that served as the foundation for my thesis. In addition, he had also provided continuous feedback on improving both the Acronimps game and this paper. Moreover, he was the one who analyzed the data gathered from the game evaluation study and shared the resulting graphs and tables with me. He had also designed the experimental materials such as the mnemonics, lecture, surveys and quizzes. Without him, I could never have made it even close to this far.

Second, I would like to thank Dr. Alvine Boaye Belle and Dr. Andrée-Ann Cyr, the chair and internal member of my examining committee, respectively. Alongside Dr. Oyibo, they had taken the time to read my thesis, set up the oral exam, listen to my presentation, ask questions to provide me with the opportunity to prove my knowledge, and create a list of revisions to help improve my paper. Thanks to them, I have succeeded in writing a thesis that is worthy of publication.

Third, I would like to thank Sarah Ejaz, a fellow student in Computer Science (MSc) and a great friend and colleague, for developing the SAVE Tool that would become involved in my study, providing continuous feedback for the Acronimps game, setting up accounts for the participants, and assisting in cleaning the recall response data. She has helped me in many ways throughout our time in the program; without her, my experience would have been far worse.

Fourth, I would like to thank Isaac (Oluwamayowa) Ibidun, another friend and colleague, for presenting his own prototype RADAR game as inspiration, for providing code and continuous feedback for the Acronimps game, and for creating the database that the Acronimps game and SAVE Tool would use. Thanks to him, I avoided becoming overwhelmed by the number of tasks needed for setting up the experiment.

Fourth, I would like to thank my other friends and colleagues, Shogo Kai Toyonaga, Mohammed Abdulmawjood, Cathy (Man Chau) Ng, Marco Antonio Lopez Arriaga, Hooman Hoghooghi Esfahan, Manmeet Walia, Emmanuel Nzeakor, and Dr. Mark Atta Mensah, for providing useful feedback for the Acronimps game. Without their presence, my game would have become much inferior, and I would have been left miserable and lonely.

Fifth, I would like to thank Mr. Timothy Ojo, a product designer and friend of Dr. Oyibo, for providing additional feedback during the development of the Acronimps game. Despite not being part of York University, he had chosen to go out of his way and take time to assist me; I am very grateful for this act of selflessness.

Sixth, I would like to thank Lassonde Graduate Program Assistants Ciara Campbell and Huy Quang Doan for answering many questions I had about technical aspects about the program throughout my time in it. They have always cleared up any confusion I had, and I truly believe they have saved me from missing deadlines for important tasks out of unawareness.

Finally, I would like to thank my parents, Mr. Walter (Wa Tat) Fung and Ms. Karin (Kit Ming) Wong, for always providing support. They would ask and listen to what I had to say, and they would guide me on where to seek advice if they could not provide it. Without them, my life would be in complete disarray.

Contents

Page

Abstract ii

Publications iii

Acknowledgments iv

Contents v

List of Tables viii

List of Figures xi

1 Introduction 1
- 1.1 Problem Overview . . . 1
- 1.2 Memory . . . 1
- 1.3 Introducing Mnemonics . . . 1
- 1.4 Metacognition, Knowledge Construction, and Retrieval Practice . . . 2
- 1.5 The Effectiveness of Mnemonics . . . 3
- 1.6 The Underuse of Mnemonics . . . 3
- 1.7 A Mnemonics Serious Game as Part of a Toolkit . . . 4
- 1.8 Research Question and Hypotheses . . . 5

2 Literature Review 7
- 2.1 Scoping Review of Papers . . . 7
- 2.2 Systematic Review of Apps . . . 7

3 Method 10
- 3.1 Framework . . . 10
  - 3.1.1 Framework Overview . . . 10
  - 3.1.2 First Stage . . . 10
  - 3.1.3 Second Stage . . . 10
  - 3.1.4 Third Stage . . . 11
  - 3.1.5 Fourth Stage . . . 11
  - 3.1.6 Fifth Stage . . . 11
  - 3.1.7 Sixth Stage . . . 12
  - 3.1.8 Seventh Stage . . . 12
- 3.2 Toolkit Description . . . 12
  - 3.2.1 SAVE Tool . . . 13
  - 3.2.2 RADAR Game . . . 14

3.2.3 Quiz Notification System 15
3.3 Features of the SAVE Tool 15
3.3.1 Login 15
3.3.2 Creating and Viewing a Mnemonic 15
3.4 Acronimps, My RADAR Game 16
3.4.1 Setup 16
3.4.2 Quests 18
3.4.3 Town (Recollection) 18
3.4.4 Bridge (Association) 18
3.4.5 Castle (Decoding) 19
3.4.6 Museum (Artifacts Review) 20
3.4.7 Quest Statistics and Settings 22
3.4.8 Class Ranking 22
3.4.9 Image Credits 23
3.4.10 Profile 23
3.4.11 Acronimps Game's Implementation of IN-3PACIFIC and C-SHARP 23
3.5 User Requirements Survey 23
3.5.1 Apparatus 23
3.5.2 Participants 24
3.5.3 Procedure 24
3.6 Game Evaluation 26
3.6.1 Apparatus 27
3.6.2 Participants 27
3.6.3 Experiment Design 27
3.6.4 Procedure 27

4 Results **37**
4.1 User Requirements Survey 37
4.2 Game Evaluation Study 37
4.2.1 Inter-Rater Reliability 39
4.2.2 Topic Familiarity 39
4.2.3 Cognitive Performance 39
4.2.4 UX Variables 48

5 Discussion **49**
5.1 User Requirements Survey 49
5.2 Game Evaluation Study 51
5.2.1 Inter-Rater Reliability 51
5.2.2 Topic Familiarity 51
5.2.3 Cognitive Performance 51

5.2.4 UX Variables . . . 56
5.2.5 RQ: RADAR Game's Relative Improvement in Knowledge Acquisition and Retention 57
5.3 Limitations and Future Work . . . 58

6 Conclusion **61**

Bibliography **62**

Appendices **70**
A Lecture Slides . . . 70
B Demographic Data . . . 74
C Analysis of Test Data . . . 80
C.1 Cohen's Kappa Tables . . . 80
C.2 Mean Difference Graphs . . . 83
C.3 ANOVAs and Pairwise Comparisons . . . 84

List of Tables

1 IN-3PACIFIC Characteristics [1] . . . 8
2 C-SHARP Themes and Their Characteristics [2] . . . 9
3 Acronimps Game's Implementation of IN-3PACIFIC [1]. . . . 24
4 Acronimps Game's Implementation of C-SHARP [2]. . . . 25
5 User requirements survey items (* indicates that the item was reversed during analysis) [3]. . 26
6 Mnemonics for Krebs Cycle Substrates and Reactions. Licensed under CC BY-NC-SA. . . . 28
7 Mnemonics for Cranial Nerves and Functions. Licensed under CC BY-NC-SA. . . . 29
8 Mnemonics for Biological Organization Levels and Disciplines. Licensed under CC BY-NC-SA. . . . 29
9 Recognition questions for Cranial Nerves and Functions. . . . 30
10 Recognition questions for Biological Organization Levels and Disciplines. . . . 31
11 Recognition questions for Krebs Cycle Substrates and Reactions. . . . 31
12 Thematic analysis of participants' comments in pre-system-design user study; categories arranged in descending count order within each type. . . . 38
13 Inter-rater Cohen's Kappa scores for recall-based responses at all three timepoints. C: Control, R: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, OV: Overall . . . 39
14 Overall descriptive statistics. SD: Standard Deviation, SE: Standard Error, C: Control, R: RADAR. . . . 42
15 Overall four-way RM-ANOVA. Df: Degrees of freedom, res: residual, PCode: Participant Code. . . . 43
16 Overall three-way RMANOVAs for recall (middle) and recognition (right) individually based on condition, time, and topic. Df: Degrees of freedom, res: residual, PCode: Participant Code. 43
17 One-way ANOVA based on condition and topic overall due to interaction between them in Table 15. C: Control, R: RADAR, CN: Cranial Nerves, CF: Cranial Nerves Functions, BL: Biological Organization Levels, BD: Biological Organization Disciplines, KS: Krebs Cycle Substrates, KR: Krebs Cycle Reactions. . . . 44
18 Control's and RADAR's topic-based pairwise comparison overall with Tukey method familywise correction due to the simple effect of topic in Table 17. SE: Standard Error, Df: Degrees of freedom, CN: Cranial Nerves, CF: Cranial Nerves Functions, BL: Biological Organization Levels, BD: Biological Organization Disciplines, KS: Krebs Cycle Substrates, KR: Krebs Cycle Reactions. . . . 45
19 One-way ANOVA between condition and task with regards to T1. C: Control, R: RADAR, RCL: Recall, RCG: Recognition. . . . 45
20 One-way ANOVA between condition and task with regards to T2. C: Control, R: RADAR, RCL: Recall, RCG: Recognition. . . . 46
21 One-way ANOVA between condition and task with regards to T3. C: Control, R: RADAR, RCL: Recall, RCG: Recognition. . . . 46

22 One-way ANOVA due to interaction between condition and task in Table 15. C: Control, R: RADAR, RCL: Recall, RCG: Recognition. 47
23 Correlation between topic difficulty and group difference in performance. R: RADAR, C: Control, OV: Overall, RCL: Recall, RCG: Recognition. 47
24 Response distribution for each of the five UX variables. T1: Time 1, T3: Time 3. 48
25 Support of hypotheses (CN: Cranial Nerves, CF: Cranial Nerves Functions, BL: Biological Organization Levels, BD: Biological Organization Disciplines, KS: Krebs Cycle Substrates, KR: Krebs Cycle Reactions, MFHD: Mean Follows Hypothesis's Direction, DS: Difference Significant). 59
C1 Test averages and participant averages for Cohen's Kappa comparing 0 to other values (Left half: control, right half: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, PA: Participant Average, TA: Test Average). 80
C2 Test averages and participant averages for Cohen's Kappa comparing 1 to other values (Left half: control, right half: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, PA: Participant Average, TA: Test Average). 81
C3 Test averages and participant averages for Cohen's Kappa comparing blanks to other values (Left half: control, right half: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, PA: Participant Average, TA: Test Average). 82
C4 Topic-based pairwise comparison of combined recall and recognition scores with Tukey method familywise correction due to the main effect of topic in Table 15. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. . . 84
C5 Time-based pairwise comparison of combined recall and recognition scores with Tukey method familywise correction due to the main effect of time in Table 15. SE: Standard Error, Df: Degrees of freedom, T1: Time 1, T2: Time 2, T3: Time 3. 84
C6 One-way ANOVA based on task and topic due to interaction between them in Table 15. RCL: Recall, RCG: Recognition, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. . . 85
C7 Recall's and Recognition's topic-based pairwise comparison with Tukey method familywise correction due to the simple effect of topic in Table C6. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 85
C8 One-way ANOVA based on task and time due to interaction between them in Table 15. RCL: Recall, RCG: Recognition, T1: Time 1, T2: Time 2, T3: Time 3. 86
C9 Recall's and Recognition's time-based pairwise comparison with Tukey method familywise correction due to the simple effect of time in Table C8. SE: Standard Error, Df: Degrees of freedom, T1: Time 1, T2: Time 2, T3: Time 3. 86

C10 Recall's and Recognition's topic-based pairwise comparison with Tukey method familywise correction due to the main effect of topic in Table 16. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 86
C11 Recall's and Recognition's time-based pairwise comparison with Tukey method familywise correction due to the main effect of time in Table 16. SE: Standard Error, Df: Degrees of freedom, T1: Time 1, T2: Time 2, T3: Time 3. 86
C12 One-way ANOVA due to interaction between condition and topic in Table 16 for recall only $[F(5,850) = 4.34, p = 0.0007]$. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 87
C13 Control's and RADAR's topic-based pairwise comparison for recall with Tukey method familywise correction due to the simple effect of topic in Table C12. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 87
C14 One-way ANOVA due to interaction between condition and topic in Table 16 for recognition only. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. . . 88
C15 Control's and RADAR's topic-based pairwise comparison for recognition with Tukey method familywise correction due to the simple effect of topic in Table C14. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 88

List of Figures

1 A diagram provided by [4] showing the hierarchy in Bloom's taxonomy [5]. 2
2 Stages of the research program and key methods to realize them [6]. 12
3 Long-term learning process based on the SANKOFA framework [6], with images provided by [7] [8] [9] [10] [11] [12] [13]. 13
4 Example mnemonic cue [6] [14]. 14
5 Screenshots of the setup and story pages in the Acronimps game, with textures provided by [15] [16] [17], avatar provided by [18], course images provided by [19] [20] [21] [22] [23], and the fire image provided by [24]. Licensed under CC BY-NC-ND. 17
6 Screenshots of "Main Menu" and "Town (Recollection)," with textures provided by [15] [16] [17], avatar provided by [18], location background provided by [25], and bottom icons provided by [26] [27] [28]. Licensed under CC BY-NC-ND. 19
7 Screenshots of "Bridge (Association)," with textures provided by [15] [16] and location background provided by [29]. Licensed under CC BY-NC-ND. 20
8 Screenshots of "Castle (Decoding)" and "Aftermath," with textures provided by [15] [16] and location background provided by [30]. Licensed under CC BY-NC-ND. 21
9 Screenshots of "Museum (Artifacts Review)," with textures provided by [15] [16] [17] and images provided by [31] [32] [33]. Licensed under CC BY-NC-ND. 21
10 Screenshots of the additional features, with textures provided by [15] [16] [17], avatars provided by [18], and course image provided by [21]. Licensed under CC BY-NC-ND. 22
11 Part 1 of the study (DM stands for demographic). 33
12 Part 2 of the study. 33
13 Graphs showing mean scores for each item among the Prof group and the Self group [3], with error bars showing standard deviation and dotted line showing neutral value in range 1 to 7. 38
14 Mean familiarity scores across topics for the control and experimental groups, with error bars showing standard deviation and dotted line showing neutral value in range 1 to 7. C: Control, R: RADAR, BL: Biological Organization Levels, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 40
15 Mean recall scores for the specific topics at the three points in time and overall, with error bars showing standard deviation and dotted line showing neutral score in range 0 to 1. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 41
16 Mean recognition scores for the specific topics at the three points in time and overall, with error bars showing standard deviation and dotted line showing neutral score in range 0 to 1. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. . . 41

17 Overall mean scores of the remembering measure (aka recall - left) and understanding measure (aka recognition - right) for the three points in time, with error bars showing standard deviation and range being 0 to 1. C: Control, R: RADAR. 42
A1 A slide showing details of Krebs Cycle Substrates and Reactions. 70
A2 A slide showing details of Cranial Nerves and Functions. 70
A3 Slides showing details of Biological Organization Levels based on [34], with lion images provided by [35]. 71
A4 Slides showing details of Biological Organization Disciplines. 71
A5 A slide showing mnemonics for Krebs Cycle Substrates and Reactions. 72
A6 A slide showing mnemonics for Cranial Nerves and Functions. 72
A7 A slide showing mnemonics for Biological Organization Levels and Disciplines based on [34]. 73
B1 Pie charts showing gender distributions among the control group (top) and experimental group (bottom). 74
B2 Pie charts showing age distributions among the control group (top) and experimental group (bottom). 75
B3 Pie charts showing distributions for year of study among the control group (top) and experimental group (bottom). 76
B4 Pie charts showing distributions for course of study among the control group (top) and experimental group (bottom). 77
B5 Pie charts showing distributions for race or ethnicity among the control group (top) and experimental group (bottom). 78
B6 Pie charts showing distributions for average grades among the control group (top) and experimental group (bottom). 79
C1 Topic-based performance mean difference between each pair of times in which participants were tested for recall (aka remembering - left) and recognition (aka understanding - right), with dotted line showing no difference in range -1 to 1. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates. 83

1 Introduction

1.1 Problem Overview

It has become increasingly difficult for students of various education levels, especially university, to learn and remember what they learned due to numerous negative factors. One such factor is the way memory-intensive subjects such as biology [36], psychology [37], history [38], social science [39] [40], humanities [41] [42], and languages [43] are taught, alongside the very nature of the subjects themselves [36]. Other factors include the prominence of distracting technologies especially social media networks [44] [45], the fact that many students have to work to financially support themselves [46], the constantly-increasing body of knowledge, and the emphasis on studying individually rather than collaboratively [47]. All these factors take a heavy toll on students' memory, preventing them from achieving their full potential.

1.2 Memory

Memory is considered a low-order cognitive process in Bloom's taxonomy, but it serves as a foundation for higher-order cognitive processes, such as analysis and evaluation, making it very important in the act of learning [5]. Even without that fact, it should be obvious that memory is an extremely important cognitive function that dictates much of our daily lives, including basic actions such as eating, reading, and writing [48]. The concept of memory is more complex than it may seem. Atkinson and Shiffrin [49] describe a "three-box" model that splits memory into three types: sensory, short-term, and long-term. First, information processing begins in the sensory memory; unignored information goes to the short-term memory. Performing maintenance rehearsal (i.e., rote memorization) on information in the short-term memory keeps it there longer, otherwise it might be lost [50]. To move it into long-term memory and keep it there, a different kind of rehearsal, namely elaborative rehearsal (i.e., establishing connections between this information and prior knowledge) must be used [49] [51]. As newer information depends on older information, any information we learn may be needed for the future. Jensen describes learning and memory as "two sides of a coin" [52], asserting that they are inseparable.

1.3 Introducing Mnemonics

The great importance of learning and memory, along with the aforementioned factors creating problems for them, results in a greater urgency for techniques that enhance learning and retaining learned content. Fortunately, there are already such techniques in place; one notably successful one is the use of mnemonics. Mnemonics enhance recall by establishing connections between the new information and concepts the user is already familiar with, which means they demonstrate elaborative rehearsal. For a mnemonic to be effective however, it must demonstrate four properties: meaningfulness, elaboration, association, and visualization [44]. The problem is, as mentioned before, the modern era contains an increasing number distracting technologies causing a decline in memory ability overall [44], which means it is harder to create effective mnemonics. Thus, instructors must not only help students learn, but also assist in the methodological and metacognitive aspects of learning; in other words, they must help students learn how to learn [44].

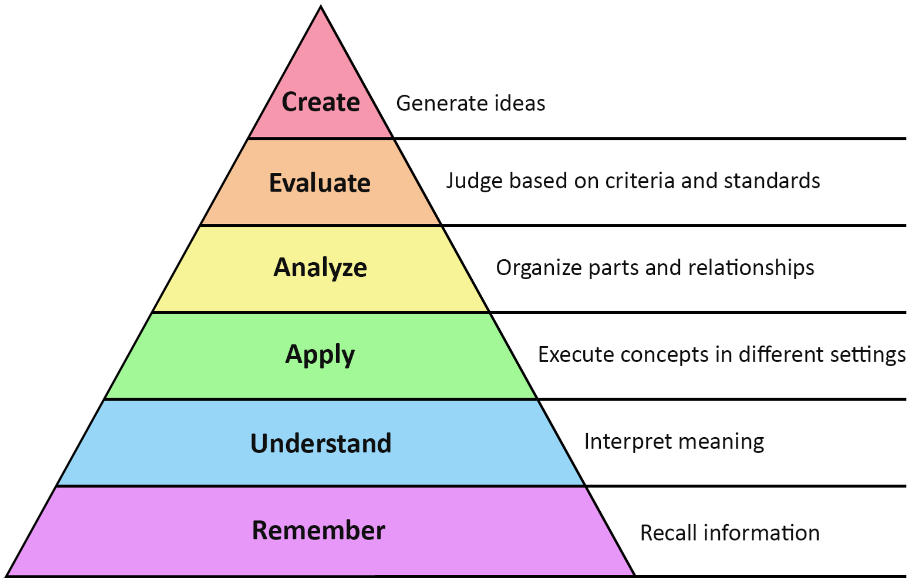


Figure 1: A diagram provided by [4] showing the hierarchy in Bloom's taxonomy [5].

Having mentioned Bloom's taxonomy [5], it is appropriate to describe it in more detail and explain mnemonics' relevance to it. Bloom's taxonomy is a hierarchical framework that classifies several learning processes into different levels of order, with higher-order processes being more complex; these processes will be listed from lowest to highest order. First, one must remember the content that is being taught. Second, they must understand what they learned or else it would become meaningless to them. Third, they must actually go apply what they learned or else it would become pointless to them in their life. Fourth, they must analyze what they learned by looking for deeper meanings within the content and making inferences. Fifth, they must evaluate what they learned, use it to defend their opinions, and gain greater autonomy. Finally, they must use what they have learned to create new data. Creating mnemonics based on the content is clearly an example of the last process, but successfully completing the previous processes is required to create a mnemonic that easily maps to the content by matching the number of major terms in the content and referencing the topic and meaning of the content. In addition, using Bloom's taxonomy on the four properties of an effective mnemonic [44] will also help create effective mnemonics. Figure 1 shows the six processes, with their heights matching their relative orders.

1.4 Metacognition, Knowledge Construction, and Retrieval Practice

Speaking of metacognition, this is when someone thinks about and even evaluates what and how they learn, hence the prefix "meta" followed by "cognition." Unfortunately, students are generally not very great at analyzing their own learning ability, which in turn, hinders said learning ability [53]. As such, the art of metacognition itself should be taught to students [54] [55] [56] [57]. In particular, according to Bransford et al. [58], it should be taught in a way that reflects the context of the other information that is to be taught, for maximum benefit. Mnemonics can be evaluated with regards to the four aforementioned properties they should have (meaningfulness, elaboration, association, and visualization) [44], meaning they can be involved with metacognition.

We will also discuss the concepts of knowledge construction and retrieval practice, as they will be relevant for later. Knowledge construction involves actively building one's understanding of a certain topic through exploration, elaboration, reflection, and interaction, all for the sake of promoting deeper learning and understanding [59]. It is related to the constructivist learning theory which focuses on a more proactive approach by building one's own knowledge and actively participating in the learning process [60] [61]. Also, since knowledge construction involves building one's own knowledge bit by bit, there is room for metacognition as well [62]. Unfortunately, students tend to ignore knowledge construction in favor of rote memorization [63] [64] [65] [66]. Yet the truth remains; despite looking simple and efficient, rote memorization is not only actually daunting [67], but also focuses on short-term knowledge retention rather than long-term. Mnemonics give long-term memory the attention it deserves, and like knowledge construction, they involve associating new information with prior knowledge, which is why we talked about knowledge construction; the two can go hand-in-hand.

Retrieval practice, like knowledge construction, also emphasizes proactivity, as users actively put their long-term memory to the test. It enhances learning [68], and may even be more beneficial than repeated study and elaborative encoding [69] [70] [71]. In addition, retrieval practice promotes gaining knowledge that is more flexible with regards to both being retrieved from long-term memory and being applied to various contexts. Moreover, because it enhances long-term memory, it can be beneficial in academia [72] [70]. It works even better with feedback, as users will then be aware whether their memory is correct or not. Mnemonics provide more benefit when paired with retrieval practice, as the latter gives them a greater advantage over repeated studying [72].

1.5 The Effectiveness of Mnemonics

We return to focusing on mnemonics. Mnemonics have been shown to help users learn and remember large amounts of information, particularly in memorization-intensive courses [73]. Mnemonics have shown to be effective in learning course materials in various disciplines including engineering [74], science [75], social science [39] [40], mathematics/statistics [76], economics [44], humanities [41] [42], and more. Research has shown that self-created mnemonics are more likely to be effective as new information to be learned can be linked to the creator's prior knowledge and experiences [74], making them more memorable and easier to relate with [74] [75] [76]. While mnemonics involve more mental effort on the front end of the encoding process, it pays off in the long run [39]; it helps to free up cognitive resources for higher-order thinking and reduce stress and anxiety during tests [76]. Despite mnemonics being shown to be effective by several studies cutting across different disciplines, instructors and students are free to pair them with other techniques if they wish [77].

1.6 The Underuse of Mnemonics

Mnemonics are not new; they have a long history of helping people remember important and easily forgotten information [75] [78]. Yet somehow, they are still underused in university education [79], particularly in memory-intensive courses. There are several reasons for this. One of such reasons is that many instructors and students are simply unaware or unfamiliar with them and/or their effectiveness [80] [53] [81]. Hence, training students and instructors alike how to use mnemonics effectively in their courses would likely increase their adoption and usage in promoting long-term memory [53] [54]. For example, research shows that students instructed to create and use mnemonics effectively showed improved long-term learning in memorization-intensive courses such as biology and psychology [82] [53]. Another reason for the underuse of mnemonics is that numerous students still dismiss mnemonics as inferior to other learning strategies [53]. Hence, demonstrating the effectiveness of mnemonics to them may lead them to give mnemonics a try. A third reason is that it may be risky for courses in their current state, particularly memory-intensive ones, to integrate mnemonics since there is already a lot of content to be taught, and spending lots of time on what may be considered as merely an alternative method goes against the generally pragmatic nature of most courses [77]. Perhaps the only hope for changing them is to come up with a method that does not take away too much time from teaching the regular material. As such, we need a mnemonics-based educational toolkit than can speak to all three reasons. First, it must demonstrate the effectiveness of mnemonics in promoting long-term learning and memory. Second, it must be easily incorporated into courses. Third, it must train instructors and students on the effective use of mnemonics. Finally, it must facilitate mnemonics creation and retrieval practice of encoded information. So far, there do not seem to be any such tools, let alone any successful ones. There also does not seem to be any research on the potential of such tools being used in courses, particularly memory-intensive ones, across all levels of education [6].

1.7 A Mnemonics Serious Game as Part of a Toolkit

To address the lack of supportive tools for instructors and students, Oyibo [6] has proposed the creation of a toolkit (named SANKOFA) that is meant to be used in various memory-intensive courses. The toolkit consists of three main components: (1) a mnemonics creation tool, referred to as the SAVE Tool, to help users effectively and easily create their own mnemonics to enhance long-term learning, (2) a serious game, referred to as the RADAR game, to help users practice retrieval of encoded information to foster episodic memory, and (3) a quiz notification system to help users practice retrieval of encoded material from time to time to strengthen acquire knowledge. The toolkit and its components will be elaborated on throughout the Method section. He has enlisted the help of several students who focused on the different components. Since I, in particular, worked on the RADAR game, I will briefly explain serious games in general and with regards to mnemonics.

Serious games are games that prioritize education over entertainment [83], which is why they are referred to as “serious.” Nevertheless, serious games can be entertaining, as fun can have a positive effect on players and their learning both in and outside any game by serving as a force of motivation [84]. Research shows that serious games hold prospects in promoting knowledge acquisition [85] and skill development through training [86]. They have been shown to to be effective in various fields, including but not limited to health education [87], medical care [88], chemistry [89], and cybersecurity [90]. Because of their effectiveness, they have been increasingly incorporated into academia [83] [91] [92] [93].

1.8 Research Question and Hypotheses

The overwhelming evidence on the effectiveness of mnemonics [73] [39] and serious games [85] [86] in promoting training and learning in various domains leads to the research question, "*How can mnemonics-based games be effectively designed to promote long-term learning and memory?*" To answer this research question, I leveraged the mnemonics-based framework proposed by Oyibo [94] called "RADAR", in conjunction with the SAVE Tool, to design, implement and evaluate a story-based game called "Acronimps" to promote knowledge acquisition and retention in university classrooms. The word "RADAR" stands for ***R**ecollection* (of created mnemonics using the SAVE Tool), ***A**ssociation* (of mnemonics with their target topics), ***D**ecoding* (of mnemonics into their target information) and ***A**rtifacts* ***R**eview* (reviewing and reflecting on the textual and visual information associated with the mnemonics).

Mnemonics serious games, especially those designed based on a framework that is replicable across courses, domains and researchers, turn out to be a rather niche topic in mnemonics game research. Based on a scoping review conducted by Oyibo and myself [95], mnemonics serious games are not so common and there is not much discussion about them in the extant literature. Moreover, the existing games on the market, including Google and Apple stores, are not particularly designed for academia, let alone based on a well-articulated framework that can be utilized by developers and researchers in the future to foster the reproduction of research findings based on similar game design characteristics and user experience [96].

Before I outline the thesis hypotheses, I will first define two key terms or concepts used throughout this thesis. The first of such terms is "knowledge acquisition," which is the amount of knowledge acquired *immediately* after studying a given educational material. It is measured as the extent to which learners can recall learned terms and/or recognize the correct answers to multiple-choice questions, both of which test their remembering and understanding of the studied material, respectively. Information recall and correct-answer recognition, as used in this thesis, correspond to remembering and understanding in Bloom's taxonomy, respectively [5]. The definition of the second term "knowledge retention" is similar to "knowledge acquisition"; the only difference is that the former is based on and assessed after a long-time period such as days and weeks, while the latter is assessed immediately after test-taking.

In my thesis experiments, knowledge acquisition was evaluated five minutes after participants finished watching a 10-minute lecture, studying with the SAVE Tool and engaging in retrieval practice with the RADAR game for a total of 45 minutes (Time 1). Moreover, knowledge retention was evaluated after a week gap without revision (Time 2) and after 15-minute revision (Time 3). The same set of questions was asked across the three points in time. Further details on each time will be provided throughout the Method section.

Based on the overwhelming evidence of the effectiveness of serious games [85] [86], retrieval practice [68] [69] [70] [71] and mnemonics [73] [39] [76], prior to my experiment aimed to test the effectiveness of the RADAR game compared with a control condition (traditional way of learning using PowerPoint), I outlined a number of hypotheses. The experiment, repeated across three points in time, tested the study participants on two cognitive tasks: recall (RCL) and recognition (RCG). It was based on a 10-minute biology video lecture that covered six important topics, namely, Biological Organization Levels (BL), Biological Organization Disciplines (BD), Cranial Nerves (CN), Cranial Nerves Functions (CF), Krebs Cycle Substrates (KS), and Krebs Cycle Reactions (KR). The hypotheses, which compare the cognitive performance of participants using the RADAR game (R) with that of those in the control condition (C), are as follows:

- Overarching RQ: Will the RADAR game improve knowledge acquisition and retention?
- "R $>$ C for CN, CF, BL, BD, KS, and KR": The RADAR game will lead to higher students' performance in each of the six topics than the control condition because of the benefits of mnemonics [73], especially when paired with retrieval practice [72].
- "R $>$ C for RCG and RCL at T1": The RADAR game will lead to improved performance in recognition and recall tasks at Time 1 compared to the control condition because of the benefits of mnemonics [73], especially when paired with retrieval practice [72].
- "R $>$ C for RCG and RCL at T2": The RADAR game will lead to improved performance in recognition and recall tasks at Time 2 compared to the control condition because of the benefits of mnemonics [73], particularly regarding long-term memory [39], and pairing them with retrieval practice [72].
- "R $>$ C for RCG and RCL at T3": The RADAR game will lead to improved performance in recognition and recall tasks at Time 3 compared to the control condition because of the benefits of mnemonics [73], particularly regarding long-term memory [39], and pairing them with retrieval practice [72].
- "R $>$ C for RCG and RCL overall": The RADAR game will lead to improved performance in recognition and recall tasks overall (regardless of time) compared to the control condition because of the benefits of mnemonics [73], particularly regarding long-term memory [39], and pairing them with retrieval practice [72].
- "Group difference for difficult topics $>$ for simple topics (overall, RCG and RCL)": The difference between the two groups' scores will be greater in difficult topics than in simple topics overall and in both recognition and recall tasks because of the ceiling effect [97].
- "RCG $>$ RCL overall, for C, and for R": The recognition task will have a higher score than the recall task overall and within both the control and experimental groups because recognition is easier than recall [98].

Additionally, those in the RADAR group were asked to provide quantitative feedback on the game with respect to their user experience and willingness to adopt the game in real-life settings. Given that the RADAR game leveraged established design guidelines such as IN-3PACIFIC [1] and C-SHARP [2], the formulated hypotheses, relating to Perceived Ease of Use (PEOU), Perceived Enjoyment (PE), Perceived Usefulness (PU), Perceived Trust (PT), and Willingness To Adopt (WTA), are as follows:

- "Mean of PEOU, PE, PU, and PT $>$ neutral value": The RADAR game will be perceived as easy to use, enjoyable, useful, and trustworthy because it was designed based on two established frameworks (IN-3PACIFIC [1] and C-SHARP [2]).
- "Mean of WTA $>$ neutral value": Students will be willing to adopt the RADAR game given its knowledge consolidation benefit, perceived ease of use, perceived enjoyment, perceived usefulness, and perceived trustworthiness.

2 Literature Review

As stated before, there has not been much discussion of mnemonics serious games. As a result, I, alongside others including Oyibo, had performed a scoping review of papers that involved creating and evaluating mnemonics serious games [95], as well as a systematic review of apps that are themselves mnemonics serious games [96].

2.1 Scoping Review of Papers

We [95] had searched ACM Digital Library, IEEE Xplore, Scopus, and Web of Science to find papers detailing studies involving creating and evaluating mnemonics serious games. Surprisingly, only 1 relevant paper was found, so we manually searched Google Scholar and used the snowballing method to find 11 more papers, resulting in a total of 12. Through these papers, we saw that after the less-than-positive results of the first two studies in 1996 [99] [100], no more studies were conducted until 2011 [101], and more began to pop up from then on. The majority of mnemonics serious games created for these studies are for personal use rather than collaborative. They have also remained prototypes instead of being released to the public; this has led us to attempt to create one that can be used by the public instead of being shelved afterwards. In addition, mnemonics in serious games are created by people rather than computers, and the people in question are the game designers rather than the players; there is little opportunity, if any, for the game to use mnemonics provided by players. This has led us to create a game that makes use of mnemonics provided by players. Moreover, mnemonics serious games have generally had a positive outcome with regard to improving users' memory, which is consistent with literature stating the benefits of both mnemonics and serious games, as described in previous sections. They are also generally found enjoyable, but not particularly memorable, which has led us to create a game that would be found memorable.

2.2 Systematic Review of Apps

We [96] had also searched the App Store and Google Play for apps that are mnemonics serious games. We found a total of 32 relevant apps. We saw that these apps started to appear in the 2010's and became more prevalent in the early 2020's. They have a wide range of subjects, consistent with what we stated in the Introduction. They tend to be puzzles or quizzes, but sometimes take a less-specific form that cannot easily be classified. The most common mnemonics used are acronyms and images, which we intend to make use of too. Some have a formal educational method while others are more casual, but even the former are not specifically meant to be used alongside academia and did not seem to be created by researchers; as researchers ourselves, we decided to create one that is meant for academia. The apps that were rated by players tend to have at least 3 stars out of 5, with quite a few having at least 4 stars. However, most of the apps fail to live up to the characteristics listed by Blumberg et al. [1], as well as the broad themes by Morris [2]. We dubbed the former IN-3PACIFIC (**I**mmersion, **N**arrative, 3 **P**resences, **A**gency or **C**ontrol, **I**dentity, **F**eedback, **I**nteractivity, **C**hallenge) and the latter C-SHARP (**C**ritical Reflection, **S**ituated in Context, **H**ands-on **A**ctivity and Participation, **R**isk/Novel Experience, **P**ragmatic/Purposeful). The former are further described in Table 1 while the latter are further described in Table 2. The only ones the majority of

Table 1: IN-3PACIFIC Characteristics [1]

Characteristic	Description
Immersion	Player's feeling present in the game
Narrative	A sense of story progression
Presence (Self)	Experience of the player's persona in the game as their actual self
Presence (Social)	Allowing players to interact with virtual actors in the game as though they were real
Presence (Spatial)	Players feeling physically embodied in the game's environment
Agency/ **C**ontrol	Allowing players to control aspects of the game including tailoring it to their preferences
Identity	Allowing players to create characters or profiles representing them
Feedback	Information that players receive from the game relating to their performance, progress, rewards, what actions they can take and the consequences of their choices
Interactivity	Allowing players to initiate and receive feedback for actions in the game, as if the game is an entity that can be communicated with
Challenge	The various degrees of complexity that are possible in a game based on the player's actions

the apps fulfill are Feedback and Interactivity (they allow players to initiate and receive feedback), Challenge (they provide a range of difficulties), and Agency or Control (they allow players to tailor the game to suit themselves). Fulfilling the majority of these characteristics became a priority, though we understood it could be rather challenging or even counter-intuitive to fulfill all or most of them.

Table 2: C-SHARP Themes and Their Characteristics [2]

Theme	Characteristic	Description
Critical Reflection	Investigation-Like	Allowing the player to act as an investigator who knows that there is room for improvement and refinement of solutions to the posed problem in the light of new knowledge
	Multisolution-Based	Providing room to solve problems in multiple ways due to their open-structured nature and context specificity
Situated in Context	Place	The setting of the game including its location, community, culture, etc.
	Time	Depicting a time period (past, present or future) in the game setting
Hands-on **A**ctivity/Participation	Learner Empowerment	Support of players' agency or control over aspects of the game including tailoring it to their preferences
	Collaboration With Others	Player being able to work with other players, e.g., through cooperation or competition
Risk/Novel Problems	Novel Experience	Throwing player into new and/or unfamiliar situations
	Challenge	The various degrees of complexity that are possible in the game based on the player's actions
	Spontaneity/Unpredictability	Periodical introduction of unexpected elements that the player must quickly adapt to
Pragmatic/Purposeful	Inquiry-Based	Allowing and encouraging player to ask themselves questions or create their own problems to solve
	Co-Construction of Meaning	Allowing players to cooperate, build knowledge collaboratively, and learn from each other

3 Method

3.1 Framework

3.1.1 Framework Overview

Oyibo [6] proposed the research methodology shown in Figure 2, which he adapted from the Design Thinking framework [102], for his toolkit. There are 7 stages for each toolkit component, including the serious game. First, a survey is to be conducted to gather user needs (he had considered interviews too, but decided sticking to a survey for convenience and wider reach). Second, the collected data is to be analyzed and used to find what problems need to be solved; this process involves creating personas and scenarios describing use cases [103] [104]. Third, brainstorming solutions to the aforementioned problem is required. Fourth, the best solutions are to be prototyped. Fifth, the prototypes are to be tested among the target users using cognitive walkthrough [105] and user experience scales [106] [107] [108] to determine which one is preferred or the most effective. Sixth, the most effective solutions are to be implemented using standard practices of software engineering. For example, we might use a version control system such as GitHub, and Agile methodology such as SCRUM [109]. Finally, the implementation is to be evaluated among the target users using usability testing and experimental methods.

3.1.2 First Stage

The first stage of Oyibo's [6] research methodology involves conducting a survey to gather user needs. The information from these surveys is to be used to motivate the processes of the later stages. However, unexpected complications during the ethics proposal meant that Oyibo had not been able to receive his HPRC Certificate until April 25, 2025. As such, surveys could not be given to human participants until then. Since it would be unwise to simply wait for the HPRC Certificate to be granted, we had decided to move on to the other stages, as they were still doable without information from potential users. We still performed the first stage after receiving the HPRC Certificate, but information collected from the survey now served primarily to create mnemonics and build upon the game rather than serve as its initial foundation. Further details of the survey are provided in Section 3.5.

3.1.3 Second Stage

The second stage of Oyibo's [6] research methodology involves analyzing data from the first stage and using it to figure out the problems that need to be solved. However, as stated before, the first stage had been postponed because of the ethics complications. Thus, the second stage needed an alternate source of information and had to be made more informal. Instead of using data from recruited participants, I listened to suggestions from Oyibo and our colleagues during and outside of our weekly group meetings, and also followed IN-3PACIFIC [1] and C-SHARP [2] guidelines. In particular, in the systematic review that Oyibo and I [96] conducted, we found that many existing mnemonics serious games failed to make players feel present in the game, create a narrative, and promote learning together with others. It was important that my game would include these characteristics.

3.1.4 Third Stage

The third stage of Oyibo's [6] research methodology involves brainstorming solutions to the problem discussed in the second stage. I had mentioned that many existing mnemonics serious games did not provide senses of presence or narrative; I decided to allow the player to create an identity for themselves in the game and participate in a story using that identity. I had two ideas for a possible plot: One was that the player would be a worker in a factory that assembles word-shaped structures based on mnemonics and then ship them off for sale to help people remember things more easily. The other was that they would be a magician in a medieval fantasy setting and banish monsters, whose names are mnemonics, back to their own realm by saying their names; I based this idea on the famous tale of Rumpelstiltskin, an imp who challenges a woman to guess his name within three days or else he will steal her child. Fortunately, she overhears him singing a song that features his name, she tells him his name the next day, and he either angrily flees or tears himself apart depending on the version. Oyibo, our colleagues, and myself greatly preferred the medieval fantasy over the factory, so I immediately went with the former. Regarding challenging and motivating one another to learn, I implemented a leaderboard where players see each other's total points obtained and feel motivated to keep playing. Further details of the game are provided in Section 3.4.

3.1.5 Fourth Stage

The fourth stage of Oyibo's [6] research methodology involves prototyping the best solutions among those in the third stage. I chose to use Figma because of my experience with it in the past. I would create frames for the various screens I planned to have in my game. If a screen would have changing states, such as an idle screen, a correct-answer screen, and a wrong-answer screen in a level, I would have a frame for each screen. I first created a low-fidelity version, which only had the basic layouts, text, and shapes, then upon approval, I created a high-fidelity version, which added colours, images, and reusable components. The high-fidelity version also included more screen state variations and multiple navigations to closely simulate an actual playable game.

3.1.6 Fifth Stage

The fifth stage of Oyibo's [6] research methodology involves testing prototypes with the target users using cognitive walkthrough [105] and user experience scales [106] [107] [108] to see which one they consider the best. However, the ethics complications meant that we could not involve human participants other than Oyibo and our colleagues, as well as Timothy Ojo, a product designer and friend of Oyibo. We also made this stage more informal, just as we did for the second stage. Instead of formal cognitive walkthroughs, I would regularly show Oyibo and our colleagues my Figma prototype during and outside our weekly meetings and asked them what ideas for improvement or new features they had; I would use their comments to improve my singular prototype rather than making several prototypes.

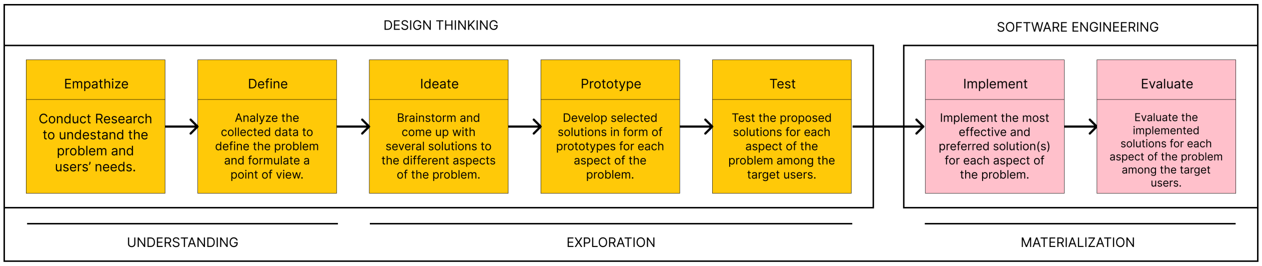


Figure 2: Stages of the research program and key methods to realize them [6].

3.1.7 Sixth Stage

The sixth stage of Oyibo's [6] research methodology involves implementing the solutions based on the prototypes using standard practices of software engineering. Though I did not end up using SCRUM [109], I did end up behaving in an Agile-like manner. As previously stated, I would participate in weekly meetings with Oyibo and his other students or research assistants. There, I would show my game (especially the changes made in the latest week) to them just as I showed my Figma prototype, and I would receive ideas for improvement or new features. I would also upload and update my code in a GitHub repository.

3.1.8 Seventh Stage

The seventh stage of Oyibo's [6] research methodology involves evaluating the implementation among the target users using usability testing and experimental methods. In my case, that would mean letting participants play my game and then comparing their performances in a test about topics that the mnemonics were for to the performances of people who did not play my game. This would be my main study. The goal was to determine whether the game's handling of mnemonics leads to greater improvements in learning and content retention than traditional methods such as focus on slides. Further details are provided in Section 3.6.

3.2 Toolkit Description

Figure 3, provided by Oyibo [6], shows the learning process, which begins in the classroom and culminates in the retrieval of learned information from long-term memory in the future. In particular, the right half shows the SANKOFA toolkit. As he explains, the term "SANKOFA" itself is an acronym which stands for "**S**ave **A**ll **N**ew **K**nowledge **O**ptimally and **F**etch **A**ccurately." It is based on the Ghanaian Akan term "Sankofa," which literally means, "to go back and fetch it" (p. 127) [110]. It relates to the Ghanian proverb, "It is not wrong to go back for that which you have forgotten" [111] [7]. As Oyibo [6] also explains, the SANKOFA Toolkit implements two main theoretical frameworks: SAVE and RADAR - both of which are derived from the 12 principles of memory [112], particularly selection, association, visualization, elaboration, review, and feedback. As such, he decided to name its components after these key memory principles; though I did not work on both of them, I will be explaining the implementation of both frameworks for readers' convenience and information. First, there is the SAVE (**S**elect, **A**ssociate, **V**isualize and Verbalize, and **E**laborate) tool. This is the mnemonics creation tool; it includes a training module that teaches students and instructors how to create effective mnemonics that can help them "go back and fetch forgotten knowledge" [113]. Second, there

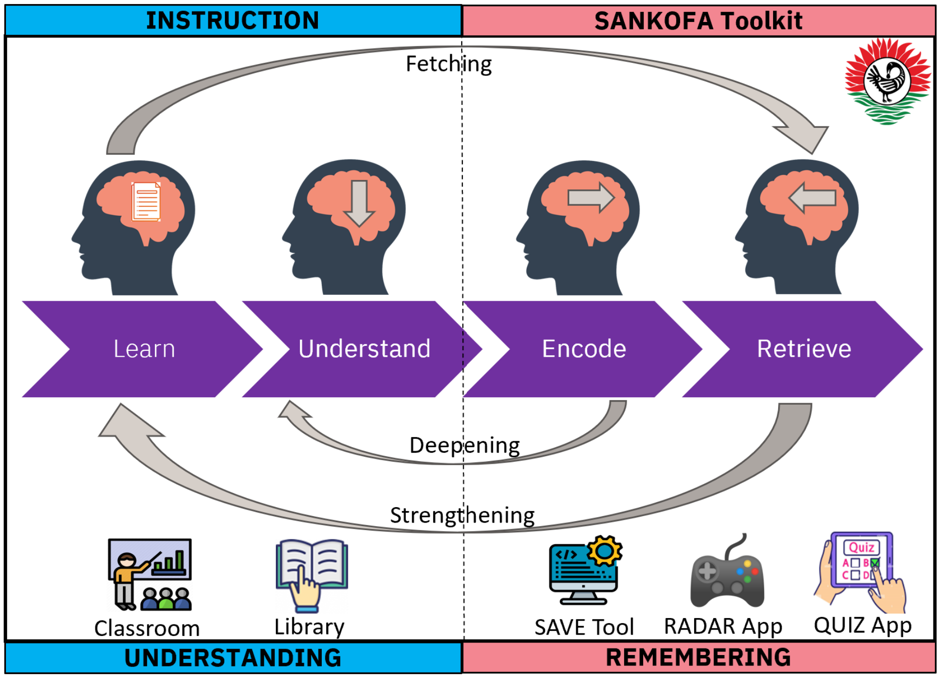


Figure 3: Long-term learning process based on the SANKOFA framework [6], with images provided by [7] [8] [9] [10] [11] [12] [13].

is the 5R (**R**etrieve, **R**eceive feedback, **R**eview, **R**eflect, and **R**einforce) system, which metamorphosed into the RADAR (**R**ecollection, **A**ssociation, **D**ecoding, and **A**rtifacts **R**eview) framework (to be discussed later) which I implemented in the form of a mnemonics serious game, and the quiz notification system which has not been implemented yet. The SAVE, RADAR, and quiz notification systems will help both students and instructors better encode content using mainly acronyms, acrostics, and images, and fetch said content when they need to. Oyibo [6] cautioned that the SANKOFA toolkit is meant to serve as a supplement in courses, not a replacement for methods that are already in use. As such, students must first learn the course content regularly, and then have the toolkit assist them in reinforcing the acquired knowledge; likewise regarding instructors teaching the content regularly. In other words, both instructors and students must still fulfill their regular duties as if the toolkit were not there.

3.2.1 SAVE Tool

As Oyibo [6] explains, the SAVE Tool involves **S**electing information such as a presentation slide to encode into long-term memory using a mnemonic, **A**ssociating this information with one's existing knowledge and experiences, **V**isualizing and Verbalizing the association, and **E**laborating on the information to be encoded, hence "**SAVE**." An example he provides involves reordering the components of web accessibility into a meaningful mnemonic, "**ADA CUBE**" (**A**uthoring tool, **D**evelopers, **A**ssistive Technology, **C**ontent, **U**sers, **B**rowser, and **E**valuation Tools), as shown in Figure 4. By reordering the components of Web Accessibility and using the first letters of each word to form a meaningful mnemonic phrase "ADA CUBE" (through elaborative rehearsal), the first word of which references Ada Lovelace, the first computer programmer,

Figure 4: Example mnemonic cue [6] [14].

learners will be able to recollect the target information easily and fast in the future. The SAVE Tool will help encode this list through the elaboration process, giving its user an advantage at remembering those components of web accessibility early on against someone who relies on rote memorization. As Oyibo [6] describes, this particular encoding method will enhance episodic memory, complementing semantic memory; he even refers to it as mental time travel, or the ability to mentally place oneself in the past or future [48]. The SAVE Tool also supports low-order, middle-order, and high-order thinking in Bloom's taxonomy [5], through remembering, analyzing via elaboration and association, and creating meaningful mnemonics respectively. The SAVE Tool was not developed by me, but rather by Sarah Ejaz [4], a fellow student of Oyibo. However, the SAVE Tool is still a key part of my study.

3.2.2 RADAR Game

The RADAR game emphasizes retrieval practice by providing chances for reviewing encoded information and reflecting on feedback. Oyibo [6] asserts that the type of active engagement provided by a game can help foster episodic memory and thus strengthen long-term memory of the target material [114] [115] [116]. As stated by Oyibo [6], the game follows the RADAR framework (**R**ecollection, **A**ssociation, **D**ecoding, **A**rtifacts **R**eview) for designing mnemonics serious games that deal with acronyms and acrostics [6]. The **R**ecollection component tests players' ability to remember the mnemonics themselves. **A**ssociation tests their ability to link mnemonics to their associated topics, called target topics. **D**ecoding tests their ability to remember what mnemonics stand for, i.e., the target information. **A**rtifacts **R**eview allows players to view the artifacts (details) of their mnemonics, which they had provided in the SAVE Tool. The word "Artifacts" was originally chosen because it means "tools," which mnemonics certainly are, and it helps contribute to the nice-sounding acronym "RADAR." There are two possible approaches to the "RAD" in "RADAR": One is to focus on one mnemonic at a time and recollect, associate, and decode it back-to-back before reviewing it immediately afterwards. This would be treated as one short level, and there would be many such levels depending on the number of mnemonics. We will call this approach "One-mnemonic-at-a-time" (One-MAT). The other approach is to recollect all the mnemonics in a given module all at once (one at a time), and then associate, decode and review them. In other words, the player attends to all of the mnemonics in a given module in each stage before moving to the next stage. There would be fewer levels, but they would be much longer, depending on the number of mnemonics. We will call this approach “All-mnemonics-at-a-time” (All-MAT). I chose to go with All-MAT as it was more appropriate to my game’s story; I will explain in further detail later. Either way, **R**ecollection must be done before **A**ssociation, which must be done before **D**ecoding, which must be done before **A**rtifacts **R**eview.

Recollection, **A**ssociation, and **D**ecoding each have five different cognitive tasks to choose from; these tasks determine how players are supposed to provide answers. Visual Recognition involves choosing the correct image among several, while Verbal Recognition involves choosing the correct text among several; these two are considered low-difficulty due to the correct answer being in plain sight. Visual Cued Recall involves typing the correct text associated with a given image that is only partly shown, while Verbal Cued Recall involves typing the correct text when given part of it; these two are considered medium-difficulty due to only a bit of aid in coming up with the correct answer. Verbal Free Recall involves typing the correct text without aid; this is considered high-difficulty. There is an idea for another high-difficulty cognitive task, Visual Free Recall, but that would involve drawing the correct image and not only would drawing take a long time, but current technologies likely cannot analyze the correctness of a drawing in a short time to avoid making the player wait.

3.2.3 Quiz Notification System

As for the quiz notification system (which is currently not implemented), Oyibo recommends to “send spaced-out tests (one question at a time) on encoded information (mnemonics and their meanings) at user-preset intervals on a daily basis.” [6] The SAVE Tool, RADAR game, and quiz notification system are based on key memory principles and will help cement users’ memory of what they learned [117] [118].

3.3 Features of the SAVE Tool

In this section, I will explain the features of the SAVE Tool [4] in further detail. In particular, I will focus on logging in, creating a mnemonic, and viewing a mnemonic.

3.3.1 Login

As part of the SANKOFA Toolkit, the SAVE Tool requires a SANKOFA account to use it [4]. From the “Welcome to SANKOFA” page of the SANKOFA website, the user must click the “Login” button in the top right corner. This takes them to the “Log In” page, where they must enter the email address and password associated with their account. For the sake of the experiment, we provided accounts to participants. In the future, however, we plan to include a feature that allows signing up for SANKOFA.

3.3.2 Creating and Viewing a Mnemonic

After logging in, the user is taken back to the “Welcome to SANKOFA” page, but now sees their name near the top right corner, as well as a “SAVE Tool” icon near the top middle [4]. To create a mnemonic, the user must first click the “SAVE Tool” icon. They then need to choose a course, which takes them to the “My

Toolbox" page, showing the modules (which are lecture topics of a course) of the chosen course. From here, they must click "Create Mnemonic," choose one of the course's modules, and click "Next." Doing so takes them to the "Workbench."

The user's first action in the "Workbench" is to **S**elect what target information (the main content that needs to be remembered) to encode. They must upload an image of a presentation slide that contains the information, then they must provide the target topic (a topic that is more specific than the module and that the mnemonic is meant to help the user for), as well as the target information itself (which is also what the mnemonic stands for) [4].

Next, the user must **A**ssociate the target information with existing knowledge and **V**isualize (as well as verbalize) the association. They must select the type of mnemonic and upload an image to represent the mnemonic [4]. Then, for each mnemonic term, they must type it, its target term, and upload an image for the target term. As they do this, the entirety of the mnemonic will be shown in a blue bar. Whenever they need more rows (one for each mnemonic term), they can simply click "Add Row."

Afterwards, the user must **E**laborate on the target information. They must first provide a verbal cue (a conceptual connection between the mnemonic and the target information) [4]. Then they must enter tidbits, which are additional pieces of information that further explain the target topic or information. If they need to enter more tidbits, they can simply click "Add Tidbit."

When finished, the user can click "Save" and receive an alert telling them their mnemonic has been created [4]. Clicking "OK" shows a page confirming that the mnemonic has been saved, and the user can click "Go Back to My Toolbox." Once back at the toolbox, all they need to do to view their mnemonic is click the module to reveal its list of existing mnemonics, and then click the mnemonic to reveal all of its artifacts.

3.4 Acronimps, My RADAR Game

The particular RADAR game I have created is called "Acronimps," which is an obvious pun on "acronyms" and "imps," but also "acrostics" (which the game accommodates) and "nimps" (a term often associated with smallness or minimality, fitting for mnemonics used to encapsulate large amount of information in a concise and compact way). In the Acronimps game, the player takes the role of a magician who banishes the titular monsters invading the player's land back to their realm by saying their names, i.e., the acronyms and/or acrostics themselves that the player has created and stored in memory using the SAVE Tool.

3.4.1 Setup

First, the player must view the "Disclaimer" page, which asserts that the mnemonics and images shown do not reflect personal opinions. Clicking the right arrow takes them to the "Log In" page, in which they must log in with the email address and password of their SANKOFA account. Then, in the "Course" page, they must look at a list of courses with mnemonics and choose which course's mnemonics to practice. Next, in the "Module" page, they must look at a list of the chosen course's modules and how many mnemonics each module has. They must select which module's mnemonics they wish to test themselves on (they must choose a module with at least five mnemonics because having fewer than five would make the Quests too short and

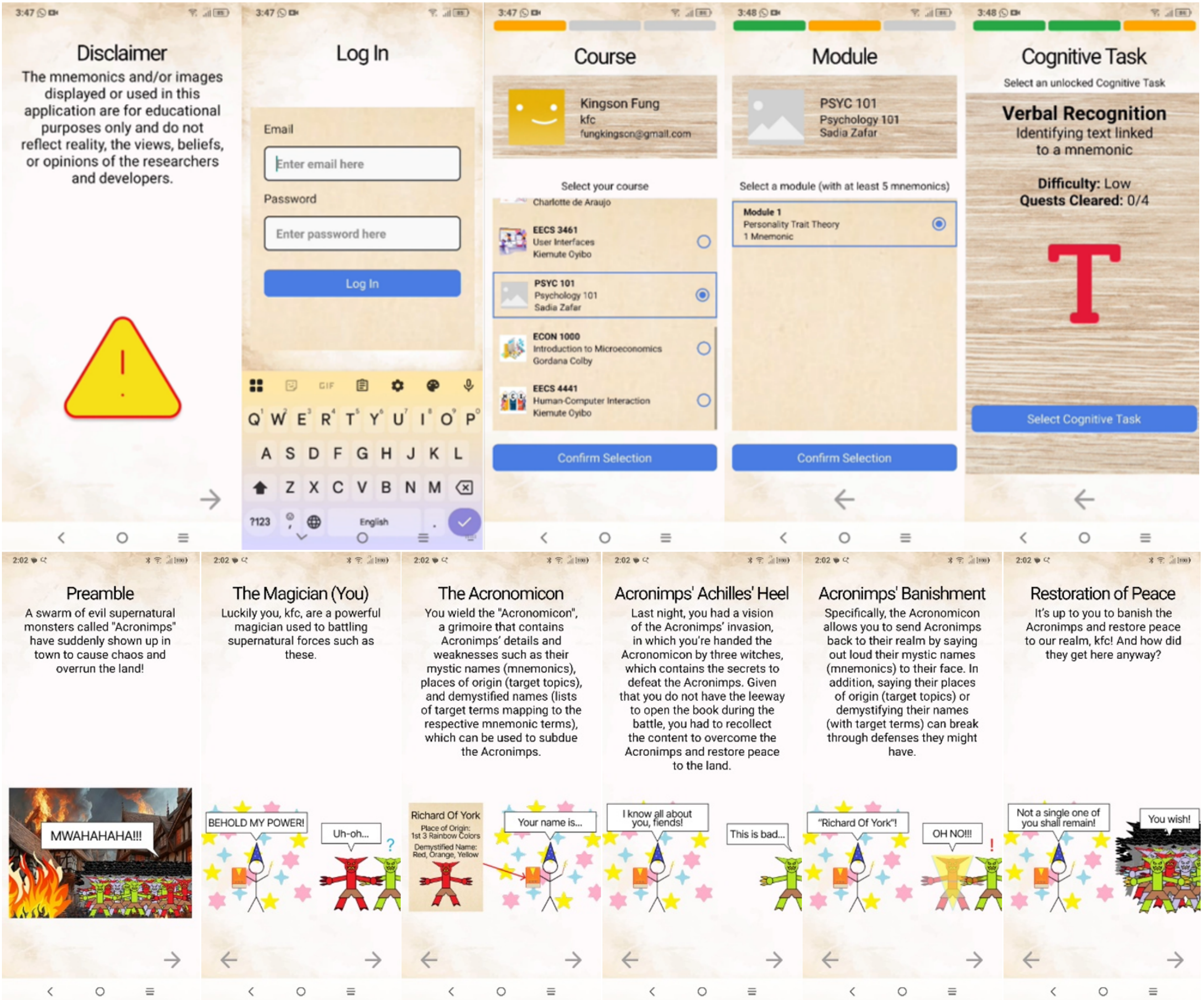


Figure 5: Screenshots of the setup and story pages in the Acronimps game, with textures provided by [15] [16] [17], avatar provided by [18], course images provided by [19] [20] [21] [22] [23], and the fire image provided by [24]. Licensed under CC BY-NC-ND.

easy). Afterwards, in the "Cognitive Tasks" page, they must look at a description of each cognitive task (Verbal Recognition is currently the only implemented one) and select one for practice. Figure 5 shows an example of the setup process. After selecting a cognitive task, the player is shown the story, which is split into several pages. The story is that evil monsters called Acronimps are attacking the town, and a powerful magician must use a spellbook called the Acronomicon to banish them back to their realm by saying their names (mnemonics) to them, while also finding out how they arrived in the first place. Clicking arrows takes the player to the next or previous page. After seeing all pages, the player arrives at the "Main Menu." This screen shows levels that the player must complete for the chosen cognitive task; they are my implementation of "RADAR" and are called Quests: "Town (Recollection)," "Bridge (Association)," "Castle (Decoding)," and "Museum (Artifacts Review)." "Main Menu" also contains icons for "Quest Statistics and Settings," "Class Ranking," "Image Attribution," and "Profile," all of which have home buttons taking the player back to "Main Menu."

3.4.2 Quests

As previously stated, the four Quests, “Town (Recollection),” “Bridge (Association),” “Castle (Decoding),” and “Museum (Artifacts Review),” are my implementations of “RADAR” and the main levels of the game that need to be completed. Choosing a Quest for the first time shows a page explaining the story of the Quest, which then leads to a preview page that includes details on the type of Acronimp that will be faced, instructions for what to do in the Quest itself, a button that allows users to “Change Quest Settings,” a button to “Go to Quest,” a back button to view the story page again, and a home button leading to the “Main Menu.” The exception is “Museum (Artifacts Review),” in which the story page leads straight to the Quest, as Acronimps are not involved and there is no concept of winning or losing. Choosing a Quest from the second time onward skips the story page. In all Quests other than “Museum (Artifacts Review),” the player is being timed and sees so; they also see how many Acronimps are left to defeat, as well as the player’s own HP (a wrong answer decreases HP by 1, and reaching 0 HP results in a loss and requires restarting the Quest). The player can pause the Quest and view a page similar to the preview page, but with buttons to either “Quit Quest” or “Resume Quest” instead. Winning a Quest shows a page with a congratulatory message, number of XP (points used in the game) received, stars received (up to 5, with 1 less for each mistake), buttons to either go to “Main Menu” or “Go to Next Quest,” and a home button. The exception is “Castle (Decoding),” whose win page replaces the aforementioned buttons with a “View Aftermath” button that leads to a page describing the aftermath of the story and showing buttons to “Replay Castle” or “Go to Review,” and a home button. Losing a Quest shows a page that motivates the player to not give up and shows buttons to go to “Quest Settings” or “Replay Quest,” and a home button. I did not force players to remain in their current Quest until they completed it or to take the next Quest immediately after completing it, as they might wish to take a look at other parts of the game that are available to them or practice earlier Quests.

3.4.3 Town (Recollection)

“Town (Recollection)” is the first Quest to be completed. As part of the story, the magician encounters Acronimps burning a town. These Acronimps are the weakest, as they have no special defenses. They attack one-by-one, and the player sees four possibilities for the current Acronimp’s name by automatically using a type of so-called Acrovision spell. One is the correct answer, and the other three replace one or two words in the correct answer with synonyms. Choosing the correct answer instantly defeats the Acronimp and makes it yell out in rage; it also grants a maximum of 30 XP (10 less for each consecutive wrong answer with a minimum of 0), which are points that contribute to the leaderboard. If an Acronimp is defeated and it is not the last one, another one rushes into the screen to replace it. A wrong answer leads to the Acronimp striking the player and turning the screen red, along with taunting the player. Figure 6 shows progression through the Quest and how to get to it from “Main Menu.”

3.4.4 Bridge (Association)

“Bridge (Association)” is the second Quest to be completed in a chosen cognitive task. “Town (Recollection)” must have already been cleared. As part of the story, the magician encounters Acronimps on a bridge

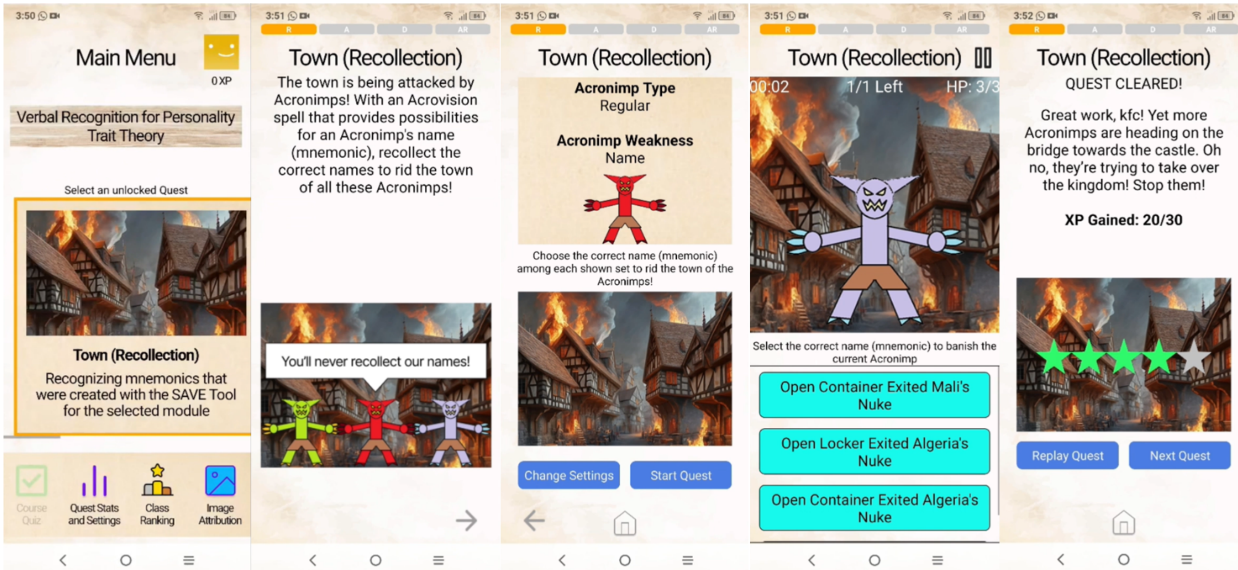


Figure 6: Screenshots of "Main Menu" and "Town (Recollection)," with textures provided by [15] [16] [17], avatar provided by [18], location background provided by [25], and bottom icons provided by [26] [27] [28]. Licensed under CC BY-NC-ND.

leading to the castle. Similarly to the previous Quest, the player confronts one Acronimp at a time, but the Acronimp is protected by a shield that renders it immune to the effects of the player saying its name. The shield must be broken by saying the place of origin (target topic) that corresponds with the Acronimp's name. Fortunately, the player automatically uses an Acrovision spell to see the current Acronimp's name, as well as places of origin for all Acronimps still in the area (meaning the list gets shorter when fewer Acronimps are remaining), and the player only needs to choose the one for the Acronimp in the screen. Once the shield is broken from a correct answer, the Acronimp is defeated and yells out in rage (since the player sees its name), and the player receives a maximum of 30 XP (10 less for each consecutive wrong answer with a minimum of 0). If an Acronimp is defeated and it is not the last one, another one rushes into the screen to replace it. A wrong answer leads to the Acronimp striking the player and turning the screen red, along with taunting the player. Figure 7 shows progression through the Quest.

3.4.5 Castle (Decoding)

"Castle (Decoding)" is the final Quest that involves Acronimps. "Bridge (Association)" must have already been cleared. As part of the story, the magician approaches the castle and encounters Elite Acronimps, whose banishment hinges on demystifying their names (by providing the correct target information) instead of simply saying them. The player confronts one Elite Acronimp at a time, and automatically uses an Acrovision spell to see the current Elite Acronimp's name with one of its parts in red; the spell also shows four possibilities for the red part's demystification (target term), with one being the correct answer while the other three start off with the same first letter and likely a few more, as well as being of the same part of speech (e.g., noun). Choosing the correct answer either makes the Elite Acronimp groan in pain and makes the next part of its name be the red one (if the current red part is not the last part), or it defeats the Elite

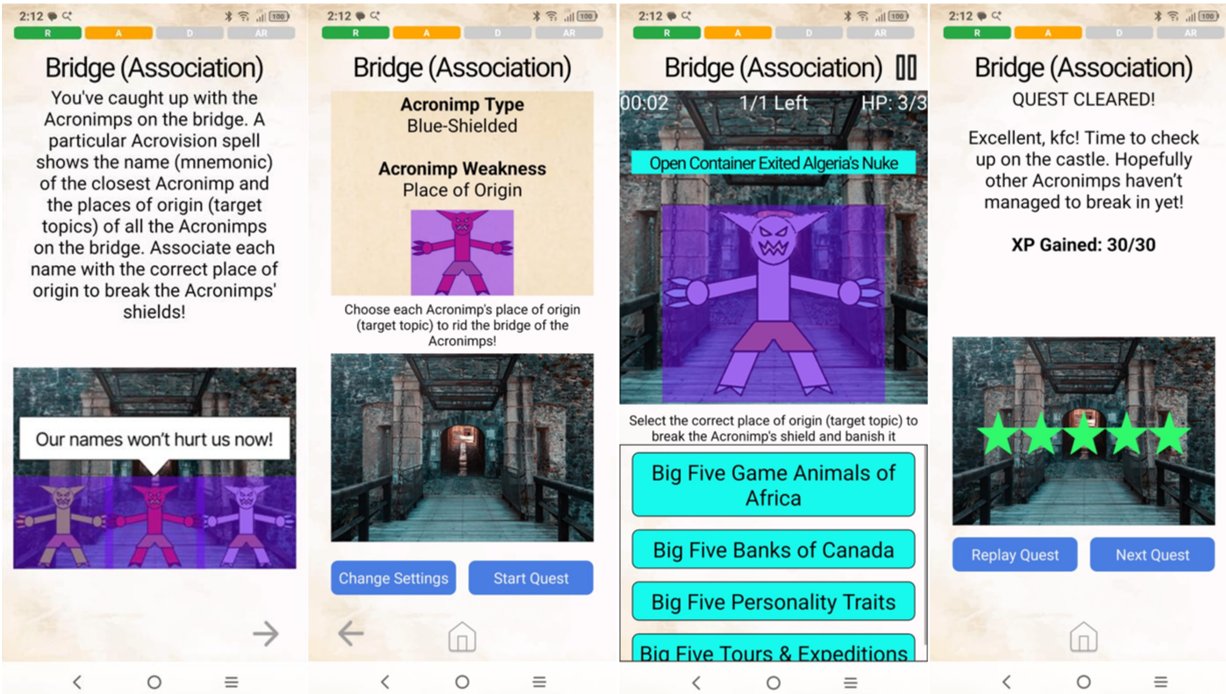


Figure 7: Screenshots of "Bridge (Association)," with textures provided by [15] [16] and location background provided by [29]. Licensed under CC BY-NC-ND.

Acronimp and makes it yell out in rage (if the current red part is the last part); either way, the player also receives a maximum of 30 XP for each part (10 less for each consecutive wrong answer with a minimum of 0), meaning an Elite Acronimp can give more than 30 XP. It is important to note that the number of Elite Acronimps left is shown (just like how other Quests show the number of regular Acronimps left), not the number of parts left. If an Elite Acronimp is defeated and it is not the last one, another one rushes into the screen to replace it. A wrong answer at any part of a name leads to the Elite Acronimp striking the player and turning the screen red, along with taunting the player. Figure 8 shows the progression of the Quest and "Aftermath."

3.4.6 Museum (Artifacts Review)

"Museum (Artifacts Review)" is the final Quest. As part of the story, the magician is rewarded with an invitation from the king to attend the grand opening of the Acronimps museum, which commemorates the monsters' defeat. The player is able to check all their mnemonics in the chosen module and their associated artifacts. The mnemonics are arranged horizontally in alphabetical order, and the player can scroll to see those that are not currently in the screen. For each mnemonic, the player can scroll vertically to view the mnemonic itself (and its visual cue), target information, target topic (and its visual cue), verbal cue, and tidbits. Figure 9 shows the various artifacts that can be seen in the Quest.

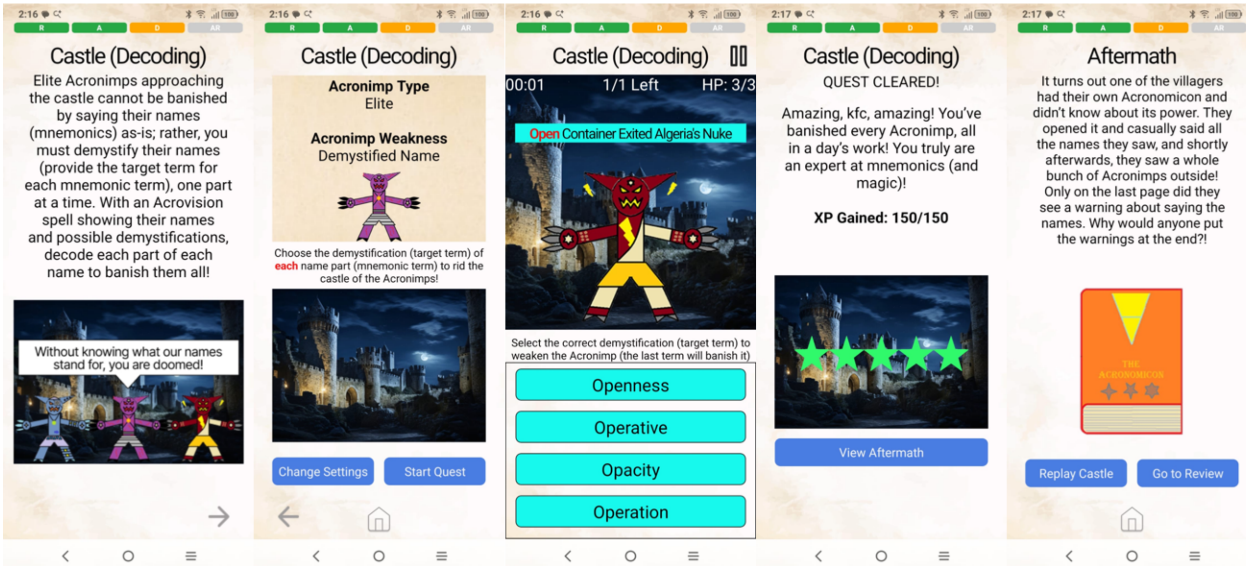


Figure 8: Screenshots of "Castle (Decoding)" and "Aftermath," with textures provided by [15] [16] and location background provided by [30]. Licensed under CC BY-NC-ND.

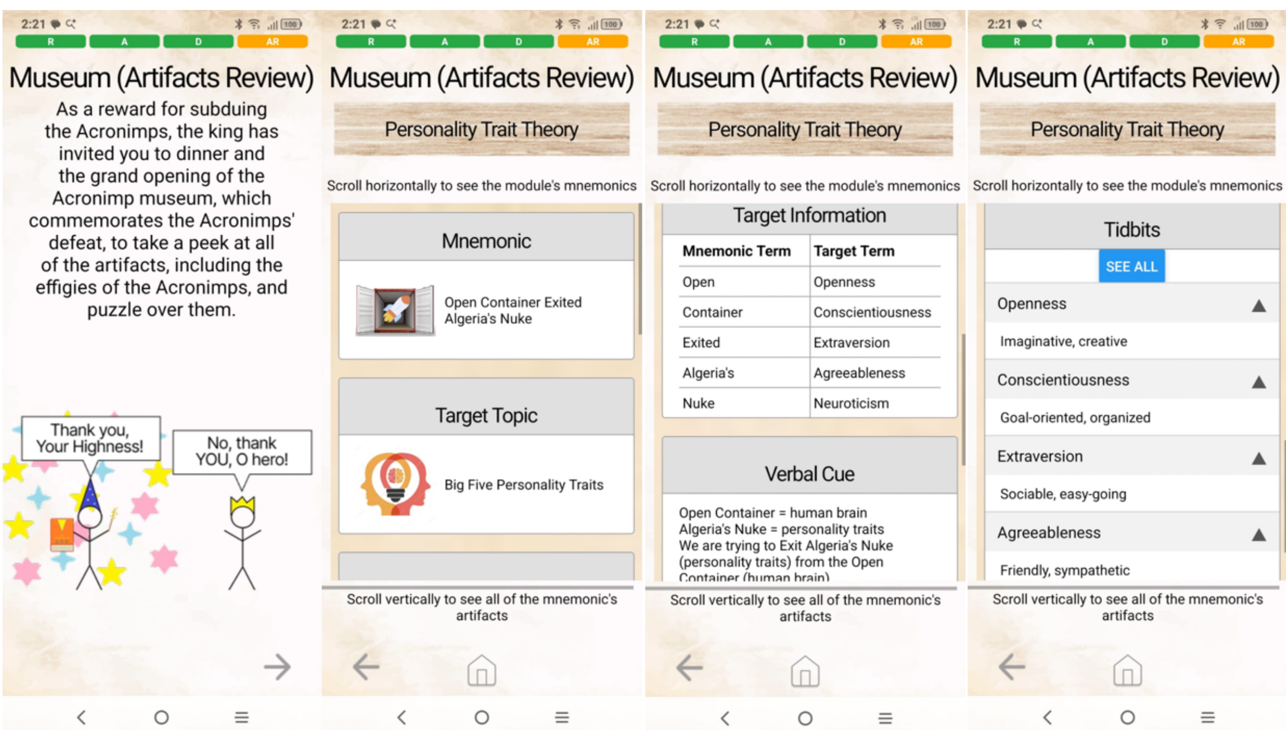


Figure 9: Screenshots of "Museum (Artifacts Review)," with textures provided by [15] [16] [17] and images provided by [31] [32] [33]. Licensed under CC BY-NC-ND.

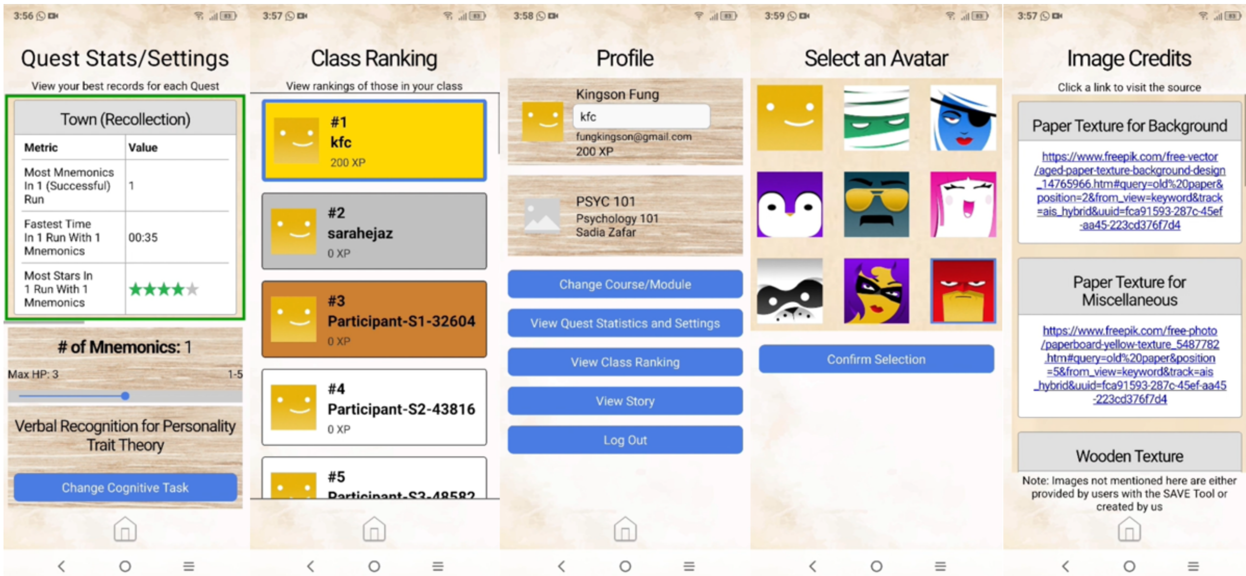


Figure 10: Screenshots of the additional features, with textures provided by [15] [16] [17], avatars provided by [18], and course image provided by [21]. Licensed under CC BY-NC-ND.

3.4.7 Quest Statistics and Settings

The "Quest Statistics and Settings" screen allows the player to see performance records for the Quests, set their maximum HP, and change what cognitive task to practice (though as stated before, the only cognitive task that has been implemented is Verbal Recognition). More specifically, for each Quest (other than "Museum (Artifacts Review)") in the current cognitive task, the player is shown the highest number of mnemonics they used in a successful attempt, the shortest amount of time they took with that many mnemonics, and the highest number of stars they received with that many mnemonics. Below, there is also a section that shows the number of mnemonics the player is currently using, and the section contains a slider that allows the player to set their maximum HP from 1 to 5 inclusive; the default is 3, but daredevils can choose as low as 1 to enjoy the thrill of not leaving any room for error while those who might not be as confident can choose up to 5 to give themselves some breathing room. Another section tells the player the cognitive task they are currently practicing, and it provides a "Change Cognitive Task" button if they wish to switch to a different one; this takes them back to the "Select Cognitive Task" screen. Figure 10 shows the "Quest Statistics and Settings" page, as well as those for the pages explained below.

3.4.8 Class Ranking

The "Class Ranking" screen shows the leaderboard for the current class. To be precise, it shows a list of players who enrolled in the current class; each such player's profile picture, username, and XP are shown (we avoided showing their full name to maintain some degree of anonymity). Players are ranked by XP for the class, with those having more XP for the class being higher up on the list. It should be noted that a certain player can be enrolled in more than one class and have different XP for each such class, meaning they can outrank another player in one class and be outranked by that same player in another class. The leaderboard is only about the class that the player is currently practicing mnemonics for.

3.4.9 Image Credits

Because many of the images I used for the backgrounds, avatars, and icons were taken from online sources, I included an “Image Credits” screen. In it, I stated each image that is not created by me or provided by a user of the SAVE Tool (i.e., the mnemonic visual cues and target topic visual cues), and I included a link that takes the player to where I got the image from if they click on it.

3.4.10 Profile

The “Profile” screen shows the player’s avatar, full name, username, email, and XP for the class they are currently practicing mnemonics for. The class’s course code, name, and instructor are also shown. The username is directly editable in this page, while the avatar can be changed by clicking it and arriving at a “Select an Avatar” screen, which allows them to choose from one of nine avatars. There are also five buttons below. “Change Course/Module” takes the player back to the “Course” screen, and they have to go through the “Module” screen immediately afterwards, just as they did in the setup. “View Mnemonic Details” and “View Quest Statistics and Settings” takes them to the “Mnemonic Details” and “Quest Statistics and Settings” pages respectively. “View Story” lets them view the initial story pages once more. “Log Out” logs the player out and takes them back to the “Log In” screen.

3.4.11 Acronimps Game’s Implementation of IN-3PACIFIC and C-SHARP

It is common practice to assess a game against a specific set of requirements [119]. As stated in Section 2.2, my Acronimps game takes IN-3PACIFIC [1] and C-SHARP [2] into account. Aside from Multisolution-Based and Co-Construction of Meaning, the Acronimps game implements all the remaining characteristics, as shown in Tables 3 and 4. The latter also explains how Multisolution-Based and Co-Construction of Meaning could be implemented in the future.

3.5 User Requirements Survey

As previously stated, the first stage of Figure 2 involves a survey, namely a user requirements survey. In this particular case, Oyibo [6] instructed that such a survey in YorkU is to be used to gather the requirements data for the mnemonics serious game. As previously mentioned however, because of ethics complications, the survey needed to be postponed, so its information served to create mnemonics and build upon the game rather than serving as its foundation. Regardless, the survey was meant to investigate learning strategies currently used in memory-intensive courses, technology-driven tools, mnemonic devices’ potential, and how willing students are to use all of these.

3.5.1 Apparatus

The survey is online, on Google Forms. Data from the participants is saved on Google Sheets. Because of its online nature, the survey was presumably done on various types of devices. Monetary compensation was

Table 3: Acronimps Game's Implementation of IN-3PACIFIC [1].

Characteristic	Implementation
Immersion	First-person view leads the player to feel immersed in the game itself
Narrative	A clear story, as well as progression through a fixed order of Quests
Presence (Self)	The player acts as the magician confronting the Acronimps
Presence (Social)	Confronting Acronimps, which are separate entities
Presence (Spatial)	The Quests' different locations provide a sense of travelling through the world
Agency/ **C**ontrol	The player can select which course and module's mnemonics will be used, as well as set their HP from 1 to 5
Identity	The player can set their username and avatar, with the former even being incorporated into the story
Feedback	The game clearly tells the player whether they are right or wrong, tells them if they have won or lost, and provides tidbits as feedback during "Castle (Decoding)"
Interactivity	The game is essentially teaching the player how to combat Acronimps and guides the player on where to go next
Challenge	"Castle (Decoding)" is more complex than the previous Quests, with very taxing memory tests and more actions required to take down an Elite Acronimp than a regular one

provided through Interac e-Transfer. R was used to analyze the data.

3.5.2 Participants

In the SANKOFA project proposal, Oyibo [6] stated that the survey would be distributed to $n = 100$ students. He explained that "the survey sample size was calculated using $n = z^2(\hat{p}(1-\hat{p}))/\varepsilon^2$ [120], where z is the z-score (1.96), ε is the margin of error (8%), and $\hat{p}$ is the estimated YorkU student population proportion using cognitive enhancers in memory-intensive courses (0.22). The estimated value of 0.22 was based on the finding that approximately 22% of university students used cognitive enhancers to augment cognitive performances while studying [44]. The calculation ($n = 1.962(0.22)(0.78)/0.082$) gives $n \approx 100$" [6]. We received a total of 122 responses from participants, who were each offered \$10 for the time spent on taking the survey.

3.5.3 Procedure

The survey was disseminated to students of any field through the graduate students email mailing list, while posters were put up on campus to get the attention of additional students. Consent was obtained from each participant in the survey itself before they could move on to the rest of the survey. After enough participants had done the survey, the survey was then closed and participants were then given the promised \$10. Though we intended to analyze the data, we could only analyze a portion of it, given the time constraint imposed by the delayed ethics approval and the urgency of the other parts of the study. We analyzed the items shown in Table 5 [3], with regards to comparisons between students who preferred to create their own mnemonics and students who preferred mnemonics given by professors. For each of the items, students expressed agreement

Table 4: Acronimps Game's Implementation of C-SHARP [2].

Theme	Characteristic	Implementation
Critical Reflection	Investigation-Like	Acrovision spells' abilities essentially serve as clues; the player can also examine and reflect on artifacts during "Museum (Artifacts Review)"
	Multisolution-Based	Not currently, but incorporating the remaining cognitive tasks would allow the player to respond in different ways
Situated in Context	Place	The game features typical places you would expect in a medieval fantasy setting
	Time	Medieval times, albeit heavily fictionalized to provide a fantasy setting
Hands-on **A**ctivity/Participation	Learner Empowerment	The player can select which course and module's mnemonics will be used, as well as set their HP from 1 to 5
	Collaboration With Others	The ranking system demonstrates this in the form of indirect competition
Risk/Novel Problems	Novel Experience	Continuously using different words as attacks can be considered a novel concept
	Challenge	"Castle (Decoding)" is more complex than the previous Quests, with very taxing memory tests and more actions required to take down an Elite Acronimp than a regular one
	Spontaneity/Unpredictability	Each of the first three Quests provides a new twist compared to the previous, and the order of questions and answers is random; the revelation about a villager accidentally summoning Acronimps by saying their names just before "Museum (Artifacts Review)" is also a surprise twist in the story
Pragmatic/Purposeful	Inquiry-Based	The player essentially needs to answer questions based on content they entered into the SAVE Tool
	Co-Construction of Meaning	Not currently, but including a chat function and a mnemonic sharing function would achieve this

Table 5: User requirements survey items (* indicates that the item was reversed during analysis) [3].

Item Name	Item Statement
Ease of Creation	I find it easy to create helpful mnemonics on my own.
Challenge of Creation	I find it challenging to create useful mnemonics on my own.*
Study	I find mnemonics useful in my study.
Learning	I find mnemonics helpful in my learning.
Retention	My use of mnemonics helps me retain learned content long term.
Recall	My use of mnemonics helps me recall learned information.
Knowledge	My use of mnemonics enhances my knowledge.
Passing	My use of mnemonics helps me pass memorization-intensive courses.
Creation Tool	If a professor in one of my memory-intensive courses provides an interactive tool that makes it easy for me to create mnemonics that enhance my learning and recall, I will use the tool.
Retrieval Practice Tool	If a professor in one of my memory-intensive courses provides an educational game that helps me rehearse retrieval of the target information of the mnemonics that I have created, I will use (play) the game to enhance my learning and recall.

on a 7-point Likert Scale, with 1 for "Strongly Disagree" and 7 for "Strongly Agree."

In addition, we also conducted qualitative analysis on how users of mnemonics tools would like for them to be. We asked participants "What feature(s) would you like the tool that helps you create your own mnemonics to have?", referring to the SAVE Tool, and "What feature(s) would you like the educational game that allows you to practice retrieval of the target information using mnemonics to have?", referring to the RADAR game. Both can also involve feelings about mnemonics in general. Looking at the responses, we created various categories and counted the responses that referred to each one. In particular, we created mnemonic-based categories named Meaningfulness, Humour, and Relevance to Topic, which had a considerable number of relevant responses and were also important factors to consider when creating a memorable mnemonic. We also created RADAR-based categories named Repetition, Interactive, Progress Tracking, Straightforward, Visuals, Fun, Feedback, Matching-Based, and Rewards, which also had a considerable number of relevant responses and were important factors to consider during game creation. Sarah Ejaz and I each worked individually and then resolved disagreements, keeping track of the Cohen's Kappa for each category.

3.6 Game Evaluation

As previously stated, the seventh stage involves evaluating the solution created in the sixth stage with target users. In this particular case, my Acronimps game was tested with students in an effort to compare its success with those of rote memorization. Further details are provided below.

3.6.1 Apparatus

The implementation of my Acronimps game is mobile-based. However, the various surveys and tests are on Google Forms with response data being saved on Google Sheets, and the SAVE Tool may be accessed on various types of devices. Monetary compensation was provided through Interac e-Transfer. In addition, R was used to analyze the data.

3.6.2 Participants

A screening survey was given out to students of any field through the graduate students email mailing list, while posters were put up on campus to get the attention of additional students. Oyibo [6] stated that there were to be 2 participant groups, each with $n = 25$ participants, leading to a total of $n = 50$. However, we got 1 group with 23 participants and another group with 32; the difference in number was because some participants in the latter group appeared to heavily favour the Acronimps game over the SAVE Tool (which I will explain later), did not take the experiment seriously, or they had special needs that could not be accommodated by us. Participants were each paid $50 for the experiment.

3.6.3 Experiment Design

The design of the experiment is 2 group conditions (rote memorization vs. mnemonics with SAVE Tool and Acronimps game) X 2 tasks (recall vs. recognition) X 3 repeated memory tests (immediate vs. delayed vs. post-revision) X 6 topics (Cranial Nerves vs. Cranial Nerve Functions vs. Biological Organization Levels vs. Biological Organization Disciplines vs. Krebs Cycle Substrates vs. Krebs Cycle Reactions). Figures A1, A2, A3, and A4 in Appendix A show the exact slides we used to explain these topics to participants.

3.6.4 Procedure

Before explaining the study, I will explain the content that all participants needed to learn, as well as the mnemonics that were provided to the experimental group. Table 6 shows the mnemonics for the Krebs Cycle Substrates (KS) and Reactions (KR). For the reaction in row n, the substrate in row n undergoes it and gets converted to the substrate in row $n+1$. Table 7 shows the mnemonics for the Cranial Nerves (CN) and Functions (CF). The mnemonic for CF is adapted from the popular mnemonic “Some Say Marry Money, But My Brother Says Big Brains Matter More” in the gray literature. For the nerve in row n, its corresponding function is also in row n. Table 8 shows the mnemonics for the Biological Organization Levels (BL) and Disciplines (BD). For the level in row n, the discipline that studies it is also in row n. Figures A5, A6, and A7 show the exact slides we used to explain these mnemonics to participants. When creating the mnemonics, we had to consider a trade-off between short and simple mnemonics that were easier to memorize but had less orthographic and/or semantic similarity (i.e., similar spelling and/or meaning) [121] to the topic/information on one hand, and long and complex mnemonics that were more difficult to memorize but had more orthographic and/or semantic similarity to the target topic/information on the other hand. We decided to use the latter since they could help participants avoid mixing up which target information is for which topic. For example, “Oldfactory Optician’s Oculi Trouble Trigonometrically, Abductor’s Facial

Table 6: Mnemonics for Krebs Cycle Substrates and Reactions. Licensed under CC BY-NC-SA.

Substrate	Mnemonic	Reaction	Mnemonic
Acetyl CoA	Acetic	Condensation	Condemned
Citrate	City	Isomerization	Isometrician (exerciser)
Isocitrate	Isobel's	Oxidative decarboxylation	Oxton (name)
Ketoglutarate	Kettle gluttonously	Oxidative decarboxylation	Overloaded
Succinyl CoA	Succinctly	Phosphorylation	Pharoah's
Succinate	Sucks	Oxidation	Ox (with)
Fumarate	Fumy	Hydration	Hydraulic
Malate	Malt	Oxidation	Oxide
Oxaloacetate	Oxaloacetic		

Vestibule Glossing pharyngeally Vagina Accessory Hypoglossal" has higher orthographic similarity to the CN terms than the well-known "On Old Olympus' Towering Top, A Finn And German Viewed Some Hops," as well as higher semantic similarity (being about nerves instead of Olympus) [121]. Longer mnemonics with terms sharing several initial letters with their target terms can help participants remember the target terms themselves or least obtain more partial marks. All six of our mnemonics have plenty of terms sharing more than just the first letter of their respective target terms; some mnemonic terms are their own target terms, such as "Motor" for CF and "Cell" for BL.

We perceive the following topics as simple: BL (because many of terms have been studied in elementary school), and BD (because many of its terms are commonly known and/or relate closely to their respective BL terms, such as "Anatomy" being mentioned a lot in everyday life and "Cellular Biology" relating to "Cell"). Conversely, we perceive the following topics as difficult: CN (with complex words such as "Vestibulo-cochlear" and "Glossopharyngeal"), CF (for having detailed functions despite us only requiring "Sensory," "Motor," or "Both"), KS (with complex words such as "Ketoglutarate" and "Oxaloacetate"), and KR (with complex words such as "Oxidative Decarboxylation" and "Phosphorylation"). Though this may be subjective, we decided to back it up by asking participants to answer the question "I knew the topic, (name of topic), prior to taking the experiment" for each topic on a 7-point Likert Scale (1 for "Strongly Disagree," 7 for "Strongly Agree"). For each topic, we calculated each group's average familiarity score, which ranges from 1 to 7. We created a graph of the two groups' familiarity scores for all six topics, which is shown in Section 4.2.2.

As part of the study, participants were tested on these topics, but not their corresponding mnemonics, though the mnemonics are certainly intended to help the experimental group. Participants first needed to recall all terms for each topic in order, by typing them without aid. Next, they engaged in multiple-choice questions and needed to recognize the correct answer among each set. Table 9 shows the recognition questions for Cranial Nerves and Functions, as well as their four possible answers. Table 10 shows the recognition questions for Biological Organization Levels and Disciplines, as well as their four possible answers. Table 11 shows the recognition questions for Krebs Cycle Substrates and Reactions, as well as their four possible

Table 7: Mnemonics for Cranial Nerves and Functions. Licensed under CC BY-NC-SA.

Nerve	Mnemonic	Function	Mnemonic
Olfactory	Oldfactory	Sensory	Seers
Optic	Optician's	Sensory	Say
Oculomotor	Oculi (eyes)	Motor	Move
Trochlear	Trouble	Motor	Motor
Trigeminal	Trigonometrically,	Both	But
Abducens	Abductor's (Ab doctor's)	Motor	Motorist
Facial	Facial	Both	Bosses
Vestibulocochlear	Vestibule (eyes)	Sensory	Say
Glossopharyngeal	Glossing pharyngeally (near the neck)	Both	Big
Vagus	Vagina	Both	Brain
Accessory	Accessory	Motor	Matters
Hypoglossal	Hypoglossal (unshiny)	Motor	Most

Table 8: Mnemonics for Biological Organization Levels and Disciplines. Licensed under CC BY-NC-SA.

Level	Mnemonic	Discipline	Mnemonic
Atom	At	Chemistry	Chemist
Molecule	Moldova's (country sandwiched between Romania and Ukraine)	Molecular Biology	Mollahs (experts)
Organelle	Original	Cellular Biology (Cytology)	Callout
Cell	Cell,	Cellular Biology (Cytology)	Cellular
Tissue	Tutsi	Histology	Historians
Organ	Organized	Anatomy	Analyzing
Organ System	Organic	Physiology	Physical
Organism	Orgy	Biology	Bones
Population	Popular,	Population Biology	Populating
Community	Commanding	Community Ecology	Compton (Hip-Pop city in California),
Ecosystem	Economic	Ecology	Echoing
Biosphere	Boom	Biogeography	Bigotry

Table 9: Recognition questions for Cranial Nerves and Functions.

Question	Answers (With First One Correct)
Cranial nerves play roles in all EXCEPT:	Digesting food, Smelling, Seeing, Moving the face
Where are the cranial nerves located in the nervous system?	Peripheral nervous system, Central nervous system, Autonomic nervous system, Somatic nervous system
Which of these is TRUE about cranial nerves?	They transmit sensory and motor information between the brain and upper body, They originate in the spinal cord, They control digestion, They mostly regulate hormone release
Which cranial nerve does not originate from the brain?	Accessory nerve, Optic nerve, Facial nerve, Trigeminal nerve
The glossopharyngeal nerve is best described as:	Sensory and Motor, Motor, Sensory, Autonomic
Which function does the hypoglossal nerve serve?	Motor, Sensory, Motor and Sensory, Autonomic
The vestibulocochlear nerve has what type of function?	Sensory, Motor, Sensory and Motor, Autonomic
The trochlear nerve is primarily responsible for:	Motor control, Taste, Smell, Balance

answers. Recognition was done after recall because doing it first would show participants some of the correct answers for recall and thus decrease the time between seeing them and needing to type them without aid.

Moreover, participants will need to complete surveys at the end. Ten of the questions (grouped into five pairs) from the UX survey were involved in the scope of this study. Participants will need to express on a 7-point Likert Scale (1 for "Strongly Disagree," 7 for "Strongly Agree") how much they agree with the following ten statements: "I find the steps to use the Acronimps game easy to follow" and "The Acronimps game was confusing to use" (for the variable named Perceived Ease of Use, or PEOU), "I found the Acronimps game useful to learn memory-intensive course content" and "The Acronimps game will help me to prepare for memory-intensive course exams" (for Perceived Usefulness, or PU), "I found using the Acronimps game enjoyable" and "I did not find using the Acronimps game fun" (for Perceived Enjoyment, or PE), "The Acronimps game is trustworthy" and "I think the Acronimps game does not have my best interest in mind" (for Perceived Trustworthiness, or PT), and "The features of the Acronimps game convince me to adopt it to study for a memory-intensive course exam" and "The features of the Acronimps game motivate me to use it in my future memory-intensive courses" (for Willingness to Adopt, or WTA). These questions were adapted from a paper on COVID Alert [122], while the variables they are associated with (Ease of Use, Usefulness, Enjoyment, and Willingness to Adopt), were adapted from the HUED-CUP UX framework [123]. Demographic information was also collected from each participant, though we did not analyze whether there were any correlations between them and test scores. For year of study, any response other than "1st year," "2nd year," "3rd year," "4th year," "Master's," or "PhD," which are the typical years of study in university, was listed as "Other." Likewise, for course of study, any response other than "Computer Science," "Psychology," or "Biology," which were the three courses involved in our study, was listed as "Other." Figures B1 to B6 (in Appendix B) show distributions among the control and experimental groups for gender,

Table 10: Recognition questions for Biological Organization Levels and Disciplines.

Question	Answers (With First One Correct)
Which level describes a single living being made of organ systems?	Organism, Organ, Tissue, Population
What is a group of different species living together in one place?	Community, Population, Organism, Biosphere
What do you call a group of organisms of the same species in one area?	Population, Community, Ecosystem, Organism
Which level includes specialized structures within a cell?	Organelle, Tissue, Molecule, Organ
Which discipline studies groups of similar cells?	Histology, Anatomy, Cellular Biology, Ecology
Which field studies organelles like mitochondria and chloroplasts?	Cellular Biology, Histology, Anatomy, Molecular Biology
Which discipline studies groups of the same species in one area?	Population Biology, Community Biology, Histology, Cellular Biology
Which discipline focuses on systems like the respiratory system?	Physiology, Anatomy, Ecology, Molecular Biology

Table 11: Recognition questions for Krebs Cycle Substrates and Reactions.

Question	Answers (With First One Correct)
What is required for the cycle to repeat?	Regeneration of Oxaloacetate, Production of FADH2, Release of CO2, Presence of Malate
Which substrate is the direct precursor to Malate?	Fumarate, Succinate, Citrate, α-Ketoglutarate
What is a key byproduct when Isocitrate is oxidized?	Citrate, FADH2, ATP, CO2
What is the product of the hydration of Fumarate?	Malate, Oxaloacetate, Citrate, Succinyl-CoA
Which step produces ATP?	Oxidation of Isocitrate, Conversion of Malate to Oxaloacetate, Conversion of Succinyl-CoA to Succinate, Formation of Citrate from Acetyl-CoA
What transformation happens when Succinate becomes Fumarate?	Oxidation, Isomerization, Hydration, Decarboxylation
What happens to Malate in Step 8?	It is oxidized to regenerate Oxaloacetate, It undergoes phosphorylation to form ATP, It is hydrated to form Citrate, It is decarboxylated to produce Fumarate
What happens to Succinyl-CoA in Step 5?	It is phosphorylated to produce Succinate and ATP, It is oxidized to produce NADH, It undergoes hydration to form Malate, It combines with Oxaloacetate to form Citrate

age, year of study, course of study, race or ethnicity, and average grades. They only include demographics of participants who provided demographics and were not excluded from performance analysis (explained in Section 4.2.3. While some demographics were relatively balanced, such as average grades in both groups, others were rather unbalanced, such as the predominance of South Asians in both groups.

Now I will explain the study itself in full detail. As shown in Figure 11, in Part 1 of the study, all participants were first given details on the study and asked to sign a consent form whose link was sent through email; this took up to 10 minutes. Next, they performed different actions depending on which group they are in. (1) The control group watched a 10-minute lecture on Biology topics before spending 45 minutes reading a physical version of the slides and taking notes as they saw fit. (2) The experimental group watched a 10-minute training video on using the SAVE Tool. Here, the topic was Big Five Personality Traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) and the provided mnemonic was "Open Container Exited Algeria's Nuke." The participants then spent the next 10 minutes using the SAVE Tool and entering this mnemonic's artifacts. They accessed the SAVE Tool through a link we sent through email; we also sent the images they needed in a folder attached to the email. Next, they watched a 10-minute training video on using the Acronimps game with the same mnemonic (I removed the minimum-five-mnemonics requirement for the experiment) and spent the next 10 minutes playing it themselves. They accessed the game through a link to an apk file (for Android) or through scanning a QR code and playing it on Expo Go, which they were told to download before the session (for iOS or Android). They then watched the same 10-minute Biology lecture as the control group, before reading the physical version for 10 minutes and spending 35 minutes practicing the Biology mnemonics with the SAVE Tool and Acronimps game as they saw fit. When using the SAVE Tool to study, all they had to fill in for the Biology mnemonics were the mnemonic terms, as all the other details of each mnemonic were provided by us to save time. Because some of the first few participants in the experimental group heavily ignored the SAVE Tool in favour of the Acronimps game and then did very poorly on the test, we asked all the participants afterwards to use the SAVE Tool first and then test themselves on recalling at least half of the target information right on paper or a device without aid. Once they did so, they could move on to the Acronimps game. For both groups, when their preparation time was up, they took a 5 minute break before being presented with a test (whose link was provided through email) on the Biology lecture, which lasted up to 45 minutes and tested their immediate memory. We will call this test T1. The test's recall section involved recalling each of the six topics' target information as shown in Tables 6, 7, and 8 (not the mnemonics, as stated before). The recognition section involved selecting the correct answers to the questions shown in Tables 9, 10, and 11. They all completed a quick demographic survey afterwards, with the experimental group also taking UX surveys on the SAVE Tool and Acronimps game. The total time for the surveys (whose links were provided through email) was 10 minutes, though it was not strict and never needed to be enforced.

As shown in Figure 12, in Part 2 of the study, the participants also returned approximately one week after the day they did Part 1. We chose one week since Murre and Dros [124] confirmed that memory of given information can decline considerably over a week; though the decline starts fast, it slows down later, and we wanted to give participants time to incorporate their return to the lab into their schedule. Both groups were first given details on the study and asked to sign a consent form; this took up to 10 minutes. They

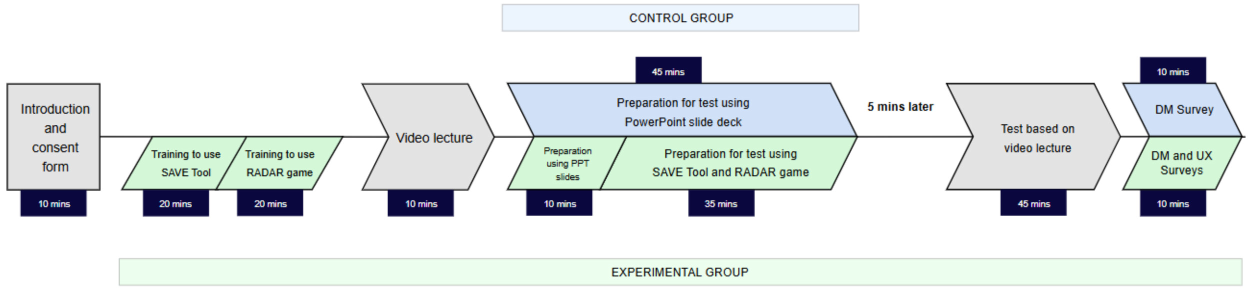


Figure 11: Part 1 of the study (DM stands for demographic).

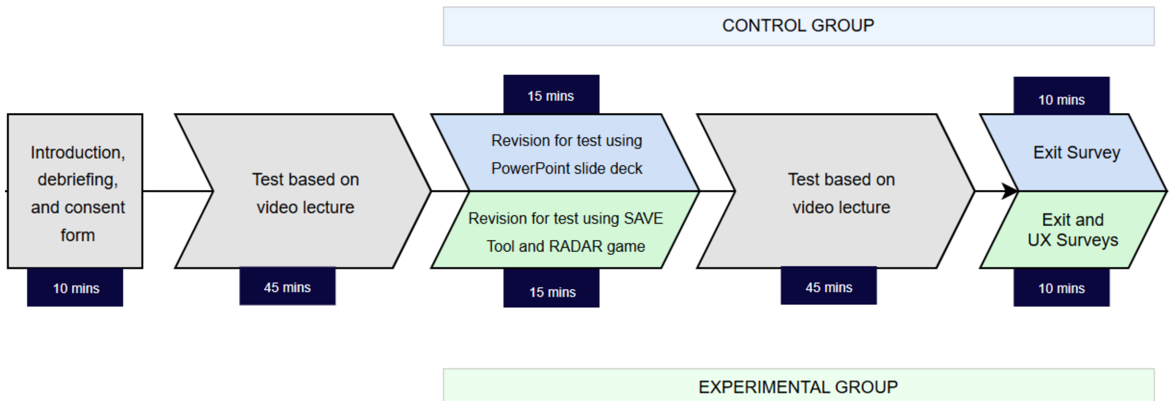


Figure 12: Part 2 of the study.

then took the same test as the one they took in Part 1, once again for up to 45 minutes, to test their delayed memory. We will call this test T2. Afterwards, the control group took 15 minutes to read the physical version of the Biology lecture slides while the experimental group took 15 minutes to practice mnemonics with the SAVE Tool and Acronimps game. During the 15-minute preparation, the experimental group was once again told to use the SAVE Tool first before playing the Acronimps game. However, they simply had to view the information that they typed in Part 1 (or fill in any incomplete spots), and they did not need to test themselves on the target info. Both groups then spent up to 45 minutes taking the same test a third time to test how effective their respective methods are at refreshing their memory. We will call this test T3. After T3, the experimental group took UX surveys on the SAVE Tool and Acronimps game immediately afterwards. Both groups concluded with an exit survey. Once again, the total time for the surveys was 10 minutes but it was not strict.

Because there are many possible answers for the recall section of the test, we needed to clean the data first. On Google Sheets, we inserted empty rows between response rows to contain either blanks, 0s, or 1s. For each individual answer, we used the following rules:

- If there are alphabetic cues at the beginning of the response that maps to the correct answer, e.g., “Hydr” (in the response) and “Hydration” (full correct answer), leave the response unchanged and leave a blank underneath.
- If there is a word or part of a word in the response that maps to the correct answer which is a phrase, e.g., “Ecology” (response) and “Community Ecology” (correct answer), leave the response unchanged

and leave a blank underneath.

- If the response is the correct answer presented in another way, change the response to the correct answer and input a 1 underneath. The 1 is meant to signify a positive modification to the answer due to feeling that the participant likely understands the deeper meaning of the correct answer. For example, if the correct answer is "Optic" (in the context of Cranial Nerves) and the participant typed "Optic nerve," remove "nerve" and input a 1 underneath. Similarly, for the optic nerve's function, which should be "Sensory," if the participant typed "Vision," change it to "Sensory" and input a 1 underneath because they have went above and beyond.
- If the response has coincidental alphabetic cues not at the beginning of the word/phrase (response) matching the answer, e.g., "Condensation" (response) and "Isomerization" (correct answer), or "Ecology" and "Biology," replace the answer with a 9 (because no question's correct answer contains numbers) and input a 0 underneath. The 0 is meant to signify a negative modification to the answer due to feeling that the participant is not on the right track at all. Do the same thing if there are no cues at all, e.g., "N/A" (response) and "Motor" (correct answer).
- For CF, BD, and KR, the correctness of participants' responses depends on what they typed for CN, BL, and KS respectively. If a response in position n for CF/BD/KR is not associated with the n-th response for CN/BL/KS (e.g., they said "Olfactory" first for CN and "Motor" first for CF when the latter should be "Sensory"), replace the n-th response for CF/BD/KR with a 9 and put a 0 underneath. If the n-th response for CN/BL/KS had already been converted to a 9, then use what the correct answer should be for comparison.

Because each participant needed to do a total of three tests (T1, T2, and T3), and there are two participant groups, there are a total of six distinct tests (albeit with the same set of questions). For each test's recall section, Sarah Ejaz and I each cleaned the data individually, and afterwards we compared our results. Because Cohen's Kappa can only be done with two different values, we calculated three types of Cohen's Kappa values based on different pairwise comparisons: 0 vs. 1 or blank, 1 vs. 0 or blank, and blank vs. 0 or 1. We also resolved the disagreements themselves to create a definitive data sheet that was involved in the data analysis. In addition, for each type of Cohen's Kappa, we created a table containing each participant's three Cohen's Kappa values (one for each test they took) and calculated the average. We also calculated the average of each of the six distinct tests. We then calculated the average of test averages and the average of participant averages; the two are not necessarily equal due to the existence of blanks caused by each participant only doing up to three out of six distinct tests. In total, there are six overall averages (i.e., averages of averages) that needed to be considered.

In the data analysis itself, we used orthographic similarity to determine each individual answer's score. Oyibo [125] advised to use orthographic similarity to eliminate subjective judgment, especially in cases where participants' answers seem to be misspelled versions of correct answers. Orthographic similarity ignores semantics and focuses on correct spelling to provide a score between 0 and 1 inclusive. Our particular method is Levenshtein distance, which is the minimum number of edits (e.g., replacements, deletions) needed to reach the correct answer; more edits result in lower scores. However, because some participants misinterpreted the answers and format we desired but still made it clear they fully understood the meaning of the answer, we needed to positively modify this answer and input a 1 underneath. An example would be the aforementioned “Vision” against “Sensory” for the optic nerve; vision was explicitly stated as the exact function for it, so it would be unfair to give them a non-perfect score for it, let alone consider them completely wrong. Conversely, because those who provided an answer that happens to share letters with the correct answer but is clearly not on the right track should not receive partial marks for this sort of coincidence, we needed to replace their answer with a 9 and input a 0 underneath. An example would be if they said “Motor” when the correct answer is “Sensory;” both contain “or” near the end but are completely different concepts. Without our intervention, they would have received partial marks for “or.” Order also matters during score calculations. For example, the first response for CN will be compared against “Olfactory.” However, we did not replace responses in the wrong position with a 9 beforehand, so as to allow participants to receive minimal but non-zero marks for providing what would possibly be mapped to a correct answer in another position. For example, if a participant said ”Optic” first for CN instead of second as it should have been, we would leave it because it is a correct answer for another position in CN and so the participant can receive a small amount of marks from sharing “O” with “Olfactory.” Note that this only applies to CN/BL/KS; as stated before, CF/BD/KR depends directly on CN/BL/KS.

For each task of each test, each participant would have an individual mean, and there would be a mean of these means. However, participants who missed all three of their tests, mistook them as open-book (as we did not think to invigilate during the very beginning of our study), seemed to not take the study seriously, or seemed to have special needs we could not accommodate were excluded from the analysis. In addition, whenever a participant missed a task of a particular test but still submitted responses in at least one task of at least one test, or heavily favoured the Acronimps game over the SAVE Tool during preparation, they would simply receive their group’s mean (ignoring their own score) for the particular task. Moreover, those who received a score of less than 0.1 for either of their T1 tasks or less than 0.2 for either of their T3 tasks for unclear reasons were considered outliers and received their group’s mean (ignoring their own score) for the particular task. The reason why we did this replacement in these situations was because the rest of their data did not look out of the ordinary, so we did not want to let what was possibly an outlier non-indicative of their typical behaviour skew results, and we did not want to disregard all their data. We also chose 0.1 as a threshold for T1 based on heuristics from seeing that most participants’ scores were considerably above that, and simply removing particular scores instead of using a threshold could lead to accusations of “cherry picking.” 0.2 was chosen for T3 because we saw and anticipated that participants would score higher on T3 than T1, so we once again used heuristics. We had no threshold for T2 because we saw and anticipated that participants would score the lowest on T2; we would not be surprised to see someone forget the content altogether in the one-week gap. With every test’s task’s mean established, we then performed a four-way RM-ANOVA analysis involving task, condition, topic, and time. We performed post-hoc analyses for any main effects and simple effects, as well as one-way ANOVAs for any interactions. We also performed three-way RM-ANOVA analysis based on each task, as well as post-hoc analyses for any main effects and simple effects, alongside one-way ANOVAs for any interactions. Some of the resulting tables and figures would be used to show whether our hypotheses are supported, which would be elaborated on in Section 5.

For the UX questions, because their responses involved a 7-point Likert Scale, we averaged the two scores (from the two questions) from each participant for each variable at each time. For negative questions, we subtracted the actual score from 8 to get the score for a positive version of the question. We then categorized averages above 4 as “Agree,” 4 as “Neutral,” and less than 4 as “Disagree.” We created response distributions for each variable at each time based on these three categories. For each variable, we also compared each time’s uncategorized distribution (i.e., still involving 1 to 7) to a randomly generated normal distribution with the same number of data points ($n = 30$), mean of 4, and standard deviation of 1; we hoped to find a significant difference between the normal distribution and the actual one to statistically validate our UX-related hypotheses.

The data analysis and visualizations used some of the same tools as those in a previous paper that Oyibo wrote about dark patterns [126]. In particular, we used several libraries in R. We used the “ggplot2” library [127] to create the graphs shown in this paper. Since our data was not normal, we used the “ARTool” library [128] to perform the Aligned Rank Transform for Non-parametric Factorial Analyses [129].

4 Results

In this section we present the results of the user requirements survey and game evaluation study. For the former, we present graphs comparing the mean scores for each item in Table 5 among a group that prefers professor-created mnemonics and a group that prefers self-created mnemonics, as well as a table showing the count of themes elicited from the qualitative analysis. For the latter, we present a table of Cohen's Kappa values showing the level of inter-rater reliability with respect to the experiment data coding, topic familiarity scores for the control and RADAR groups, group-specific cognitive performance scores for each topic, and various ANOVA tables.

4.1 User Requirements Survey

Of the 122 participants, 46 (37.70%) said "I prefer mnemonics created by the professor over mine" (we will refer to them as "Prof" for convenience, 62 (50.82%) said "I prefer mnemonics created by myself over those created by my professor" (referred to as "Self"), and 14 (11.48%) said "I do not like using mnemonics;" regarding the items listed in Table 5, we compared each of the Prof and Self groups' mean scores. As shown in Figure 13 [3], Ease of Creation and Challenge of Creation were grouped into the Expectancy category. Study, Learning, Retention, Recall, Knowledge, and Passing were grouped into the Value category. Creation Tool and Retrieval Practice Tool were grouped into the Adoption Category. The Self group expressed greater agreement with the Value items, Ease of Creation, and the reversal of Challenge of Creation. Conversely, they expressed slightly less agreement with the Adoption items. For Ease of Creation and the reversal of Challenge of Creation, the Prof group had scores below the neutral value of 4, while the Self group had scores above the neutral value. For the Value and Adoption items, both groups consistently had scores above the neutral value.

Table 12 shows the participant count and Cohen's Kappa for the notable categories formed during qualitative analysis. Visuals and Fun have a particularly high number of relevant responses, with 15/122 (12.30%) participants. Straightforward has the lowest, with 3/122 (2.46%). Also, the Cohen's Kappa values are rather varied, from 0.6541 (for Relevance to Topic) to 1 (for numerous categories such as Meaningfulness).

4.2 Game Evaluation Study

In this section we present the results of the game evaluation study. First, we present inter-rater reliability tables showing the various averages for each of the three types of Cohen's Kappa values. Then, we present the mean scores for topic familiarity. Next, we present graphs showing cognitive performance in each group with regards to task, time, and topic. Then, we present overall descriptive statistics. Afterwards, we present a table showing the results of a four-way RM-ANOVA involving condition, task, topic, and time; we also present three-way RM-ANOVAs for recall and recognition involving condition, topic, and time. Finally, we present tables and graphs corresponding to each hypothesis and show whether it is supported with significant differences.

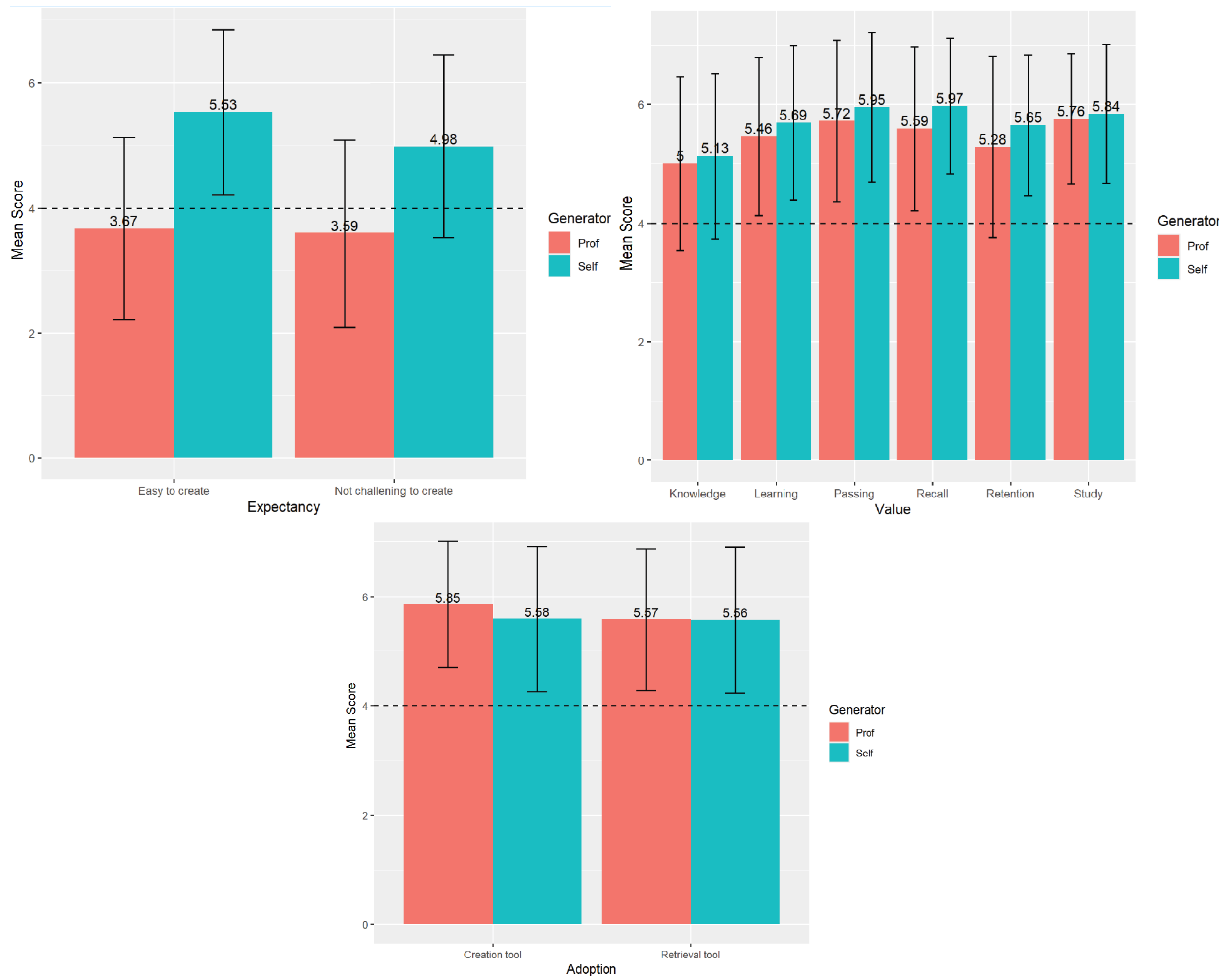


Figure 13: Graphs showing mean scores for each item among the Prof group and the Self group [3], with error bars showing standard deviation and dotted line showing neutral value in range 1 to 7.

Table 12: Thematic analysis of participants' comments in pre-system-design user study; categories arranged in descending count order within each type.

Type	Category	Count (%)	Cohen's Kappa
Mnemonic-Based	Meaningfulness	6 (4.91%)	1.0000
	Relevance to Topic	6 (4.91%)	0.6541
	Humour	5 (4.10%)	0.7421
RADAR-Based	Visuals	15 (12.30%)	0.8751
	Fun	15 (12.30%)	0.8751
	Interactive	10 (8.20%)	0.9429
	Repetition	10 (8.20%)	0.9429
	Progress Tracking	6 (4.91%)	0.9048
	Feedback	5 (4.10%)	1.0000
	Rewards	5 (4.10%)	1.0000
	Matching-Based	4 (3.28%)	1.0000
	Straightforward	3 (2.46%)	1.0000

4.2.1 Inter-Rater Reliability

Table 13 shows each group's test average for each of the three types of Cohen's Kappa at each timepoint. While the majority of values are above 0.9000, the T1, T2, and overall values for the control group regarding "1 vs. Other" range from 0.7877 to 0.8418. In addition, Tables C1 (0 vs. other), C2 (1 vs. other), and C3 (blank vs. other) in Appendix C.1 show the inter-rater Cohen's Kappa score for the coding of each participant's recall-based responses at all three timepoints, respectively.

Table 13: Inter-rater Cohen's Kappa scores for recall-based responses at all three timepoints. C: Control, R: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, OV: Overall

	C				R			
Type	T1	T2	T3	OV	T1	T2	T3	OV
0 vs. Other	0.9494	0.9431	0.9492	0.9472	0.9513	0.9480	0.9697	0.9563
1 vs. Other	0.8153	0.7877	0.9223	0.8418	0.9644	0.9834	0.9360	0.9613
Blank vs. Other	0.9379	0.9413	0.9760	0.9517	0.9469	0.9583	0.9747	0.9600
Average	0.9009	0.8907	0.9492	0.9136	0.9542	0.9632	0.9601	0.9592

4.2.2 Topic Familiarity

Figure 14 shows that only BL and BD, the only topics which we had considered simple, have familiarity scores no less than 3.42 for both groups. Meanwhile, the other four topics, which we had considered difficult, all have familiarity scores no greater than 2.73 for both groups. For each topic, both groups had scores below the neutral value 4, except BL in which the RADAR group had a score of 4.10. For all topics, the RADAR group had a higher score than the control group, with the difference being most prominent in KS.

4.2.3 Cognitive Performance

In this section, we present graphs showing the mean performance for each group with regards to task, time, and topic. We also present a four-way RM-ANOVA table involving condition, topic, time, and task, as well as three-way RM-ANOVA tables for each task involving condition, topic, and time. In addition, we present various one-way ANOVA tables to determine whether our hypotheses are supported.

Visualization of Cognitive Performance: Figures 15 and 16 show each group's score for each topic at each time and overall, for recall and recognition respectively. For recall, the vast majority of the RADAR group's scores are higher than their corresponding control group scores, with a few exceptions, such as BD for T1. For recognition, there are far more occasions in which the RADAR group had a lower score than the control group. In addition, for recall, the scores in T2 are generally visibly much lower than those in T1 and T3. For recognition, this is not the case.

Figure 17 shows each group's score for each task at each time. For recall, the RADAR group visibly has a higher score than the control group at all three times. For recognition however, while the RADAR group starts off with a higher score than the control group at T1, it ends up having lower scores than the control

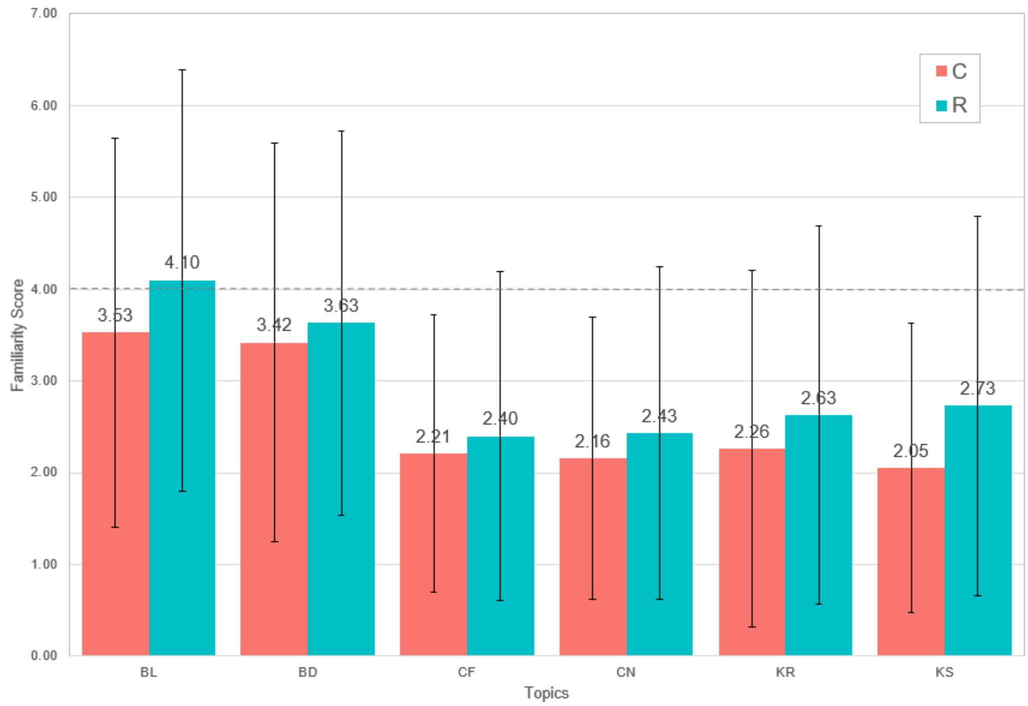


Figure 14: Mean familiarity scores across topics for the control and experimental groups, with error bars showing standard deviation and dotted line showing neutral value in range 1 to 7. C: Control, R: RADAR, BL: Biological Organization Levels, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

group at T2 and T3. Yet for all three times, the gap between the recognition scores for the two groups is smaller than the gap between the corresponding recall scores for the two groups.

Overall Descriptive Statistics: For each time a particular group performed a particular task, we also present how many participants did so, as well as the mean, standard deviation, median, min, max, range, skew, kurtosis, and standard error. The statistics are all shown in Table 14. Note that C has $n = 22$ even though I had stated there were 23 participants in the control group; this is because one such participant seemed to have mistaken the test as open-book based on having so many perfect answers suddenly followed by multiple "N/A" answers as if they spent so much time searching for answers and then ran out of time. R has $n = 30$ even though I had stated there were 32 participants; one participant had to leave for an emergency before reaching T1 and they did not return, while another could not complete the tests properly within the allotted time (they had visibly rushed as time ran out). All three of these participants were excluded from analysis.

Four-Way RM-ANOVA: The result of the four-way RM-ANOVA shows that there is a main effect of topic $[F(5,1750) = 46.44, p < 0.001]$, time $[F(2,1750) = 111.87, p < 0.001]$, and task $[F(1,1750) = 129.28, p < 0.001]$. There are also interactions between condition and topic $[F(5,1750) = 4.69, p < 0.001]$, condition and task $[F(1,1750) = 15.21, p < 0.001]$, topic and task $[F(5,1750) = 26.75, p < 0.001]$, and time and task $[F(2,1750) = 44.31, p < 0.001]$ (Table 15).

There is no need for a task-based post-hoc analysis because there are only two different tasks: recall

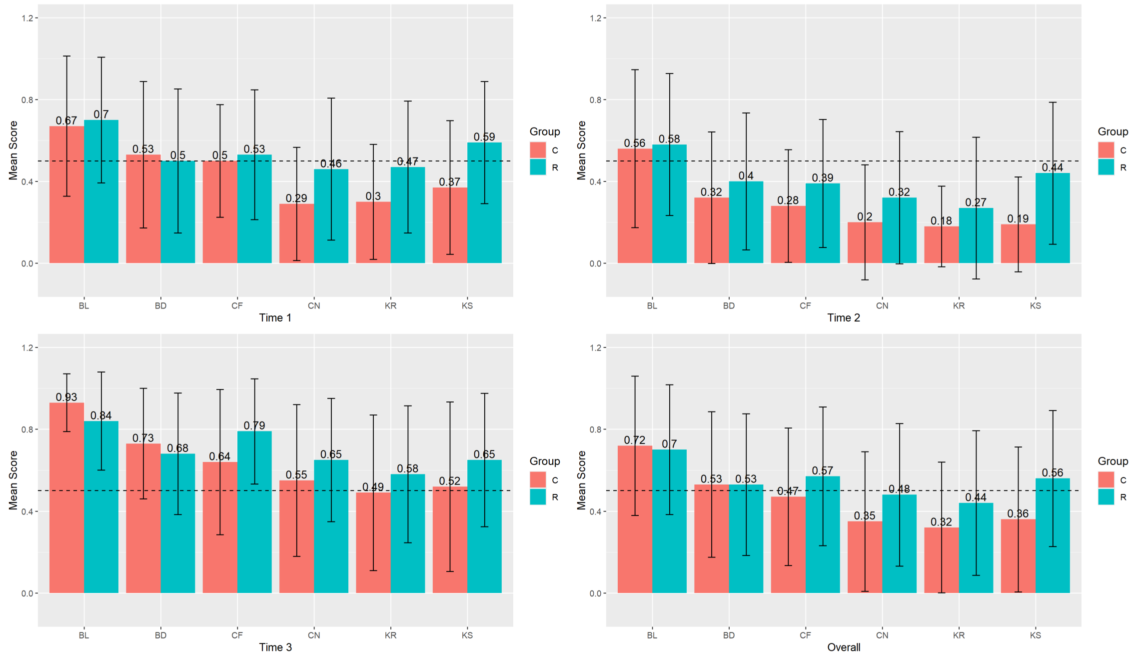


Figure 15: Mean recall scores for the specific topics at the three points in time and overall, with error bars showing standard deviation and dotted line showing neutral score in range 0 to 1. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

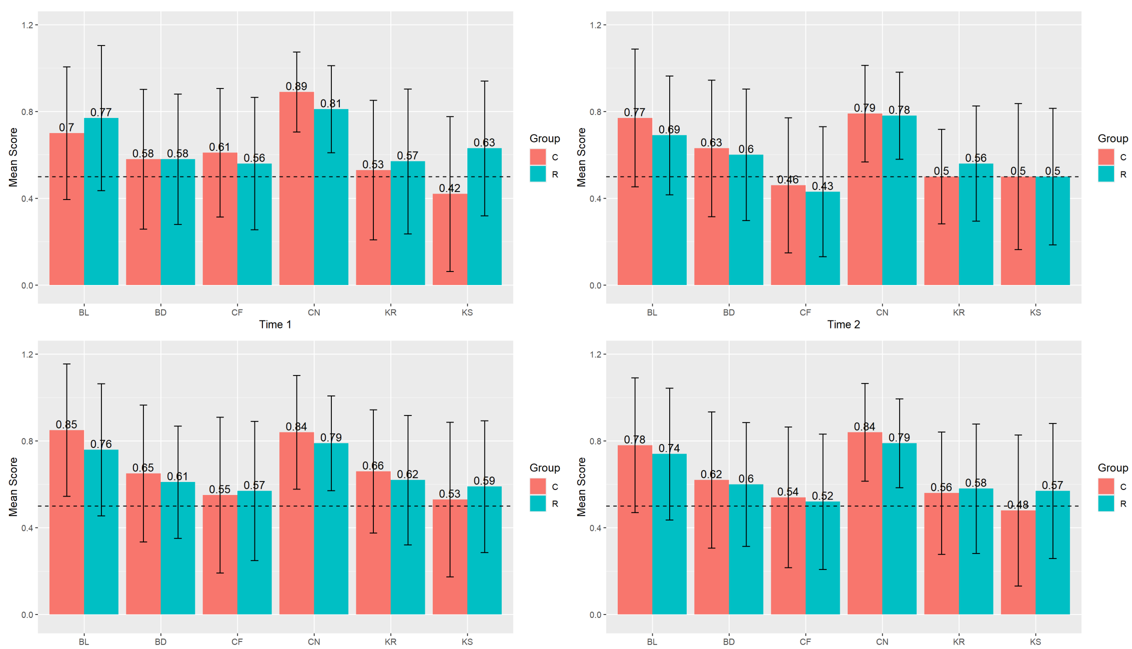


Figure 16: Mean recognition scores for the specific topics at the three points in time and overall, with error bars showing standard deviation and dotted line showing neutral score in range 0 to 1. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

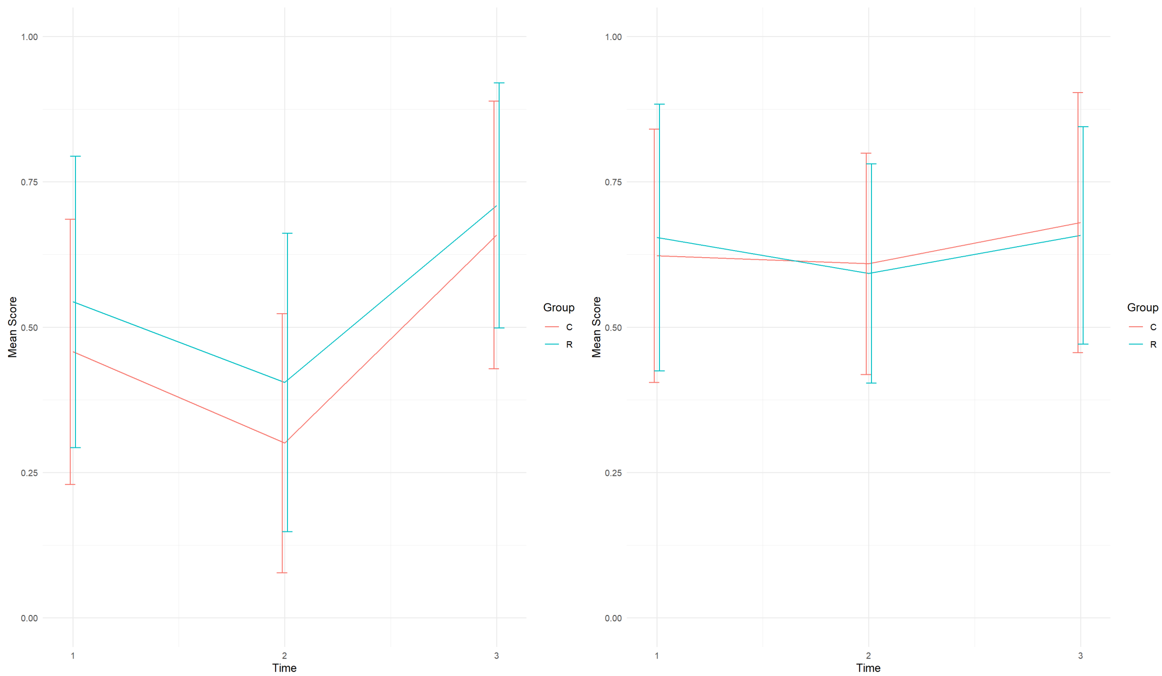


Figure 17: Overall mean scores of the remembering measure (aka recall - left) and understanding measure (aka recognition - right) for the three points in time, with error bars showing standard deviation and range being 0 to 1. C: Control, R: RADAR.

Table 14: Overall descriptive statistics. SD: Standard Deviation, SE: Standard Error, C: Control, R: RADAR.

Group	n	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE	Time	Task
C	22	0.45	0.23	0.41	0.16	0.87	0.70	0.30	-1.42	0.05	1	RCL
R	30	0.54	0.25	0.57	0.13	0.98	0.85	-0.16	-1.02	0.05	1	RCL
C	22	0.29	0.22	0.32	0.00	0.65	0.65	0.16	-1.47	0.05	2	RCL
R	30	0.40	0.26	0.39	0.02	0.96	0.94	0.50	-0.62	0.05	2	RCL
C	22	0.64	0.24	0.58	0.24	0.99	0.75	0.06	-1.59	0.05	3	RCL
R	30	0.70	0.22	0.70	0.26	1.00	0.74	-0.30	-0.98	0.04	3	RCL
C	22	0.62	0.22	0.62	0.29	1.00	0.71	-0.09	-1.29	0.05	1	RCG
R	30	0.65	0.23	0.65	0.25	1.00	0.75	-0.19	-1.00	0.04	1	RCG
C	22	0.61	0.19	0.60	0.25	0.92	0.67	0.04	-1.13	0.04	2	RCG
R	30	0.59	0.19	0.59	0.29	0.96	0.67	0.09	-1.04	0.03	2	RCG
C	22	0.68	0.22	0.67	0.33	1.00	0.67	0.07	-1.37	0.05	3	RCG
R	30	0.66	0.19	0.66	0.29	0.96	0.67	-0.26	-0.81	0.03	3	RCG

(0.51) and recognition (0.64). As such, the aforementioned main effect of task $[F(1,1750) = 129.28, p < 0.001]$ is clearly between these two. Tables C4 and C5 regarding the remaining main effects are shown in Appendix C.3 due to not relating to our hypotheses.

Table C6 shows the interaction between topic and task; Table C7 shows the two task's topic-based pairwise comparison due to the simple effect of topic in the former table. These two tables are shown in Appendix C.3 due to not relating to our hypotheses.

Table C8 shows the interaction between time and task; Table C9 shows the two tasks' time-based pair-

Table 15: Overall four-way RM-ANOVA. Df: Degrees of freedom, res: residual, PCode: Participant Code.

Term	Error	Df	Df.res	F-value	p-value
Condition	PCode	1	50	0.52	0.4728
Topic	Within	5	1750	46.44	0.0000
Time	Within	2	1750	111.87	0.0000
Task	Within	1	1750	129.28	0.0000
Condition:Topic	Within	5	1750	4.69	0.0003
Condition:Time	Within	2	1750	2.08	0.1257
Topic:Time	Within	10	1750	0.81	0.6198
Condition:Task	Within	1	1750	15.21	0.0001
Topic:Task	Within	5	1750	26.75	0.0000
Time:Task	Within	2	1750	44.31	0.0000
Condition:Topic:Time	Within	10	1750	0.93	0.5022
Condition:Topic:Task	Within	5	1750	1.30	0.2609
Condition:Time:Task	Within	2	1750	0.77	0.4616
Topic:Time:Task	Within	10	1750	0.98	0.4551
Condition:Topic:Time:Task	Within	10	1750	0.68	0.7483

wise comparison due to the simple effect of time in the former table. These two tables are shown in Appendix C.3 due to not relating to our hypotheses.

Three-Way RM-ANOVAs for Recall and Recognition: The result of the three-way RM-ANOVA for recall shows that there is a main effect of topic $[F(5,850) = 34.15, p < 0.001]$ and time $[F(2,850) = 142.27, p < 0.001]$. There is also an interaction between condition and topic $[F(5,850) = 4.34, p < 0.001]$ (Table 16). The result of the three-way RM-ANOVA for recognition shows that there is a main effect of topic $[F(5,850) = 44.72, p < 0.001]$ and time $[F(2,850) = 7.82, p < 0.001]$ (Table 16).

Table 16: Overall three-way RMANOVAs for recall (middle) and recognition (right) individually based on condition, time, and topic. Df: Degrees of freedom, res: residual, PCode: Participant Code.

Term	Error	Df	Df.res	F-value	p-value	F-value	p-value
Condition	PCode	1	50	1.87	0.1772	0.00	0.9446
Topic	Within	5	850	34.15	0.0000	44.72	0.0000
Time	Within	2	850	142.27	0.0000	7.82	0.0004
Condition:Topic	Within	5	850	4.34	0.0007	1.65	0.1455
Condition:Time	Within	2	850	0.47	0.6282	2.64	0.0722
Topic:Time	Within	10	850	0.77	0.6546	1.18	0.2991
Condition:Topic:Time	Within	10	850	0.64	0.7828	1.11	0.3519

Tables C10 and C11 show the two tasks' topic-based and time-based pairwise comparisons respectively due to the main effects of topic and time in Table 16. They are shown in Appendix C.3 due to not relating to our hypotheses.

Tables C12 shows an interaction between condition and topic for recall; Table C13 shows the two groups' topic-based pairwise comparison for recall due to the simple effect of topic in the former table. They are shown in Appendix C.3 due to not relating to our hypotheses.

Table C14 shows an interaction between condition and topic for recognition; Table C15 shows the two groups' topic-based pairwise comparison for recognition due to the simple effect of topic in the former table. They are shown in Appendix C.3 due to not relating to our hypotheses.

RADAR Vs. Control for Each Topic: Here, we show whether the Acronimps game leads to greater performance in each of the six topics than the control condition. As seen in Table 17, the experimental group has higher overall scores than the control group for CN (0.63 vs. 0.59) and CF (0.55 vs. 0.51), but neither difference is significant (the former has $p = 0.4863$ and the latter has $p = 0.2905$). The hypotheses that "R>C for CN and CF" are not supported (non-significant results), though the means follow their direction. The experimental group has lower scores than the control group for BL (0.72 vs. 0.75) and BD (0.56 vs. 0.58), though neither difference is significant ($p = 0.2308$ for the former, $p = 0.8178$ for the latter). "R>C for BL and BD" are not supported (non-significant results), and the means follow the opposite direction of the hypothesis. The experimental group has a higher score than the control group for KS (0.57 vs. 0.42), with the difference being significant $p < 0.001$. "R>C for KS" is supported. The experimental group has a higher score than the control group for KR (0.51 vs. 0.44), but the difference is not significant ($p = 0.0744$), although almost marginal. "R>C for KR" is not supported (non-significant results), though the means follow the hypothesis's direction. Table 18 follows up on Table 17 by showing the two groups' topic-based pairwise comparison due to the simple topic effect found there.

Table 17: One-way ANOVA based on condition and topic overall due to interaction between them in Table 15. C: Control, R: RADAR, CN: Cranial Nerves, CF: Cranial Nerves Functions, BL: Biological Organization Levels, BD: Biological Organization Disciplines, KS: Krebs Cycle Substrates, KR: Krebs Cycle Reactions.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of topic	CN	0.59	0.63	$F(1,310) = 0.49$, $p = 0.4863$
	CF	0.51	0.55	$F(1,310) = 1.12$, $p = 0.2905$
	BL	0.75	0.72	$F(1,310) = 1.44$, $p = 0.2308$
	BD	0.58	0.56	$F(1,310) = 0.05$, $p = 0.8178$
	KS	0.42	0.57	$F(1,310) = 13.72$, $p = 0.0003$
	KR	0.44	0.51	$F(1,310) = 3.20$, $p = 0.0744$
	Topic Effect	$F(5,765) = 24.13$, $p = 0.0000$	$F(5,1045) = 16.25$, $p = 0.0000$	

Table 18: Control's and RADAR's topic-based pairwise comparison overall with Tukey method familywise correction due to the simple effect of topic in Table 17. SE: Standard Error, Df: Degrees of freedom, CN: Cranial Nerves, CF: Cranial Nerves Functions, BL: Biological Organization Levels, BD: Biological Organization Disciplines, KS: Krebs Cycle Substrates, KR: Krebs Cycle Reactions.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
CN - CF	0.59	0.51	62.74	22.58	765	2.78	0.0620	0.63	0.55	87.59	26.19	1045	3.34	0.0110
CN - BL	0.59	0.75	-99.27	22.58	765	-4.40	0.0002	0.63	0.72	-83.10	26.19	1045	-3.17	0.0193
CN - BD	0.59	0.58	23.77	22.58	765	1.05	0.8997	0.63	0.56	76.20	26.19	1045	2.91	0.0429
CN - KS	0.59	0.42	115.01	22.58	765	5.09	0.0000	0.63	0.57	72.21	26.19	1045	2.76	0.0654
CN - KR	0.59	0.44	99.88	22.58	765	4.42	0.0002	0.63	0.51	121.84	26.19	1045	4.65	0.0001
CF - BL	0.51	0.75	-162.00	22.58	765	-7.17	0.0000	0.55	0.72	-170.70	26.19	1045	-6.52	0.0000
CF - BD	0.51	0.58	-38.97	22.58	765	-1.73	0.5150	0.55	0.56	-11.40	26.19	1045	-0.44	0.9980
CF - KS	0.51	0.42	52.27	22.58	765	2.32	0.1892	0.55	0.57	-15.39	26.19	1045	-0.59	0.9919
CF - KR	0.51	0.44	37.14	22.58	765	1.65	0.5688	0.55	0.51	34.24	26.19	1045	1.31	0.7811
BL - BD	0.75	0.58	123.03	22.58	765	5.45	0.0000	0.72	0.56	159.30	26.19	1045	6.08	0.0000
BL - KS	0.75	0.42	214.28	22.58	765	9.49	0.0000	0.72	0.57	155.31	26.19	1045	5.93	0.0000
BL - KR	0.75	0.44	199.15	22.58	765	8.82	0.0000	0.72	0.51	204.94	26.19	1045	7.82	0.0000
BD - KS	0.58	0.42	91.24	22.58	765	4.04	0.0008	0.56	0.57	-3.99	26.19	1045	-0.15	1.0000
BD - KR	0.58	0.44	76.11	22.58	765	3.37	0.0102	0.56	0.51	45.64	26.19	1045	1.74	0.5039
KS - KR	0.42	0.44	-15.13	22.58	765	-0.67	0.9852	0.56	0.51	49.63	26.19	1045	1.89	0.4059

RADAR vs. Control for Recognition and Recall at T1: Here, we show whether the Acronimps game leads to greater performance in recognition and recall at Time 1 compared to the control condition. As seen in Table 19 and Figure 17, the experimental group has a higher overall score than the control group for T1, for both recall (0.54 vs. 0.45 with $p < 0.05$) and recognition (0.65 vs. 0.62 with $p = 0.3460$). "R>C for RCG at T1" is not supported (non-significant results), though the means follow the hypothesis's direction. However, the hypothesis "R>C for RCL at T1" is supported.

Table 19: One-way ANOVA between condition and task with regards to T1. C: Control, R: RADAR, RCL: Recall, RCG: Recognition.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of task	RCL	0.45	0.54	$F(1,310) = 6.72$, $p = 0.0100$
	RCG	0.62	0.65	$F(1,310) = 0.89$, $p = 0.3460$
	Task Effect	$F(1,241) = 30.20$, $p = 0.0000$	$F(1,329) = 20.96$, $p = 0.0000$	

RADAR vs. Control for Recognition and Recall at T2: Here, we show whether the Acronimps game leads to greater performance in recognition and recall at Time 2 compared to the control condition. As seen in Table 20 and Figure 17, for T2, the experimental group only has a higher overall score than the control group for recall (0.40 vs. 0.29 with $p < 0.01$), but not for recognition (0.59 vs. 0.61 with $p = 0.6188$). "R>C for RCG at T2" is not supported (non-significant results), with the means following the opposite direction of the hypothesis. "R>C for RCL at T2" is supported.

Table 20: One-way ANOVA between condition and task with regards to T2. C: Control, R: RADAR, RCL: Recall, RCG: Recognition.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of task	RCL	0.29	0.40	$F(1,310) = 9.42$, $p = 0.0023$
	RCG	0.61	0.59	$F(1,310) = 0.25$, $p = 0.6188$
	Task Effect	$F(1,241) = 105.41$, $p = 0.0000$	$F(1,329) = 55.10$, $p = 0.0000$	

Table 21: One-way ANOVA between condition and task with regards to T3. C: Control, R: RADAR, RCL: Recall, RCG: Recognition.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of task	RCL	0.64	0.70	$F(1,310) = 1.26$, $p = 0.2627$
	RCG	0.68	0.66	$F(1,310) = 1.00$, $p = 0.3181$
	Task Effect	$F(1,241) = 6.21$, $p = 0.0134$	$F(1,329) = 1.67$, $p = 0.1966$	

RADAR vs. Control for Recognition and Recall at T3: Here, we show whether the Acronimps game leads to greater performance in recognition and recall at Time 3 compared to the control condition. As seen in Table 21 and Figure 17, for T3, the experimental group has a higher overall score than the control group for recall (0.70 vs. 0.64 with $p = 0.2627$) but not recognition (0.66 vs. 0.68 with $p = 0.3181$). "R>C for RCG at T3" is not supported (non-significant results), with the means following the opposite direction of the hypothesis."R>C for RCL at T3" is not supported (non-significant results), though the means follow the hypothesis's direction.

RADAR vs. Control for Recognition and Recall Overall: Here, we show whether the Acronimps game leads to greater performance in recognition and recall overall (regardless of time) compared to the control condition. Table 22 shows that there is no significant condition effect on recognition $[F(1,934) = 0.08, p = 0.7815]$. "R>C for RCG overall" is not supported (non-significant results), with the means following the opposite direction of the hypothesis. However, there is a significant condition effect on recall $[F(1,934) = 14.97, p = 0.0001]$, thus supporting "R>C for RCL overall."

Difference Between RADAR and Control for Difficult Topics Vs. for Simple Topics: Here, we show whether the difference between the two groups' scores is greater in difficult topics than in simple topics

Table 22: One-way ANOVA due to interaction between condition and task in Table 15. C: Control, R: RADAR, RCL: Recall, RCG: Recognition.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of task	RCL	0.46	0.55	$F(1,934) = 14.97$, $p = 0.0001$
	RCG	0.64	0.63	$F(1,934) = 0.08$, $p = 0.7815$
	Task Effect	$F(1,769) = 86.25$, $p = 0.0000$	$F(1,1049) = 30.45$, $p = 0.0000$	

overall and within both recognition and recall. As stated before, we consider BL and BD to be simple topics while CN, CF, KS, and KR are considered to be difficult. Table 23 shows that on the overall level, the p-value of 0.0645 shows that "Group difference for difficult topics $>$ for simple topics (overall)" is not supported (non-significant results); the differences for the difficult topics combined and for the simple ones combined are unknown. Likewise, the p-value of 0.5372 for recognition shows that "Group difference for difficult topics $>$ for simple topics (RCG)" is not supported (non-significant results); with the combined differences once again being unknown. However, the p-value of $0.0061 < 0.01$ for recall does support "Group difference for difficult topics $>$ for simple topics (RCL)." All the individual recall-based differences can be seen in Figure 15 while the recognition-based ones can be seen in Figure 16. Similarly, Figure C1 (in Appendix C.2) shows the average mean difference for the topics for both tasks at each time.

Table 23: Correlation between topic difficulty and group difference in performance. R: RADAR, C: Control, OV: Overall, RCL: Recall, RCG: Recognition.

R	C	Level	Cor	p-value
0.59	0.55	OV	0.78	0.0645
0.55	0.46	RCL	0.94	0.0061
0.63	0.64	RCG	0.32	0.5372

Recognition vs. Recall: Here, we show whether the recognition task has a higher score than the recall task overall and within both the control and experimental groups. The overall score for recognition is 0.64, compared to recall's overall score of 0.51; Table 15 shows that the difference is significant ($p < 0.001$), indicating that "RCG>RCL overall" is supported. In addition, Table 22 shows that among the control group, recognition's score is 0.64 while recall's score is 0.46, and the difference is significant ($p < 0.001$); this supports "RCG>RCL for C." Likewise, for the experimental group, recognition's score is 0.63 while recall's score is 0.55, with the difference being significant ($p < 0.001$); this supports "RCG>RCL for R."

4.2.4 UX Variables

In this section, we present Table 24, which shows response distributions for each of the five RADAR UX variables at T1 and T3, with scores of 5 to 7 counting as "Agree", 4 as "Neutral", and 1 to 3 as "Disagree." We also compare each variable's two means (one for T1, one for T3) with the neutral value of 4 in order to determine whether the UX-based hypotheses are supported.

Table 24: Response distribution for each of the five UX variables. T1: Time 1, T3: Time 3.

Variable	Response	T1 Count (%)	T3 Count (%)
Perceived Ease of Use	Agree	25 (83.33%)	26 (86.67%)
	Neutral	3 (10%)	2 (6.67%)
	Disagree	2 (6.67%)	2 (6.67%)
Perceived Usefulness	Agree	20 (66.67%)	19 (63.33%)
	Neutral	1 (3.33%)	4 (13.33%)
	Disagree	9 (30%)	7 (23.33%)
Perceived Enjoyment	Agree	26 (86.67%)	23 (76.67%)
	Neutral	1 (3.33%)	5 (16.67%)
	Disagree	3 (10%)	2 (6.67%)
Perceived Trustworthiness	Agree	23 (76.67%)	24 (80%)
	Neutral	6 (20%)	5 (16.67%)
	Disagree	1 (3.33%)	1 (3.33%)
Willingness to Adopt	Agree	21 (70%)	21 (70%)
	Neutral	4 (13.33%)	1 (3.33%)
	Disagree	5 (16.67%)	8 (26.67%)

Perceived Ease of Use, Perceived Enjoyment, Perceived Usefulness, and Perceived Trustworthiness: For each variable, we compared its overall two means (with the seven possible values and from the two times) to a randomly generated normal distribution with a mean of approximately 4, the same number of participants ($n = 30$), and standard deviation of 1. For PEOU, there was a significant difference between the normal distribution and the actual distribution both times (4.00 vs. 5.86 for T1 with $p < 0.001$, 4.00 vs. 5.74 for T3 with $p < 0.001$), supporting "Mean of PEOU > neutral value." For PE, there was a significant difference between the normal distribution and the actual distribution both times (4.00 vs. 5.86 for T1 with $p < 0.001$, 4.00 vs. 5.78 for T3 with $p < 0.001$), supporting "Mean of PE > neutral value." For PU, there was a significant difference between the normal distribution and the actual distribution (4.00 vs. 4.98 for T1 with $p < 0.05$, 4.00 vs. 4.84 for T3 with $p < 0.01$), supporting "Mean of PU > neutral value." For PT, there was a significant difference between the normal distribution and the actual distribution both times (4.00 vs. 5.72 for T1 with $p < 0.001$, 4.00 vs. 5.62 for T3 with $p < 0.001$), supporting "Mean of PU > neutral value."

Students' Willingness to Adopt RADAR: For the UX question about WTA, we compared its overall two means (with the seven possible values and for the two times) to a randomly generated normal distribution with a mean of approximately 4, the same number of participants ($n = 30$), and standard deviation of 1. There was a significant difference between the normal distribution and the actual distribution (4.00 vs. 5.10 for T1 with $p < 0.001$, 4.00 vs. 4.72 for T3 with $p < 0.01$), supporting "Mean of WTA > neutral value."

5 Discussion

In this section we discuss the results of the user requirements survey and game evaluation study. In particular, we provide explanations for why these results are the way they are. We also discuss how the support of the hypotheses (or lack thereof) can be used to answer the research question. We also discuss limitations of our study and future work.

5.1 User Requirements Survey

It makes sense that as shown in Figure 13 [3], the Prof group would somewhat disagree with mnemonics being easy and not challenging to create as much as the Self group would. Those who do not find mnemonics easy or not challenging to create would prefer for someone else, especially a trusted individual such as a professor, to do it for them. Meanwhile, those who at least somewhat agree that mnemonics are easy and not challenging to create (or at least do not disagree) would be more likely to do it themselves, as it is not as much of a perceived inconvenience to them. As for why each group's scores for Ease of Creation and the reversal of Challenge of Creation are not quite the same (such as 5.53 and 4.98 respectively for the Self group), they most likely did not simply calculate 8 minus the former to get the latter as they may have thought the two questions were not simply opposites or else there would be redundancy. Regardless, the observation that the highest mean is 5.53 indicates that neither group is particularly confident in creating mnemonics, which justifies the creation of a mnemonics creation tool that can guide them.

It is unsurprising that as shown in Figure 13 [3], for all the Value items, the Self group would express greater agreement than the Prof group. Because the Self group finds creating mnemonics to not be much of an issue, they would actually do so, and as mnemonics have been shown to benefit users' memory [73], they likely received better experiences (especially higher test scores) and understood that their mnemonics are the ones to credit. Conversely, the Prof group likely refrained from creating as many mnemonics, and thus would have less experience using mnemonics, especially if their professor did not give them any. As such, they would not give as much credit to mnemonics. Regardless, the observation that the lowest score is 5 indicates that both groups at least somewhat agree that mnemonics are overall valuable in numerous ways, which justifies the inclusion of mnemonics in memory-intensive courses. Doing so would also allow mnemonics to prove themselves to students who did not like using mnemonics.

It is expected that as shown in Figure 13 [3], both groups share nearly the same high amount of agreement (the lowest being 5.56) with regards to Creation Tool and Retrieval Practice Tool. However, the Prof group still has slightly higher scores for the two, which may be because those who currently do not feel confident creating mnemonics would wish for a tool that can assist them, as well as a tool that can test them with the mnemonics. Even so, no matter how difficult they think creating mnemonics is, if a trusted individual such as a professor provides them with tools that specialize in using mnemonics to enhance learning and recall, they would probably have faith in these tools. In addition, they would also assume that either the professor, the creation tool, or both would properly guide them through the process of creating an effective mnemonic. Such an assumption is not unfounded, whether or not they know this; there already exists prior research [44] that had provided a set of defined properties that a mnemonic needs to be effective:

meaningfulness, elaboration, association, and visualization. A professor could simply incorporate this into their mnemonic-related teachings. Regardless, the observation that the lowest score is 5.56 indicates that both groups at least somewhat agree to use such tools, meaning these tools would not be rejected by many students upon their inclusion into memory-intensive courses.

Regarding the qualitative analysis categories, though all their counts may not look very high at first glance, with Straightforward even having only 3/122 (2.46%) participants talk about it, it should be noted that the questions were very open-ended, leading to a great variety in what participants' responses would be about, so even having only 3 responses discussing the same subject is something to consider, let alone 15 (for Visuals and Fun). Also, while some Cohen's Kappa values may appear less than impressive (particularly 0.6541 for Relevance to Topic), it should be noted that when the number of responses where both Sarah Ejaz and I said "yes" to counting as relevant to the category is very low, even a low number of disagreements can drastically decrease the Cohen's Kappa. As such, the low Cohen's Kappa does not indicate that the results for the category are unreliable.

Our mnemonic and game creation took these categories into consideration. The Meaningfulness and Relevance to Topic categories in Table 12 led us to ensure that our mnemonics reflect their target topics, such as "Seers Say Move Motor But Motorist Bosses Say Big Brain Matters Most" clearly reflecting the Sensory and Motor functions, as well as the brain where the cranial nerves reside. This allows the mnemonic to be more closely related to the content. The Humour category also led us to incorporate humour into some of our mnemonics, such as "Condemned Isometrician Oxton Overloaded Pharoah's Ox (with) Hydraulic Oxide" bringing to mind an inflated ox, which may be considered rather cartoonish, making it memorable. Repetition led us to allow repeating a Quest before moving to the next one, in case players feel the need to practice a Quest. Interactive led us to make the game seem like a guide while the Acronimps seem like actual enemies causing problems for players, so that the players would not feel alone. Progress Tracking led us to show progress bars letting players know how far they have gone and how much more is left, so they would have an idea of how much more they need to do. Straightforward led us to provide a linear path for the players to decrease the chances of them getting lost. Visuals led us to ensure the game is full of unique images to maintain players' attention rather than overwhelming them with lots of text. Fun led us to seek to make the game an entertaining experience to complement its educational purpose through the other actions we discussed and will discuss. Feedback led us to ensure the game tells players whether their answers were right or wrong, and in the case of "Castle (Decoding)", we added tidbits to give them additional information related to the target information. Matching-Based led us to give players the opportunity to match mnemonics with their target topics in "Bridge (Association)", and match mnemonic terms with their target terms in "Castle (Decoding)." Rewards led us to implement the XP system where players get more XP for doing better, which would motivate them to give it their all and keep going to rank above other players in "Class Ranking;" the stars obtained from the first three Quests can also motivate players to aim for a perfect score.

5.2 Game Evaluation Study

In this section we discuss the results of the data analysis for the game evaluation study pertaining to our hypotheses. First, we explain the various averages for each of the three types of Cohen's Kappa values. Then, we discuss how the mean scores for topic familiarity justify our simple-difficult topic classification. Next, we discuss implications of overall descriptive statistics. Then, we discuss the results of a four-way RM-ANOVA involving condition, task, topic, and time; we also discuss the results of the three-way RM-ANOVAs for recall and recognition involving condition, topic, and time. Afterwards, we provide explanations for why each hypothesis was or was not statistically supported. Finally, we use the results relating to our hypotheses to answer our research question.

5.2.1 Inter-Rater Reliability

As seen in Table 13, the majority of Cohen's Kappa values are above 0.9000, indicating a generally very high level of agreement throughout the cleaning. However, the T1, T2, and overall values range from 0.7877 to 0.8418, which are notable exceptions. A possible reason for these relatively low values could be that since the control group consistently fared worse at recall than the RADAR group especially at T1 and T2, it became harder to discern when the control group participants understood the deeper meanings of the correct answer (and should receive a 1), as opposed to when they were completely on the wrong track (and should receive a 0) or on the right track but not accurate enough (and should receive a blank). Even so, the range of values from 0.7877 to 0.8418 is still rather decent for Cohen's Kappa and indicates a somewhat high level of agreement.

5.2.2 Topic Familiarity

As stated in Section 4.2.2, the simple topics (BL and BD) have familiarity scores no less than 3.42 for both groups, while the four difficult topics (CN, CF, KS, and KR) all have familiarity scores no greater than 2.73 for both groups. Along with the differences in complexity of the topics' target information, the gap between the lowest score for a simple topic and the highest score for a difficult topic further justifies our classification.

5.2.3 Cognitive Performance

In this section, we will discuss the results relating to cognitive performance. In particular, we will discuss the overall descriptive statistics, the four-way and three-way RM-ANOVAs, and the results pertaining to each hypothesis.

Overall Descriptive Statistics: There are many inferences that can be made from Table 14. For example, the observation that each row's mean and median are similar suggest that each group's data for each task at each time is generally not very skewed; this is backed up by how the greatest absolute value of skew in any row is 0.30. The observation that each row's range is no less than 0.65 indicates a great variety in participants' performance even within the same group, time, and task. This is unsurprising given how we recruited all sorts of participants regardless of their general abilities; we would expect some to fare much better than others. The observation that the standard deviations, standard errors, and ranges for recall tend to be greater than those for recognition indicate that there is more variety for recall in general. This makes sense given that recognition narrows down possible answers and decreases the variety within responses, while recall does no such thing. The observation that kurtosis values range from -1.59 to -0.62 indicate somewhat light-tailed distributions for any task done by any group at any time, meaning there are fortunately not many outliers. The observation that both minimum and maximum scores for recognition are generally higher than their recall counterparts indicates that it tends to be easier to achieve higher scores and avoid lower ones in recognition; this is supported by the means, which were used to support the hypotheses of recognition having higher scores than recall.

Four-Way RM-ANOVA: The main effects of topic, time, and task seen in Table 15 are fully expected. We had anticipated that participants' performance would vary across topics due to some being considered simple while others are considered difficult, that they would fare considerably worse in T2 than in T1 or T3, and that they would fare considerably better at recognition than recall. We had also predicted several interactions. Regarding the interaction between condition and topic, not only did we expect the RADAR group to fare better than the control group within any given topic, we expected different topics to have vastly different scores within either condition. Regarding the interaction between condition and task, we expected the RADAR group to fare better than the control group within either task, and we expected recognition to have a higher score than recall within either condition. Regarding the interaction between topic and task, we expected participants to fare better at recognition than recall within any topic, and we expected different topics to have vastly different scores within either task. Regarding the interaction between task and time, we expected participants to fare considerably worse in T2 within either task, and we expected them to fare better at recognition within each time.

Three-Way RM-ANOVAs for Recall and Recognition The main effects of topic and time within both tasks as shown in 16 is expected. Within either task, we anticipated that participants' performance would vary across topics due to some being considered simple while others are considered difficult, and that they would fare considerably worse in T2 than in T1 or T3. The interaction between condition and topic within recall is unsurprising, as not only did we expect the RADAR group to fare better at recall than the control group within any given topic, we expected different topics to have vastly different recall scores within either condition. However, the lack of statistical evidence of an interaction within recognition is surprising, as we expected similar results as those for recall. It appears that because the RADAR group did not fare significantly better at recognition overall than the control group did, this would be reflected within individual topics and counteract the major differences between topics within a given condition for recognition, thus making an interaction between condition and topic not immediately apparent.

RADAR Vs. Control for Each Topic: We had hoped that the Acronimps game would improve students' performance in all six topics because of the shown benefits of mnemonics [73], especially when paired with retrieval practice [72], but this was not the case for BL and BD. It should be noted that these two were what we considered simple topics, which are more susceptible to the ceiling effect [97]. This effect occurs when enough participants reach or approach the highest possible score on a question, likely due to the question being rather easy; there would not be much room for improvement. In our case, that would mean that there would be less room in the Biological Organization topics' questions for the Acronimps game to help participants improve upon because participants are already getting many of these questions right. Indeed, BL's control score (0.75) and BD's (0.58) are among the highest control scores. As for why the RADAR group has lower scores for BL (0.72) and BD (0.56), they could be attributed to a combination of the ceiling effect and the unfortunate circumstances regarding the recognition portion of T2 and T3 for the experimental group, namely midterm season, the game never asking players to recognize tidbits like they needed to for the test, and the prioritization of recall over recognition (extra impactful in the short 15-minute preparation for T3). These three circumstances will be elaborated upon when discussing hypotheses relating to recognition. Regardless, the differences are far from significant anyway, with $p = 0.2308$ for BL and $p = 0.8178$ for BD.

However, the Acronimps game did result in higher means for CN, CF, KS, and KR, likely because the ceiling effect [97] did not happen for them (i.e., more mistakes were made in many of these topics' questions). Thus, there would be more room for the Acronimps game to help participants improve. In particular, KS and KR, which are both about the Krebs Cycle, have the lowest control scores (0.42 and 0.44 respectively) and the smallest p-values. KS even has $p < 0.001$ which is significant and supports "R>C for KS" (though the great difference in familiarity scores for it between the two groups as shown in Figure 14 could have played a role), and KR has $p = 0.0744$ which is much closer to significance than the remaining topics, especially the simple ones). Essentially, difficult topics, which are less susceptible to the ceiling effect, appear more likely to benefit from the Acronimps game; this will be elaborated upon when discussing comparing simple topics to difficult ones.

RADAR vs. Control for Recognition and Recall at T1: The Acronimps game having a higher recognition score for T1 is expected because alongside the shown benefits of mnemonics [73] especially when paired with retrieval practice [72], the Acronimps game is centered around recognition due to its Quests being of a multiple-choice nature (other than "Museum (Artifacts Review)"), with "Castle (Decoding)" specifically asking participants about the exact target information that is involved in each test, as well as providing tidbits as feedback, many of which are also involved in each test. "Museum (Artifacts Review)" also organizes the tidbits for each topic, just as the SAVE Tool does. In addition, "Town (Recollection)", "Bridge (Association)", and "Museum (Artifacts Review)" feature the mnemonics we provided. While mnemonics have already been shown in prior research to benefit takers of memory-intensive courses [73], in our case, since several recognition questions explicitly involve order (e.g., some for KS and KR as seen in Table 6) while many others implicitly do so, the mnemonics can provide a quick reminder on what the order is.

The Acronimps game having a higher recall score for T1 (with significant results) is expected because alongside mnemonics' shown benefits [73] especially with retrieval practice [72], "Museum (Artifacts Review)" organizes all the artifacts in a similar manner to the SAVE Tool; in particular, the target information (which is exactly and solely what is involved in the recall section of each test) of each topic is arranged in a table that matches the required order. Also, all topics' artifacts are put side-by-side and require a simple horizontal scrolling to traverse from topic to topic, making it more convenient than even the SAVE Tool, which has different pages for each topic. The mnemonics' aforementioned assistance regarding order is also a big help in matching the required order; the fact that our mnemonics' terms generally match the first few letters (not just the first letter) of their corresponding target terms can also help remind participants of the target terms themselves, or at least score partial marks.

RADAR vs. Control for Recognition and Recall at T2: The Acronimps game having a lower recognition score for T2 was initially unexpected because mnemonics had been shown in prior research to be beneficial [73], particularly for long-term memory [39], and go hand-in-hand with retrieval practice [72]. However, we have found possible factors in our case that counteract this shown benefit. One possible factor is that the experimental group participated throughout four weeks of October, i.e., midterm season, while the control group participated in one week of September. It was much more difficult to recruit participants during October, as they likely prioritized preparing for midterms (and many of the ones who did sign up either rescheduled or backed out); we also recruited more participants for the experimental group instead of stopping at approximately 25 because various participants did not perform as well as we expected and could have been outliers. During the one-week gap between T1 and T2, the experimental group would have likely cast aside what they learned in Part 1 so as to concentrate on midterms, and also because we deceived them (and the control group) by claiming that Part 2 would be on a different subject so they would not actively rehearse; this deception was approved by the Office of Research Ethics. A second possible factor is that in the Acronimps game, tidbits are provided in "Museum (Artifacts Review)" and appear as feedback in "Castle (Decoding)", but players are never asked to recognize them. This absence likely detracts from the retrieval practice experience, leaving participants to only become experienced in recognizing target terms themselves instead of both target terms and tidbits. This could also explain why even at T1, the experimental group's improvement for recognition was notably less than for recall, as seen in Figure 17.

The Acronimps game having a higher recall score for T2 (with significant results) is expected because as stated before, mnemonics had been shown in prior research to help [73], particularly for long-term memory [39], and go hand-in-hand with retrieval practice [72]. Also, during Part 1, the experimental group likely spent more time recalling target information than performing recognition on any artifacts because both the SAVE Tool and "Museum (Artifacts Review)" showcase target information, and also because we explicitly told them to test themselves on target information before T1, which might have carried over to later tests. Even with the one-week gap during midterm season, the target information likely remained in their heads more than the tidbits.

RADAR vs. Control for Recognition and Recall at T3: The Acronimps game having a lower recognition score for T3 was initially unexpected because of mnemonics' shown benefits [73], particularly for long-term memory [39], and when paired with retrieval practice [72]. Yet just as in T2, there are plausible explanations. Unlike for T2, this time the one-week gap and midterms are less of a factor because participants will be focusing solely on our content during the 15-minute preparation. However, the aforementioned lack of a tidbit recognition part of the Acronips game continues to be a likely factor. In addition, the 15-

minute preparation for T3 is much shorter than the 45-minute (35 if we exclude the 10-minute slide reading) preparation for T1, making every minute extra precious. The experimental group likely prioritized recalling target information over recognizing tidbits as they likely felt that the nature of multiple-choice questions itself could help them. Moreover, I had seen various participants start from “Town (Recollection)” (the first Quest) again when moving on to the game, meaning that even less time was spent on reaching “Castle (Decoding)” and “Museum (Artifacts Review)”, where the tidbits (and target information) are featured. As such, tidbits were treated as low-priority during preparation.

The Acronimps game having a higher recall score is expected because alongside mnemonics’ shown benefits [73] especially for long-term memory [39] and when paired with retrieval practice [72], the experimental group likely prioritized recalling target information even more than they did in Part 1, for two main reasons. First, the preparation time had been reduced from 35 minutes to 15, meaning they needed to put greater effort in a short amount of time. Second, they knew that the recall section consists of all target terms and nothing more, as opposed to the recognition section, where they could not necessarily remember many of the questions there due to their length, complexity, and variety.

RADAR vs. Control for Recognition and Recall Overall: The Acronimps game having a lower recognition score overall (as well as in T2 and T3) results in a rather ironic situation; on multiple levels, the Acronimps game, whose Quests consist primarily of recognition instead of recall due to their multiple-choice nature, fails to help the experimental group gain a significant improvement over the control group when it comes to recognition, while the experimental group shows a significant improvement over the control group on multiple levels when it comes to recall. It is also surprising because of mnemonics’ shown benefits [73] especially regarding long-term memory [39] and retrieval practice [72]. However, the reasons for this are a combination of those from T2 and T3: a one-week gap during midterm season, players not being asked to recognize tidbits themselves, target information being prioritized over tidbits, and game time spent on Quests not involving tidbits.

The Acronimps game having a higher recall score overall (with significant results) is certainly expected of mnemonics’ shown benefits [73] especially for long-term memory [39] and with retrieval practice [72] alongside the game having higher recall scores for the three individual times. Even Table 17 shows notable gaps between the experimental and control groups for T1 and T2, and to a lesser extent, T3. The reasons for this significant difference are a combination of those from T1, T2, and T3: the shown benefits of mnemonics (including their help for order), the fact that the SAVE Tool and “Museum (Artifacts Review)” both showcase target information, and the prioritization of target information (both from our explicit instruction and participants’ knowledge that all target information is on the test).

Difference Between RADAR and Control for Difficult Topics Vs. for Simple Topics: The difference between the two groups not being shown to be greater for difficult topics overall was initially unexpected because we had intended for the ceiling effect to occur only for the simple topics [97]. In our case, we expected for both groups to be able to correctly answer many of the simple topics’ questions, but for the control group to incorrectly answer many of the difficult topics’ questions and the experimental group to "improve" on them by correctly answering more of them. Yet it seems this was not what happened. However, the overall result could be attributed to the situation involving recognition. Though the difference not being greater for recognition was likewise initially unexpected for the same reason, it has quickly become clear that because recognition provides a limited set of possible answers to choose from, including the correct one which may have visible clues, there would be less variety in the answers, allowing more participants to select the right one regardless of topic difficulty. In addition, prior research has shown that recognition is easier than recall [98], meaning there would not be as much room for improving upon questions involving recognition anyway. Essentially, the Acronimps game has less of a chance to help raise scores if other factors are already doing so.

The difference being shown to be greater for recall (with significant results) is expected because we had anticipated for the control group to still perform well on simple topics' questions even without recognition narrowing down their choices or cluing them in on the correct answer, but not for difficult topics' questions. Because participants would more likely be struggling on the difficult topics' questions without any aid, there would be more space for improvement, and the Acronimps game can seize the bigger opportunity to help them improve. It is likely that for recall, the ceiling effect did indeed happen for only the simple topics.

Recognition vs. Recall: Recognition being shown to have higher scores than recall overall and within each group is expected, as recognition has been shown to be easier than recall [98]. Recognition provides the correct answer and all that needs to be done is to select it, while recall requires reconstructing the correct answer in its entirety, here without aid. As such, it is expected for a simpler task to have a better score. Even if participants prioritized preparing for recall, the fact that they needed to recall numerous terms they may not have heard of before for the correct topic, in the correct order, and with the correct spelling, makes it extremely unlikely that enough of them would have kept faring around as well at recall (let alone better) as they did for recognition throughout the study. Condition factor is also clearly irrelevant to the fact that recognition is an easier and simpler task than recall.

5.2.4 UX Variables

In this section, we discuss the results regarding the five RADAR UX variables. In particular, we discuss reasons for why each UX-based hypothesis was supported with significant results.

Perceived Ease of Use, Perceived Enjoyment, Perceived Usefulness, and Perceived Trustworthiness: Participants likely found the Acronimps game easy to use because of its implementations of Narrative, Feedback, and Interactivity from IN-3PACIFIC [1]. The step-based setup followed by the story ensuring a fixed order of Quests (Narrative) makes the game rather linear and easy to follow. The Feedback after each answer, as well as the game's teachings on the various Acronimps' weaknesses and guidance on where to go next (Interactivity) also help the player avoiding getting confused. Also, participants watched a tutorial video on it and received an instruction manual while playing the game. Even if someone tries to argue that the game would be harder to use without the tutorial video and instruction manual, they are forgetting that the Acronimps game is a complex tool, and forcing someone to use a complex tool without any guidance is a terrible idea.

Though we did not analyze the UX survey question, "What feature from the Acronimps game do you like the most?", a quick look at some of the participants' responses already shows several notable reasons why they found the game enjoyable. One such reason is the story (Narrative [1]); one participant liked "the game story" and another "enjoyed the story of the Acronimps, it was very creative and very suitable." Another is the XP-based ranking system (Collaboration With Others [2]); one participant said "I like the reward system (XP and HP)", another liked "the point system", and a third "really enjoyed the leaderboard feature." A third reason is "Castle (Decoding)" (possibly Challenge [1] [2]); two participants enjoyed "decoding", a third enjoyed "the last stage of the game with the tidbits (when answered correctly)", and a fourth enjoyed "the decoding stage, where I have to recall what each letter in the acronym stands for."

Participants could have found the Acronimps game useful because of its already-perceived ease of use. They could have also acknowledged its adaptability to multiple courses and multiple modules within a course (Agency/Control [1] and Learner Empowerment [2]), given that in the "Course" screen, they had seen a list of multiple unrelated courses. Besides that, "Castle (Decoding)", which was liked by several participants as stated above, clearly contains the exact content that will be on the test, and so does "Museum (Artifacts Review)." Provided that participants have unlocked these Quests, they can go to them to quickly either practice or read them.

Participants could have found the Acronimps game trustworthy because of its already-perceived ease of use and usefulness. If they thought it was easy to use and useful, they would likely trust that this ease of use and usefulness would not hinder the game from helping them improve. In addition, they might feel like they can trust that the complexity of "Castle (Decoding)" (Challenge [1] [2]) and the twists and random nature of the Quests in general (Spontaneity/Unpredictability [2]) would prevent them from being caught-off guard by the length of the test and the questions being unknown until they appear. Moreover, participants might think that since this game is made in academia for academia, it could be trusted more.

Students' Willingness to Adopt RADAR: The provided reasons for the Acronimps game's PEOU (i.e., linear nature and training), PE (i.e., story, ranking system, decoding), PU (ease of use, adaptability, clear presentation of tested content), and PT (preparation for challenges and the unknown), all of which were based on IN-3PACIFIC [1] and C-SHARP [2], could also contribute to students' willingness to adopt it. Also, some participants may have felt that they would have done worse on the tests we gave them if they had not been given the game, though this could be purely psychological and without a known basis. In addition, participants could be more willing to adopt a tool if a professor, a trusted individual, was the one to give it to them, which they might expect to be the case.

5.2.5 RQ: RADAR Game's Relative Improvement in Knowledge Acquisition and Retention

Table 25 shows each hypothesis and whether it is statistically supported. Using the results associated with specific hypotheses, we were able to gauge the Acronimps game's relative improvement in knowledge acquisition and retention. The Acronimps game having higher scores for CN, CF, KS, and KR, but not for BL or BD, suggests that it tends to foster better performance mainly on difficult topics. The Acronimps game having higher scores for both recognition and recall at T1 shows that it indeed improves knowledge acquisition, though it allows participants to express it better through recall. The Acronimps game having a higher score for recall but not recognition at T2 shows that in the context of a re-test after a prolonged period of time and presumed lack of practice, the Acronimps game improves knowledge retention, but only allows participants to express the improvement through recall. However, Figure 17 suggests the relative improvement in recall is even greater in this instance of knowledge retention (T2) than it was for knowledge acquisition (T1). The Acronimps game having a higher score for recall but not recognition at T3 suggests that in the context of giving students the exact same test once more after a brief period of revising the content, the Acronimps game also improves knowledge retention, but only allows participants to express the improvement through recall; even so, Figure 17 suggests the relative improvement in recall is minimal here (T3). The Acronimps game having a higher score for recall but not recognition overall, along with Figure 17, shows that considering a series of repeated but identical tests to evaluate both knowledge acquisition and retention (after a prolonged period of time without practice and after a brief revision), when treating the series as a whole, the Acronimps game provides improvement but allows users to express it better (if at all) with recall. The difference between the two groups' recall scores being greater for difficult topics shows that the Acronimps game works significantly better at recall for difficult topics. Recognition having higher scores than recall overall and within each group shows that participants generally perform better at recognition than recall, whether they were playing the Acronimps game or not. The actual response distributions for the first four UX variables having significantly higher means than their respective normal distributions shows students view the Acronimps game as easy to use, enjoyable, useful, and trustworthy, which could give the game a greater opportunity to improve students' performance since they feel they can rely on it. The actual response distributions for WTA having significantly higher means than their respective normal distributions shows that students will be willing to adopt the Acronimps game, which could give it an even greater opportunity to improve their performance.

5.3 Limitations and Future Work

As with any other study, ours has limitations. First, as stated several times before, the midterm season greatly interfered with our study, as it discouraged students from signing up, caused many who did to reschedule or back out, and likely prevented those who did arrive from being at their best. As a result, we spent four weeks gathering participants for the experimental group as opposed to one for the control group. This is compounded by how prior research has shown that sustained attention and motivation are higher earlier in the semester than later [130] and the control group participated before midterm season, not after; as such, the experimental group had a twofold disadvantage based on when they participated. Second, some hypotheses were not supported with significant differences despite the means following their direction. That said, we did explain that we still have reason to believe them. Third, the fact that my Acronimps game included recognizing target information but not tidbits meant that participants were not as properly prepared for the exact questions in the recognition section of each test; maybe providing participants the opportunity could lead them to have higher recognition scores than the control group for T2, T3, and overall. Fourth, despite the rules we laid out for cleaning the recall response data, the cleaning might not be perfect, due to human

Table 25: Support of hypotheses (CN: Cranial Nerves, CF: Cranial Nerves Functions, BL: Biological Organization Levels, BD: Biological Organization Disciplines, KS: Krebs Cycle Substrates, KR: Krebs Cycle Reactions, MFHD: Mean Follows Hypothesis's Direction, DS: Difference Significant).

Hypothesis	MFHD	DS
The RADAR game will lead to higher students' performance in each of the six topics (CN, CF, BL, BD, KS, KR) than the control condition because of the benefits of mnemonics [73], especially when paired with retrieval practice [72].	✓✓××✓✓	××××✓×
The RADAR game will lead to improved performance in recognition and recall tasks at Time 1 compared to the control condition because of the benefits of mnemonics [73], especially when paired with retrieval practice [72].	✓✓	×✓
The RADAR game will lead to improved performance in recognition and recall tasks at Time 2 compared to the control condition because of the benefits of mnemonics [73], particularly regarding long-term memory [39], and pairing them with retrieval practice [72].	×✓	×✓
The RADAR game will lead to improved performance in recognition and recall tasks at Time 3 compared to the control condition because of the benefits of mnemonics [73], particularly regarding long-term memory [39], and pairing them with retrieval practice [72].	×✓	××
The RADAR game will lead to improved performance in recognition and recall tasks overall (regardless of time) compared to the control condition because of the benefits of mnemonics [73], particularly regarding long-term memory [39], and pairing them with retrieval practice [72].	×✓	×✓
The difference between the two groups' scores will be greater in difficult topics than in simple topics overall and within both recognition and recall tasks because of the ceiling effect [97].	××✓	××✓
The recognition task will have a higher score than the recall task overall and within both the control and experimental groups because recognition is easier than recall [98].	✓✓✓	✓✓✓
The RADAR game will be perceived as easy to use, enjoyable, useful, and trustworthy because it follows two established frameworks we call IN-3PACIFIC [1] and C-SHARP [2].	✓✓✓✓	✓✓✓✓
Students will be willing to adopt the RADAR game given its knowledge consolidation benefit, perceived ease of use, perceived enjoyment, perceived usefulness, and perceived trustworthiness.	✓	✓

error and subjectivity regarding cases not accounted for by the rules, especially with the sheer amount of responses that need to be checked. Even a change to the judgment of one response can change whether there is a significant difference or not. However, the errors and subjectivity cut across both the control and experimental groups, and may arguably cancel out. Fifth, the Acronimps game was not alone, but rather paired with the SAVE Tool. Though this was necessary as the game fetches data inputted with the SAVE Tool and thus depends on the SAVE Tool, some of the improvements that the experimental group showed over the control group could be attributed to the SAVE Tool instead, putting the benefits of the game further into question. Sixth, some participants in the control group were seen using mnemonics during preparation even though we had not taught them to; there may also be some we did not notice. These participants could have raised the means in the control group and diminished the difference between the group using mnemonics in the Acronimps game and SAVE Tool, and the group that we had not expected to use mnemonics at all; this raises the possibility that the Acronimps game could be more beneficial than it currently seems (if we were to compare it against a group of participants who absolutely did not use mnemonics at all). Seventh, when we had replaced a participant's score for a particular task with the group mean (ignoring their own score) for it (whether due to a lack of response here but existent responses elsewhere or due to having less than 0.1 for a T1 task or 0.2 for a T3 task), this could have been inaccurate and led to a rather different result than if there were no change, especially with the small sample sizes for the two groups. However, it can be argued that leaving the scores unchanged could also be inaccurate as they may have been outliers, and we did explain that we did not want to disregard the participants' entire data just because of one odd instance inconsistent with their typical behaviour. Also, the same judgment was applied to both groups, meaning it was unbiased towards one.

Like other studies, ours can pave the way for future work. First, the idea of a quiz notification system could be implemented, whether as an additional Quest in the Acronimps game or as a separate tool, to account for the current lack of tidbit recognition. Second, there could be the implementation of an automated system to clean response data to eliminate human error, though it may be difficult to account for all possible cases and adapt the system to tests in other situations. Third, the data from all the surveys, such as demographic information and opinions, can be analyzed to provide further insight on whether demographics may play a role in performance, as well as how students generally feel about the Acronimps game (and the SAVE Tool) and mnemonics in general. For example, the predominance of South Asians in both groups could be compared against the other races/ethnicities to check for possible correlations between them and performance. Fourth, there could be a comparison between a group of participants using only the SAVE Tool and a group using both the SAVE Tool and the Acronimps game, to determine how much the game supports the SAVE Tool. Fifth, other variations of this study can be conducted, such as with a different subject, with a comparison between instructor-created mnemonics and self-created mnemonics, or with a comparison between several RADAR games when more are created. Sixth, within the Acronimps game itself, the implementation of Multisolution-Based and Co-Construction of Meaning from C-SHARP [2] could be achieved through incorporating the remaining cognitive tasks for the former, and including a chat function and a mnemonic sharing function for the latter. Seventh, this study can be repeated but with a larger control group (such as $n = 50$) that can be split into a sub-group of participants who used mnemonics (though still without the Acronimps game) and a sub-group of those who did not. While repeating the study or conducting a variant of it, sessions can also alternate between groups instead of having one group finish before another begins, to alleviate the factor of sustained attention and motivation being higher earlier in a semester [130].

6 Conclusion

The Acronimps game, as far as we know, is the first mnemonics serious game to implement the majority of the IN-3PACIFIC and C-SHARP game characteristics proposed by Blumberg et al. [1] and Morris [2], unlike other mnemonics serious games that relegate these user-experience features to the background [96]. For example, while most previous games lack a narrative, the Acronimps game has a clearly established storyline involving traveling through areas and hunting down Acronimps, which are mnemonic proxies to be recollected and decoded. In addition, it implements and validates the RADAR framework proposed by Oyibo [6]. More importantly, it allows players to engage in the retrieval of self-created mnemonics and their associated information rather than those created by designers, researchers or instructors, which are the hallmark of most prior mnemonics-based games in the academic literature and/or on the market [95].

To the best of our knowledge, the SANKOFA toolkit [6], comprising the SAVE Tool [4] and the RADAR-based Acronimps game, is the first known mnemonics-based toolkit that is more than just a single tool for either mnemonics creation or retrieval practice. Rather, it is a holistic toolkit for both mnemonics creation and retrieval practice. In this thesis, we validated the SANKOFA framework, upon which the SAVE Tool and RADAR-based Acronimps game are grounded, using a biology-lecture-based experiment. We found that the Acronimps game led to higher differential performance in difficult topics such as Krebs Cycle than simple topics such as Biology Organization. Specifically, compared with the control condition, the RADAR-based Acronimps game significantly improved recall in the first two of three repeated tests across three points in time aimed to evaluate knowledge acquisition (immediately after studying the lecture material) and retention (after one week of no practice and after a 15-minute revision). The game also improved recognition, but only in the first test. Comparatively, students generally performed better in recognition than in recall, regardless of what method they used to prepare, with no significant difference between the experimental and control groups for most of the topics, indicating the use of mnemonics may not hamper higher cognitive processes such as understanding, which the recognition task based on multiple choice questions tested. Moreover, students found the game easy to use, enjoyable, useful, and trustworthy; as a result, they were willing to adopt it.

The statistically significant benefits of the Acronimps game in recall, as well as the demonstrated benefits of mnemonics in memory-intensive courses in general [73] and in the long run [39], serve as empirical evidence for instructors of memory-intensive courses to adopt and learn mnemonic techniques, teach them to their students, and incorporate the SANKOFA toolkit or approach to learning into their curriculum to help combat the current cognitive decline experienced in education as a result of the advent of AI tools such as ChatGPT. Developers could also contribute to the SANKOFA toolkit by building upon the Acronimps game, creating their own RADAR game, or implementing quizzes to enhance retrieval practice. Finally, the Acronimps game's strength in improving recall may help users prepare for urgent situations where they must recall and apply what they have learned without aid, such as in a medical emergency in which a guidebook is not available. To crown it all, the RADAR game has the potential to improve students' performance in memory-intensive courses. As such, it should not be shelved or ignored by the academic community.

Bibliography

[1] F. C. Blumberg, D. E. Almonte, J. S. Anthony, and N. Hashimoto, "Serious Games: What Are They? What Do They Do? Why Should We Play Them?," in *The Oxford Handbook of Media Psychology*, Oxford University Press, 12 2012.

[2] T. H. Morris, "Experiential learning–a systematic review and revision of Kolb's model," *Interactive learning environments*, vol. 28, no. 8, pp. 1064–1077, 2020.

[3] K. Oyibo, "The effect of gender and mnemonics generator on perceived difficulty, utility and adoption of mnemonics-based learning tools," 2025 (Unpublished).

[4] S. Ejaz, "Designing an interactive tool for mnemonics creation and knowledge retention," *York University*, 2025 (Unpublished).

[5] J. Stayanchi, "Higher Order Thinking through Bloom ' s Taxonomy," *Humanities Review*, vol. 22, pp. 117–124, 2017.

[6] K. Oyibo, "SANKOFA: An interactive user-centred pedagogical toolkit to promote long-term learning," *York University*, 2024 (Unpublished).

[7] Climate Justice Alliance, "Sankofa bird," Aug 2023. https://climatejusticealliance.org/workgroup/black-caucus/sankofa-bird-2/.

[8] pngimg.com, "Download png image: Brain png." https://pngimg.com/image/86552.

[9] Freepik. https://freepik.com/icons.

[10] VectorPortal, "Education free icon." https://flaticon.com/free-icon/education_13376202.

[11] Bsquare. https://bsquare.in/computerized-maintenance-management-system.html.

[12] Google, Aug 2018. https://github.com/googlefonts/noto-emoji/blob/v2018-08-10-unicode11/svg/emoji_u1f3ae.svg.

[13] bsd studio, "Online quiz color icon stock illustration." https://istockphoto.com/vector/online-quiz-color-icon-gm1172274908-325128869.

[14] prescott10, "Rubik's cube — stock editorial photography." https://depositphotos.com/editorial/rubiks-cube-5016489.html.

[15] Freepik, "Aged paper texture background design." https://freepik.com/free-vector/aged-paper-texture-background-design_14765966.htm.

[16] Freepik, "Paperboard yellow texture." https://freepik.com/free-photo/paperboard-yellow-texture_5487782.htm.

[17] Freepik, "Light brown wooden textured background." https://freepik.com/free-photo/light-brown-wooden-textured-background_3743392.htm.

[18] Netflix, "The classics." https://netflix.com.

[19] Onyxprj, "Cells collection. human blood structure micro types of anatomy science vector collection cells set. illustration blood structure, human cell microbiology." https://dreamstime.com/cells-collection-human-blood-structure-micro-types-anatomy-science-vector-collection-cells-set-cells-collection-human-blood-image180777875.

[20] Net Vector, "Web design, user interface ui and user experience ux content organization. web design development concept. vector isolated concept illustration. 3d liquid design with floral elements," Mar

2019. https://shutterstock.com/image-vector/web-design-user-interface-ui-experience-1331260901.

[21] Flanoz, "A placeholder svg," Sep 2020.

[22] Siberian Art, "Budgeting for home maintenance, renovation concept vector isometric illustration. home repairs budget composition for web banner, website page etc," Sep 2019. https://shutterstock.com/image-vector/budgeting-home-maintenance-renovation-concept-vector-1495980860.

[23] N. Kozhedub, "Flat design with people. hci - human computer interface acronym, business concept background. vector illustration for website banner, marketing materials, business presentation, online advertising," Jan 2021. https://istockphoto.com/vector/flat-design-with-people-hci-human-computer-interface-acronym-business-concept-gm1296637158-389999061.

[24] vhv.rs, "Hot vector fire text - free download fire vector, hd png download." https://vhv.rs/viewpic/hbmJomx_hot-vector-fire-text-free-download-fire-vector.

[25] Medieval Chronicles, "Medieval houses close together on fire," Dec 2023. https://medievalchronicles.com/wp-content/uploads/2023/12/medieval_houses_close_together_on_fire 3.png.

[26] Material Design, "Material 3 design kit." https://figma.com/community/file/1035203688168086460.

[27] Figma, "Simple design system." https://figma.com/community/file/1380235722331273046.

[28] Halal Lab, "Award icons - free download (SVG icons)." https://figma.com/community/file/1400438751131416829.

[29] Freepik, "Fenestrelle's fort entry, Italy." https://freepik.com/free-photo/fenestrelle-s-fort-entry-italy_9183883.htm.

[30] Freepik, "Nature landscape with castle building." https://freepik.com/free-ai-image/nature-landscape-with-castle-building_94958137.htm.

[31] IndiaMART, "Container floor image." https://indiamart.com/proddetail/container-floor-22374296988.html.

[32] Flaticon, "Missile/nuke icon." https://flaticon.com/free-icon/missile.

[33] VectorStock, "Brain and mind vector illustration." https://vectorstock.com/royalty-free-vector/intellectual-concepts-brain-mind-vector-3899306.

[34] A. Helmenstine, "Levels of organization in biology," Feb 2024. https://sciencenotes.org/levels-of-organization-in-biology.

[35] Vecteezy, "Lion pride photos & images." https://vecteezy.com/free-photos/lion-pride.

[36] J. Michael, "What makes physiology hard for students to learn? results of a faculty survey," *Advances in physiology education*, vol. 31, no. 1, pp. 34–40, 2007.

[37] M. Gruneberg, "The role of memorization techniques in finals examination preparation–a study of psychology students," *Educational Research*, vol. 15, no. 2, pp. 134–139, 1973.

[38] D. D. Massey, "Reading history: Moving from memorizing facts to critical thinking," *Improving reading comprehension of middle and high school students*, pp. 19–47, 2015.

[39] J. A. McCabe, K. L. Osha, J. A. Roche, and J. A. Susser, "Psychology students' knowledge and use of mnemonics," *Teaching of psychology*, vol. 40, no. 3, pp. 183–192, 2013.

[40] D. R. Stalder, “Learning and motivational benefits of acronym use in introductory psychology,” *Teaching of Psychology*, vol. 32, no. 4, pp. 222–228, 2005.

[41] D. Agnes and R. Srinivasan, “Fostering vocabulary memorization: Exploring the impact of ai-generated mnemonic keywords on vocabulary learning through anki flashcards,” *World Journal of English Language*, vol. 14, no. 2, 2024.

[42] L. M. Rasiban, D. Sudana, and D. Sutedi, “Using mnemonic-based applications to learning Japanese hiragana characters,” in *Eleventh Conference on Applied Linguistics (CONAPLIN 2018)*, pp. 12–16, Atlantis Press, 2019.

[43] N. M. D. Sintadewi, N. P. J. Artini, and I. Febryan, “Analysis of English learning difficulty of students in elementary school,” *International Journal of Elementary Education*, vol. 4, no. 3, pp. 431–438, 2020.

[44] S. S. Seay and H. G. McAlum, “The use/application of mnemonics as a pedagogical tool in auditing,” *Academy of Educational Leadership Journal*, vol. 14, no. 2, p. 33, 2010.

[45] C. Chen, K. Z. Zhang, X. Gong, M. K. Lee, and Y.-Y. Wang, “Preventing relapse to information technology addiction through weakening reinforcement: A self-regulation perspective,” *Information & Management*, vol. 58, no. 5, p. 103485, 2021.

[46] R. Hall, “The work–study relationship: experiences of full-time university students undertaking part-time employment,” *Journal of education and Work*, vol. 23, no. 5, pp. 439–449, 2010.

[47] B. M. Hard, J. M. Lovett, and S. T. Brady, “What do students remember about introductory psychology, years later?,” *Scholarship of Teaching and Learning in Psychology*, vol. 5, no. 1, p. 61, 2019.

[48] G. Plancher and P. Piolino, “Virtual reality for assessment of episodic memory in normal and pathological aging,” *The role of technology in clinical neuropsychology*, pp. 237–260, 2017.

[49] R. C. Atkinson and R. M. Shiffrin, “Human memory: A proposed system and its control processes,” in *Psychology of learning and motivation*, vol. 2, pp. 89–195, Elsevier, 1968.

[50] E. F. Loftus, *Memory*. Rowman & Littlefield Publishers, 1988.

[51] J. M. Gardiner, B. Gawlik, and A. Richardson-Klavehn, “Maintenance rehearsal affects knowing, not remembering; elaborative rehearsal affects remembering, not knowing,” *Psychonomic Bulletin & Review*, vol. 1, pp. 107–110, 1994.

[52] E. Jensen, *Teaching with the brain in mind*. ASCD, 2005.

[53] J. A. McCabe, “Integrating mnemonics into psychology instruction Jennifer A. McCabe, Goucher College 2010 Instructional Resource Award recipient,” 2011.

[54] P. R. Pintrich, “The role of metacognitive knowledge in learning, teaching, and assessing,” *Theory into practice*, vol. 41, no. 4, pp. 219–225, 2002.

[55] K. D. Tanner, “Promoting student metacognition,” *CBE—Life Sciences Education*, vol. 11, no. 2, pp. 113–120, 2012.

[56] W. R. Balch, “Elaborations of introductory psychology terms: Effects on test performance and subjective ratings,” *Teaching of Psychology*, vol. 32, no. 1, pp. 29–34, 2005.

[57] M. Weimer, “Deep learning vs. surface learning: Getting students to understand the difference,” 2012.

[58] J. D. Bransford, R. R. Cocking, and A. L. Brown, *How People Learn: Brain, Mind, Experience, and*

School: Expanded Edition. Washington, DC: The National Academies Press, 2000.

[59] X. Hao and X. Gu, “Analysis of key roles in large-scale online learning: Interactive participation characteristics and knowledge construction behaviour patterns,” *British Journal of Educational Technology*, vol. 55, no. 3, pp. 910–932, 2024.

[60] L. Harasim, *Learning theory and online technologies*. Routledge, 2017.

[61] K. R. Clark, “Learning theories: constructivism,” 2018.

[62] K. Y. L. Ku, S. Phillipson, and S. N. Phillipson, *Educational learning theory*, vol. 7, pp. 238–245. Netherlands: Elsevier, 2nd ed., mar 2015.

[63] S. R. Gundu, T. Chundru, V. R. Manipatruni, and V. R. R. Karibandi, “Education system on Indian campuses-a review,” *EPRA International Journal of Multidisciplinary Research (IJMR)*, vol. 10, no. 9, pp. 295–303, 2023.

[64] M. de Sousa Oliveira, “Procedural/rote learning vs meaningful learning: Which do first year undergraduate students prefer?,” in *EDULEARN23 Proceedings*, pp. 1681–1690, IATED, 2023.

[65] J. Elliott and C.-t. Tsai, “What might Confucius have to say about action research?,” *Educational Action Research*, vol. 16, no. 4, pp. 569–578, 2008.

[66] P. Phokha, C. Saepueng, and S. Jirotkul, “A Kod-Orffian methodology: Development of music education for the Thai elementary context.,” *International Journal of Learner Diversity & Identities*, vol. 30, no. 1, 2023.

[67] O. G. Keehl, *Evidence-Based Gamification and Mnemonics for Logographic Writing Systems*. University of California, Santa Cruz, 2023.

[68] C. Wiklund-Hörnqvist, S. Stillesjö, M. Andersson, B. Jonsson, and L. Nyberg, “Retrieval practice facilitates learning by strengthening processing in both the anterior and posterior hippocampus,” *Brain and behavior*, vol. 11, no. 1, p. e01909, 2021.

[69] J. D. Karpicke and J. R. Blunt, “Retrieval practice produces more learning than elaborative studying with concept mapping,” *Science*, vol. 331, no. 6018, pp. 772–775, 2011.

[70] J. D. Karpicke and H. L. Roediger III, “The critical importance of retrieval for learning,” *science*, vol. 319, no. 5865, pp. 966–968, 2008.

[71] B. Pastötter and K.-H. T. Bäuml, “Retrieval practice enhances new learning: The forward effect of testing,” *Frontiers in psychology*, vol. 5, p. 83305, 2014.

[72] H. L. Roediger and A. C. Butler, “The critical role of retrieval practice in long-term retention,” *Trends in cognitive sciences*, vol. 15, no. 1, pp. 20–27, 2011.

[73] M. A. Mastropieri and T. E. Scruggs, “Enhancing school success with mnemonic strategies,” *Intervention in school and clinic*, vol. 33, no. 4, pp. 201–208, 1998.

[74] H. Dave and S. Awasthi, “An investigation of the role of mnemonics in higher education,” in *Proceedings of International Conference on Digital Pedagogies (ICDP)*, 2019.

[75] J. G. Tullis and J. Qiu, “Generating mnemonics boosts recall of chemistry information.,” *Journal of Experimental Psychology: Applied*, vol. 28, no. 1, p. 71, 2022.

[76] M. Mocko, L. M. Lesser, A. E. Wagler, and W. S. Francis, “Assessing effectiveness of mnemonics for tertiary students in a hybrid introductory statistics course,” *Journal of Statistics Education*, vol. 25,

no. 1, pp. 2–11, 2017.

[77] A. L. Putnam, "Mnemonics in education: Current research and applications.," *Translational Issues in Psychological Science*, vol. 1, no. 2, p. 130, 2015.

[78] J. B. Worthen and R. R. Hunt, *Mnemonology: Mnemonics for the 21st century*. Psychology Press, 2011.

[79] K. Jurowski, A. Jurowska, and M. Krzeczkowska, "Comprehensive review of mnemonic devices and their applications: State of the art," *International E-Journal of Science, Medicine and Education*, vol. 9, no. 3, 2015.

[80] C. Goedhart, "Activating student metacognition: How to help students learn more effectively and efficiently," Apr 2021.

[81] A. G. Setiawan and P. K. Hakim, "Student's cognitive engagement in teaching vocabulary through mnemonic strategy," *Pedagogy: Journal of English Language Teaching*, vol. 11, no. 2, pp. 230–238, 2023.

[82] R. N. Carney and J. R. Levin, "Promoting higher-order learning benefits by building lower-order mnemonic connections," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 17, no. 5, pp. 563–575, 2003.

[83] D. Economou, M. G. Russi, I. Doumanis, M. Mentzelopoulos, V. Bouki, and J. Ferguson, "Using serious games for learning british sign language combining video, enhanced interactivity, and vr technology," *Journal of Universal Computer Science*, vol. 26, no. 8, pp. 996–1016, 2020.

[84] L. M. Baker, "Fun and games: Connecting for learning," *International Journal of Disability, Development and Education*, vol. 59, no. 1, pp. 119–123, 2012.

[85] Y. M. Arif, N. Ayunda, N. M. Diah, and M. B. Garcia, "A systematic review of serious games for health education: Technology, challenges, and future directions," *Transformative Approaches to Patient Literacy and Healthcare Innovation*, pp. 20–45, 2024.

[86] G. Lampropoulos, E. Keramopoulos, K. Diamantaras, and G. Evangelidis, "Integrating augmented reality, gamification, and serious games in computer science education," *Education Sciences*, vol. 13, no. 6, p. 618, 2023.

[87] N. Sharifzadeh, H. Kharrazi, E. Nazari, H. Tabesh, M. Edalati Khodabandeh, S. Heidari, and M. Tara, "Health education serious games targeting health care providers, patients, and public health users: scoping review," *JMIR serious games*, vol. 8, no. 1, p. e13459, 2020.

[88] M. Graafland, J. M. Schraagen, and M. P. Schijven, "Systematic review of serious games for medical education and surgical skills training," *Journal of British Surgery*, vol. 99, no. 10, pp. 1322–1330, 2012.

[89] V. J. Traver, L. A. Leiva, V. Mart´i-Centelles, and J. Rubio-Magnieto, "Educational videogame to learn the periodic table: design rationale and lessons learned," *Journal of chemical education*, vol. 98, no. 7, pp. 2298–2306, 2021.

[90] N. Micallef and N. A. G. Arachchilage, "A serious game design: Nudging users' memorability of security questions," *arXiv preprint arXiv:1709.08167*, 2017.

[91] S. Câmara Olim, V. Nisi, and T. Romão, "Periodic fable discovery: an augmented reality serious game

to introduce and motivate young children towards chemistry," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 52593–52619, 2024.

[92] A. De Gloria, F. Bellotti, and R. Berta, "Serious games for education and training," *International Journal of Serious Games*, vol. 1, no. 1, 2014.

[93] M. Ullah, S. U. Amin, M. Munsif, M. M. Yamin, U. Safaev, H. Khan, S. Khan, and H. Ullah, "Serious games in science education: a systematic literature review," *Virtual Reality & Intelligent Hardware*, vol. 4, no. 3, pp. 189–209, 2022.

[94] K. Oyibo, "RADAR: A mnemonics-based framework for designing educational serious games for retrieval practice," *York University*, 2023 (Unpublished).

[95] K. Fung and K. Oyibo, "Examining the effectiveness of mnemonics serious games in enhancing memory and learning: A scoping review," *Applied Sciences*, vol. 14, no. 23, p. 11379, 2024.

[96] K. Fung and K. Oyibo, "Examining the design characteristics of mnemonics serious games on the app stores: A systematic heuristic review," *Applied Sciences*, vol. 15, no. 14, p. 7772, 2025.

[97] M. vSimkovic and B. Träuble, "Robustness of statistical methods when measure is affected by ceiling and/or floor effect," *PloS one*, vol. 14, no. 8, p. e0220889, 2019.

[98] H. L. Hollingworth, "Characteristic differences between recall and recognition," *The American Journal of Psychology*, vol. 24, no. 4, pp. 532–544, 1913.

[99] L. M. Schoen, "Mnemopoly: Board games and mnemonics," *Teaching of Psychology*, vol. 23, no. 1, pp. 30–32, 1996.

[100] A.-S. Oyen and J. M. Bebko, "The effects of computer games and lesson contexts on children's mnemonic strategies," *Journal of experimental child psychology*, vol. 62, no. 2, pp. 173–189, 1996.

[101] C.-Y. Shen, M.-P. Chen, N.-S. Chen, T.-Y. Chuang, and Y.-C. Huang, "The effectiveness of game-based learning on students' mnemonic techniques and perceptions," in *European Conference on Games Based Learning*, p. 553, Academic Conferences International Limited, 2011.

[102] J. E. McLaughlin, M. D. Wolcott, D. Hubbard, K. Umstead, and T. R. Rider, "A qualitative review of the design thinking framework in health professions education," *BMC medical education*, vol. 19, pp. 1–8, 2019.

[103] J. Salminen, K. Wenyun Guan, S.-G. Jung, and B. Jansen, "Use cases for design personas: A systematic review and new frontiers," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2022.

[104] R. Gudjonsdottir, *Personas and scenarios in use*. PhD thesis, KTH, 2010.

[105] C. Lewis and C. Wharton, "Cognitive walkthroughs," in *Handbook of human-computer interaction*, pp. 717–732, Elsevier, 1997.

[106] M. Schrepp, "User experience questionnaire handbook," *All you need to know to apply the UEQ successfully in your project*, pp. 50–52, 2015.

[107] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908, Sage publications Sage CA: Los Angeles, CA, 2006.

[108] M. H. Phan, J. R. Keebler, and B. S. Chaparro, "The development and validation of the game user

experience satisfaction scale (guess)," *Human factors*, vol. 58, no. 8, pp. 1217–1247, 2016.

[109] S. Sachdeva, "Scrum methodology," *Int. J. Eng. Comput. Sci*, vol. 5, no. 16792, pp. 16792–16800, 2016.

[110] C. N. Temple, "The emergence of Sankofa practice in the united states: A modern history," *Journal of Black Studies*, vol. 41, no. 1, pp. 127–150, 2010.

[111] P. F. Small, M. J. Barker, and M. Gasman, *Sankofa: African American perspectives on race and culture in US doctoral education*. State University of New York Press, 2020.

[112] University of Central Florida, "9 Types of Mnemonics for Better Memory," *Student Academic Resource Center*, pp. 1–13, 2012.

[113] T. S. Project, "African tradition, proverbs, and Sankofa," 2014.

[114] S. A. Smith, "Virtual reality in episodic memory research: A review," *Psychonomic bulletin & review*, vol. 26, pp. 1213–1237, 2019.

[115] R. B. S. M. Salim, J. S. Chen, H. L. Chen, and B. Rai, "Design and development of MemoryTrail virtual reality game to study brain and memory processes in a fun and interactive manner," *Journal of Applied Learning and Teaching*, vol. 6, no. Sp. Iss. 1, 2024.

[116] P. Toril, J. M. Reales, J. Mayas, and S. Ballesteros, "Video game training enhances visuospatial working memory and episodic memory in older adults," *Frontiers in human neuroscience*, vol. 10, p. 206, 2016.

[117] P. Cowan and R. Farrell, "Using virtual reality to support retrieval practice in blended learning: An interdisciplinary professional development collaboration between novice and expert teachers," *Digital*, vol. 3, no. 3, pp. 251–272, 2023.

[118] A. B. Hostetter, E. A. Penix, M. Z. Norman, W. R. Batsell Jr, and T. H. Carr, "The role of retrieval practice in memory and analogical problem-solving," *Quarterly Journal of Experimental Psychology*, vol. 72, no. 4, pp. 858–871, 2019.

[119] K. Oyibo, *Designing culture-tailored persuasive technology to promote physical activity*. PhD thesis, 2020.

[120] "Sample size calculator: Find out the sample size," 2008.

[121] K. Oyibo, "Towards mnemonic optimization and a paradigm shift to semionics and goldenomics: A semiotic perspective," *York University*, 2025 (Unpublished).

[122] K. Oyibo and P. P. Morita, "Covid alert: factors influencing the adoption of exposure notification apps among Canadian residents," *Frontiers in Digital Health*, vol. 4, p. 842661, 2022.

[123] K. Oyibo, "Comtech: Towards a unified taxonomy of persuasive techniques for persuasive technology design," *Computers in Human Behavior Reports*, vol. 14, p. 100372, 2024.

[124] J. M. Murre and J. Dros, "Replication and analysis of Ebbinghaus' forgetting curve," *PloS one*, vol. 10, no. 7, p. e0120644, 2015.

[125] K. Oyibo, "Using AI to automate mnemonics generation and evaluation in education: A scoping review.," pp. 1–17, Pan-African AI and Smart Systems Conference, 2025.

[126] K. Oyibo, "The influence of user knowledge and usage behaviour on decision-making and perceived reputation of streaming sites that use dark patterns," *Behaviour & Information Technology*, pp. 1–20,

2025.

[127] H. Wickham, “ggplot2,” *Wiley interdisciplinary reviews: computational statistics*, vol. 3, no. 2, pp. 180–185, 2011.

[128] M. Kay and J. O. Wobbrock, “Package ‘ARTool’,” *CRAN Repository*, vol. 2016, pp. 1–13, 2016.

[129] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, “The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011.

[130] M. E. Nicholls, K. M. Loveless, N. A. Thomas, T. Loetscher, and O. Churches, “Some participants may be better than others: Sustained attention and motivation are higher early in semester,” *Quarterly journal of experimental psychology*, vol. 68, no. 1, pp. 10–18, 2015.

Appendices

A Lecture Slides

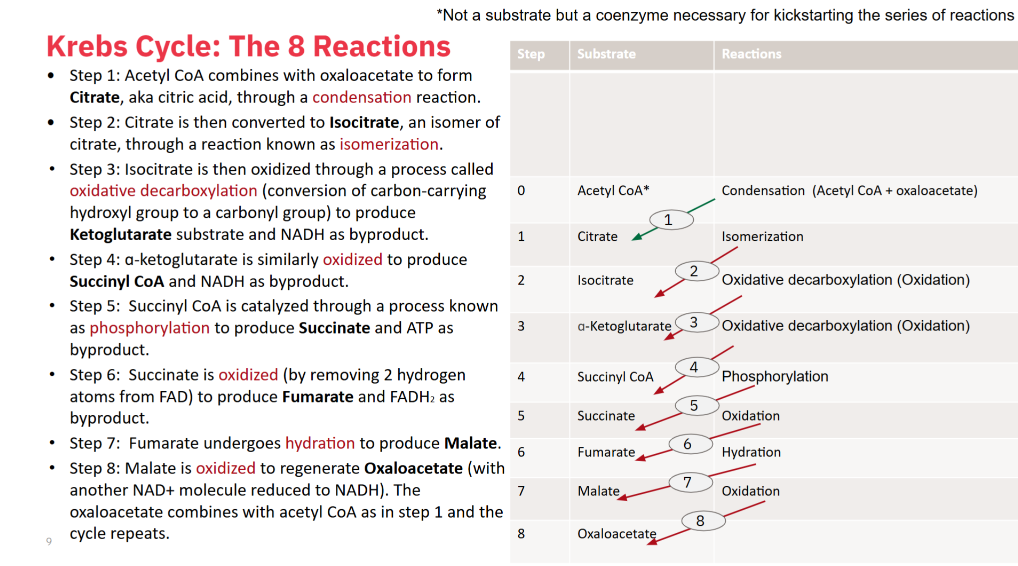
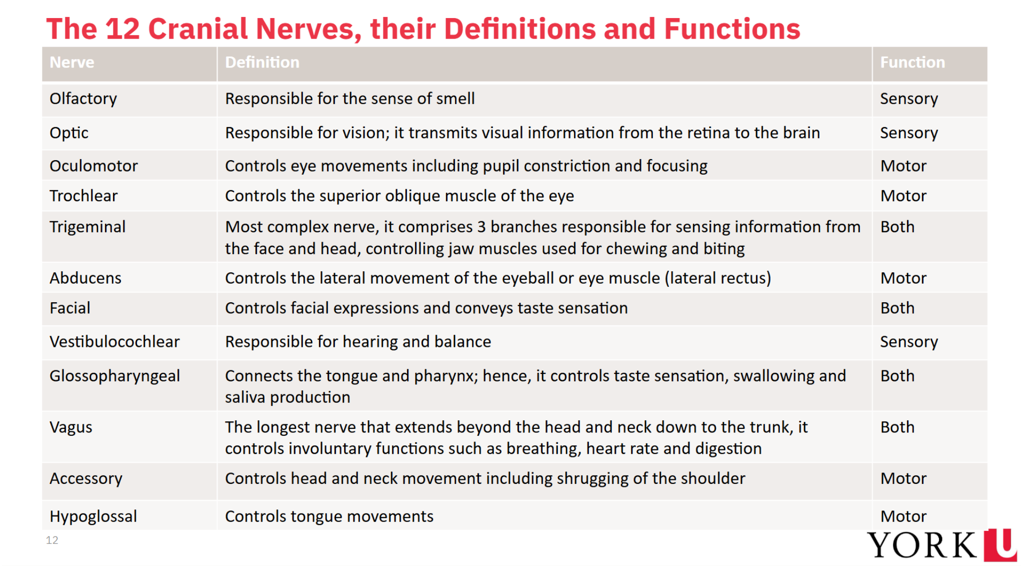


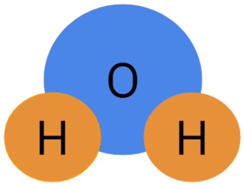
Figure A1: A slide showing details of Krebs Cycle Substrates and Reactions.

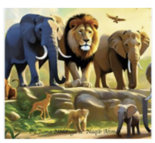


Nerve	Definition	Function
Olfactory	Responsible for the sense of smell	Sensory
Optic	Responsible for vision; it transmits visual information from the retina to the brain	Sensory
Oculomotor	Controls eye movements including pupil constriction and focusing	Motor
Trochlear	Controls the superior oblique muscle of the eye	Motor
Trigeminal	Most complex nerve, it comprises 3 branches responsible for sensing information from the face and head, controlling jaw muscles used for chewing and biting	Both
Abducens	Controls the lateral movement of the eyeball or eye muscle (lateral rectus)	Motor
Facial	Controls facial expressions and conveys taste sensation	Both
Vestibulocochlear	Responsible for hearing and balance	Sensory
Glossopharyngeal	Connects the tongue and pharynx; hence, it controls taste sensation, swallowing and saliva production	Both
Vagus	The longest nerve that extends beyond the head and neck down to the trunk, it controls involuntary functions such as breathing, heart rate and digestion	Both
Accessory	Controls head and neck movement including shrugging of the shoulder	Motor
Hypoglossal	Controls tongue movements	Motor

Figure A2: A slide showing details of Cranial Nerves and Functions.

Biological Organization: Definition





17 https://www.vecteezy.com/free-photos/lion-pride

Level	Definition
Atom	The smallest unit of chemical elements, e.g., hydrogen.
Molecule	Two or more atoms held together by chemical bonds, e.g., water.
Organelle	Specialized structure/subunit within a cell that performs specific functions, e.g., mitochondria, chloroplasts.
Cell	The basic unit of life that carries out life processes, e.g., blood cell.
Tissue	Group of similar cells performing a certain task, e.g., bone or blood, nervous, adipose tissues, etc.
Organ	Group of similar tissues performing a function, e.g., heart, lung, liver.
Organ System	Group of organs performing a function, e.g., respiratory system.
Organism	A living being or animal composed of organ systems working together, e.g., lion.
Population	A group of organisms of the **same** species living in the same area, e.g., pride of lions or people, e.g., lion pride.
Community	A group of organisms of **different** species living in the same area, e.g., lion and zebra.
Ecosystem	A group of different communities of living and non-living things interacting as a unit in their environment, e.g., lion, zebra, trees, grass, soil, etc.
Biosphere	The entire planet comprising living/non-living things and the parts of earth including animals, plants, hills, earth and sky.

Figure A3: Slides showing details of Biological Organization Levels based on [34], with lion images provided by [35].

Biological Organization: Definition & Disciplines

Level	Discipline	Definition
Atom	Chemistry	Studies the properties, composition, and behavior of atoms and how they interact to form molecules
Molecule	Molecular Biology	Examines the molecular mechanisms of biological processes, focusing on the structure and function of molecules such as DNA, RNA, and proteins
Organelle	Cellular Biology (Cytology)	Encompasses the study of cells, including their structures, functions, and organelles
Cell	Cellular Biology (Cytology)	Encompasses the study of cells, including their structures, functions, and organelles
Tissue	Histology	Examines tissues - the groups of similar cells that work together to perform specific functions
Organ	Anatomy	Examines the structure of living organisms, with a particular focus on individual organs
Organ System	Physiology	Investigates the functions of biological systems, including how organ systems operate and interact to sustain life
Organism	Biology	The science of life and living organisms, encompassing various subfields like botany, zoology, and microbiology
Population	Population Biology	The study of how populations of organisms (of the same species) change over time
Community	Community Ecology	Explores the interactions among different species that coexist in the same geographic area
Ecosystem	Ecology	Investigates the relationships between living organisms - including individuals, populations, and communities - and their physical environments
Biosphere	Biogeography	The study of the distribution of species and ecosystems in geographic space and through geological time

19 YORK U

Figure A4: Slides showing details of Biological Organization Disciplines.

Krebs Cycle: Acetyl CoA, 8 Substrates & Reactions Mnemonics

*Not a substrate but a coenzyme necessary for kickstarting the series of reactions

*Red text not part of the mnemonic, just a clarification or meaning

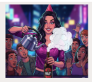
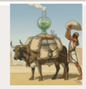
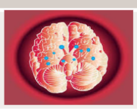
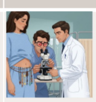
Step	Substrate	Mnemonic	Reaction	Mnemonic
				
0	Acetyl CoA*	Acetic	Condensation	Condemned
1	Citrate	City	Isomerization	Isometrician (“exerciser”)
2	Isocitrate	Isobel’s	Oxidative decarboxylation	Oxton (“name”)
3	Ketoglutarate	Kettle gluttonously	Oxidative decarboxylation	Overloaded
4	Succinyl CoA	Succinctly	Phosphorylation	Pharoah’s
5	Succinate	Sucks	Oxidation	Ox (with)
6	Fumarate	Fumy	Hydration	Hydraulic
7	Malate	Malt	Oxidation	Oxide
8	Oxaloacetate	Oxaloacetic		
	Verbal Cue (Conceptual Connection)	Imagine the kettle of Isobel, who is a city girl and "acetic" (troublesome), in place of a glass, gluttonously but succintly (gradually) sucking off the fumy malt out of the only available bottle of malt at a party in a SUBSTATE CRIB (small-state crib sounds like "Krebs' substrate"). Take note of the “Acetic” adjective and/or cue at the beginning and end of the mnemonic, which sounds like “a city,” the “ate” suffix in 7 of the 9 substrates (e.g., Isocitrate, Fumarate), and the paradox of the adverbs “Gluttonously” and “Succinctly.”	**Verbal Cue (Conceptual Connection)**	Imagine a condemned prisoner, who loves to perform isometic exercise when free, loading instead pharoah's ox with bags of hydraulic oxide while in prison as a REACTION to the nonavailability of isometric exercise equipment in prison. Notice how "Oxton", "Ox", and "Oxide" all start with "Ox" just like "Oxidative decarboxylation" and "Oxidation". "Isometrician" also sounds like "Isomerization".

Figure A5: A slide showing mnemonics for Krebs Cycle Substrates and Reactions.

12 Cranial Nerves and Functions - Mnemonics

*Red text not part of the mnemonic, just a clarification or meaning

Nerves	Mnemonic	Function	Mnemonic
			
Olfactory	Oldfactory	Sensory	Seers
Optic	Optician’s	Sensory	Say
Oculomotor	Oculi (eyes)	Motor	Move
Trochlear	Trouble	Motor	Motor
Trigeminal	Trigonometrically,	Both	But
Abducens	Abductor’s (Ab doctor’s)	Motor	Motorist
Facial	Facial	Both	Bosses
Vestibulocochlear	Vestibule (‘eyes’)	Sensory	Say
Glossopharyngeal	Glossing pharyngeally (near the neck)	Both	Big
Vagus	Vagina	Both	Brain
Accessory	Accessory	Motor	Matters
Hypoglossal	Hypoglossal (unshiny)	Motor	Most

23

Figure A6: A slide showing mnemonics for Cranial Nerves and Functions.

12 Levels of Biological Organization: Mnemonic

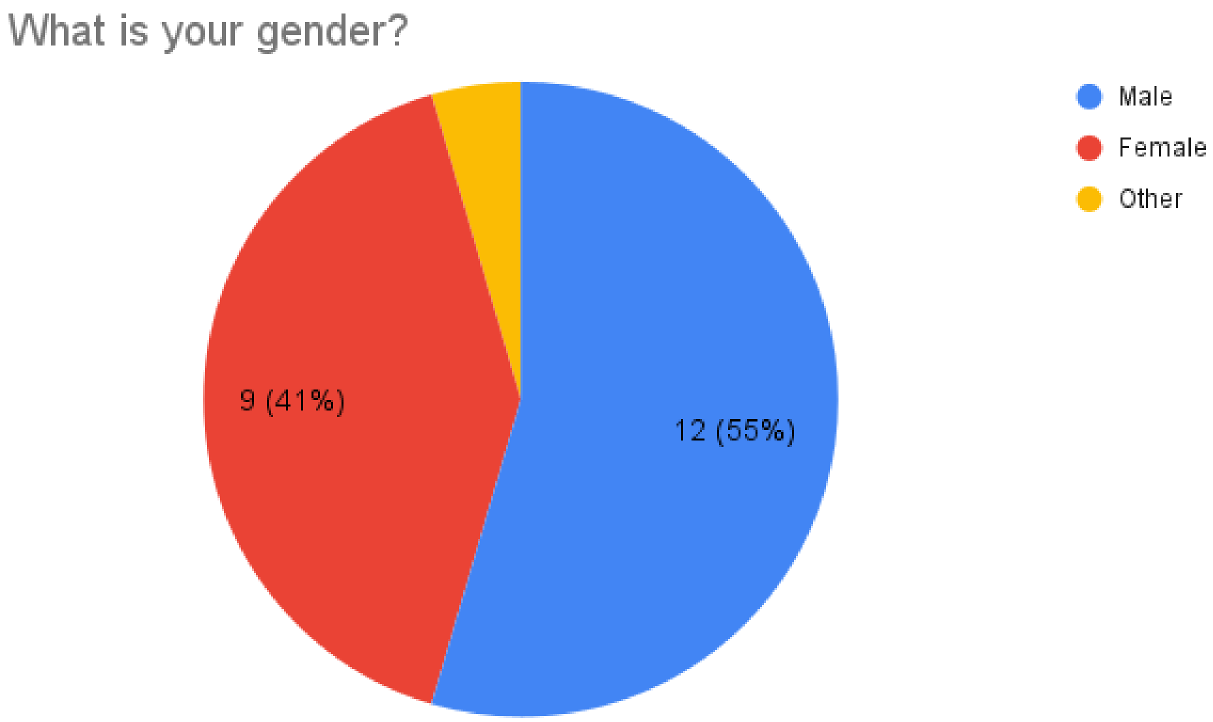
Nerves	Mnemonic	Discipline	Mnemonic
			
Atom	At	**Chemistry**	**Chemist**
Molecule	Moldova's ("country sandwiched between Romania and Ukraine")	**Mol**ecular Biology	**Mol**lahs ("Experts")
Organ**elle**	Original	Ce**ll**ular Biology (Cytology)	Ca**ll**out
Cell	Cell,	**Cell**ular Biology (Cytology)	**Cell**ular
Ti**ss**ue	Tutsi	Hi**st**ology	Hi**s**torians
Org**an**	Organized	**An**atomy	**An**alyzing
Organ **Sys**tem	Organic	Ph**ys**iology	**Phys**ical
Organism	Orgy	Biology	**B**ones
Population	Popular,	**Population** Biology	**Populating**
Community	Commanding	**Community** Ecology	**Com**pton ("Hip-Pop city in California"),
Ecosystem	Economic	**Ec**ology	**Ec**hoing
Biosphere	Boom	**Bio**geography	**Big**otry

24

Adapted from https://sciencenotes.org/levels-of-organization-in-biology/

Figure A7: A slide showing mnemonics for Biological Organization Levels and Disciplines based on [34].

B Demographic Data



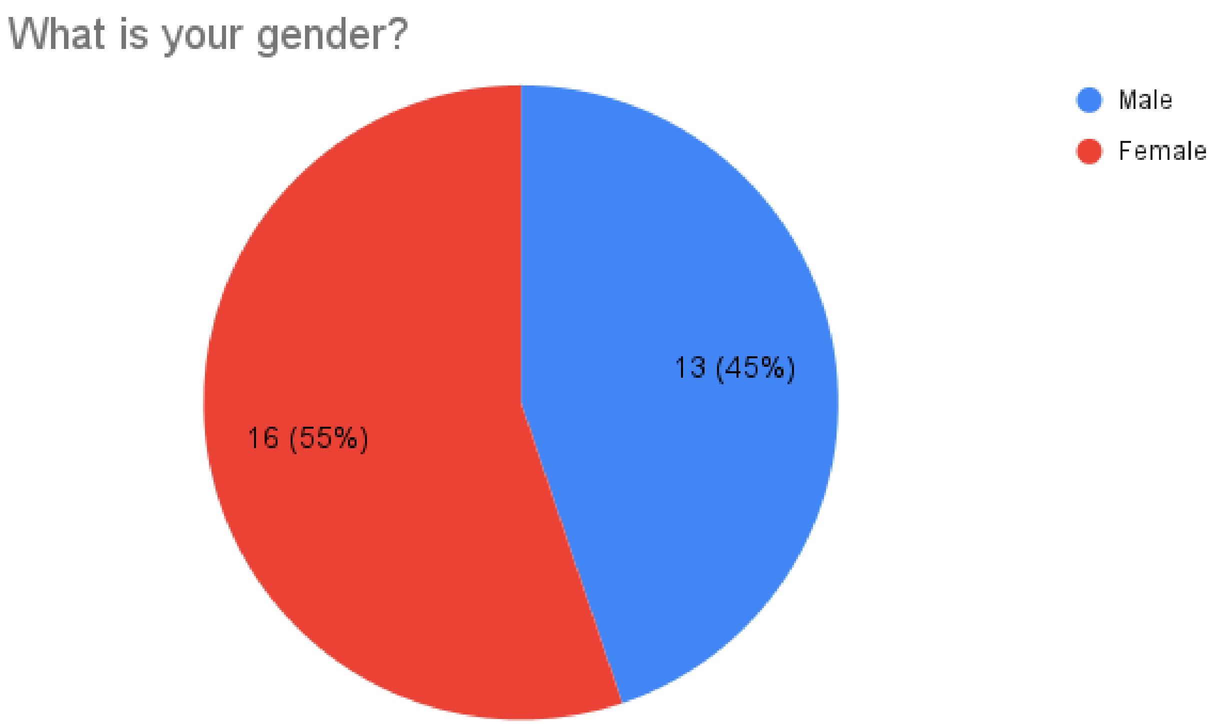


Figure B1: Pie charts showing gender distributions among the control group (top) and experimental group (bottom).

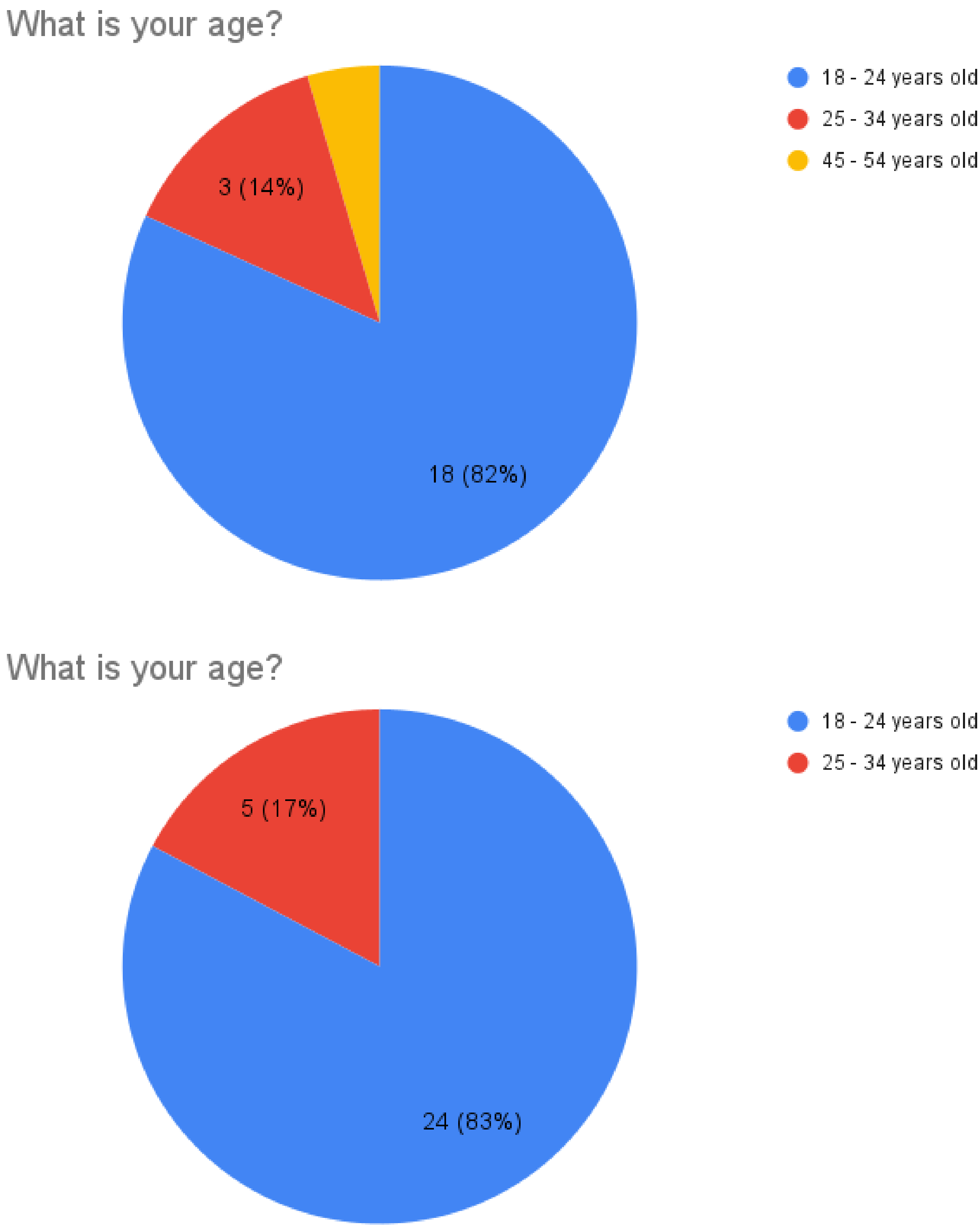


Figure B2: Pie charts showing age distributions among the control group (top) and experimental group (bottom).

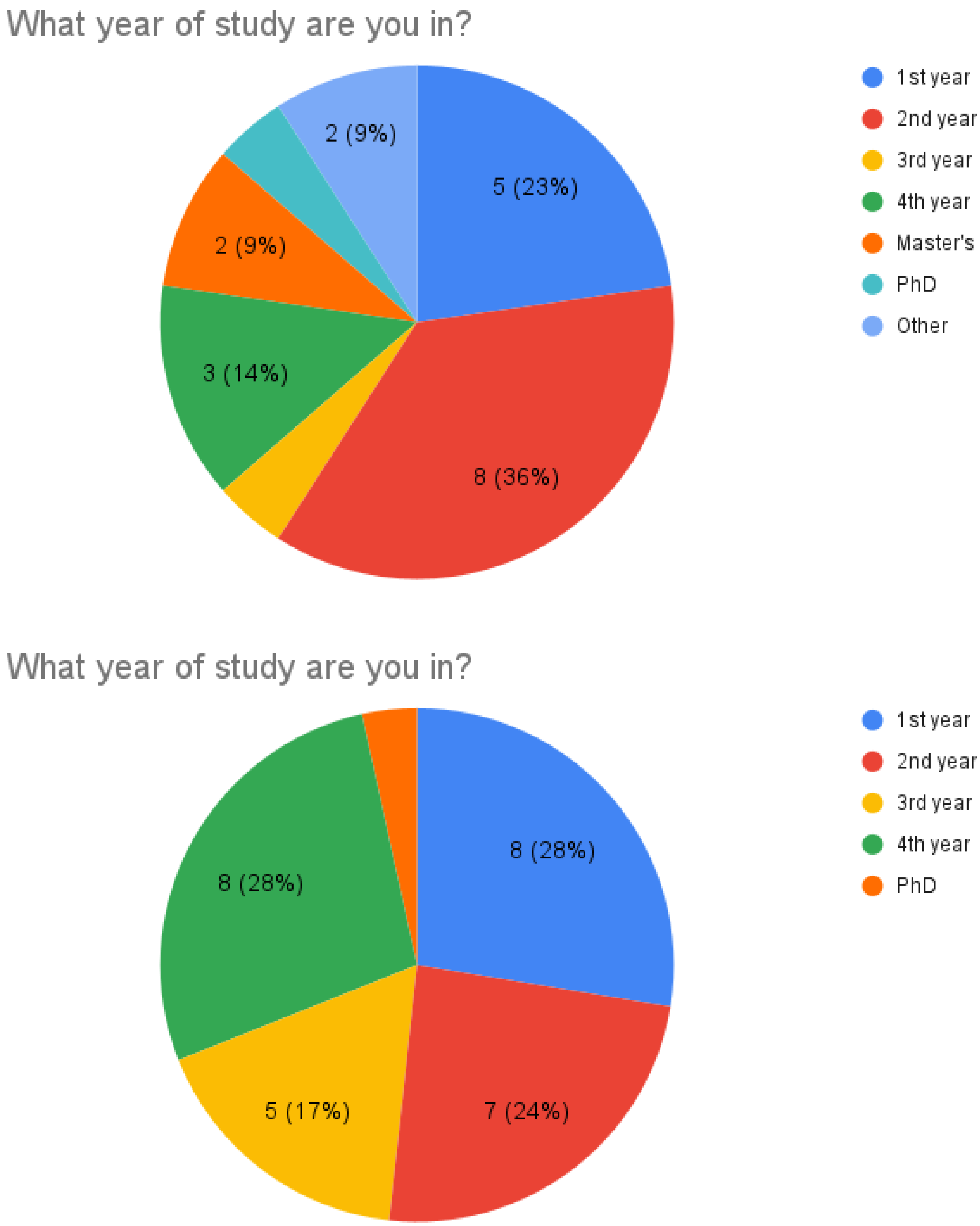


Figure B3: Pie charts showing distributions for year of study among the control group (top) and experimental group (bottom).

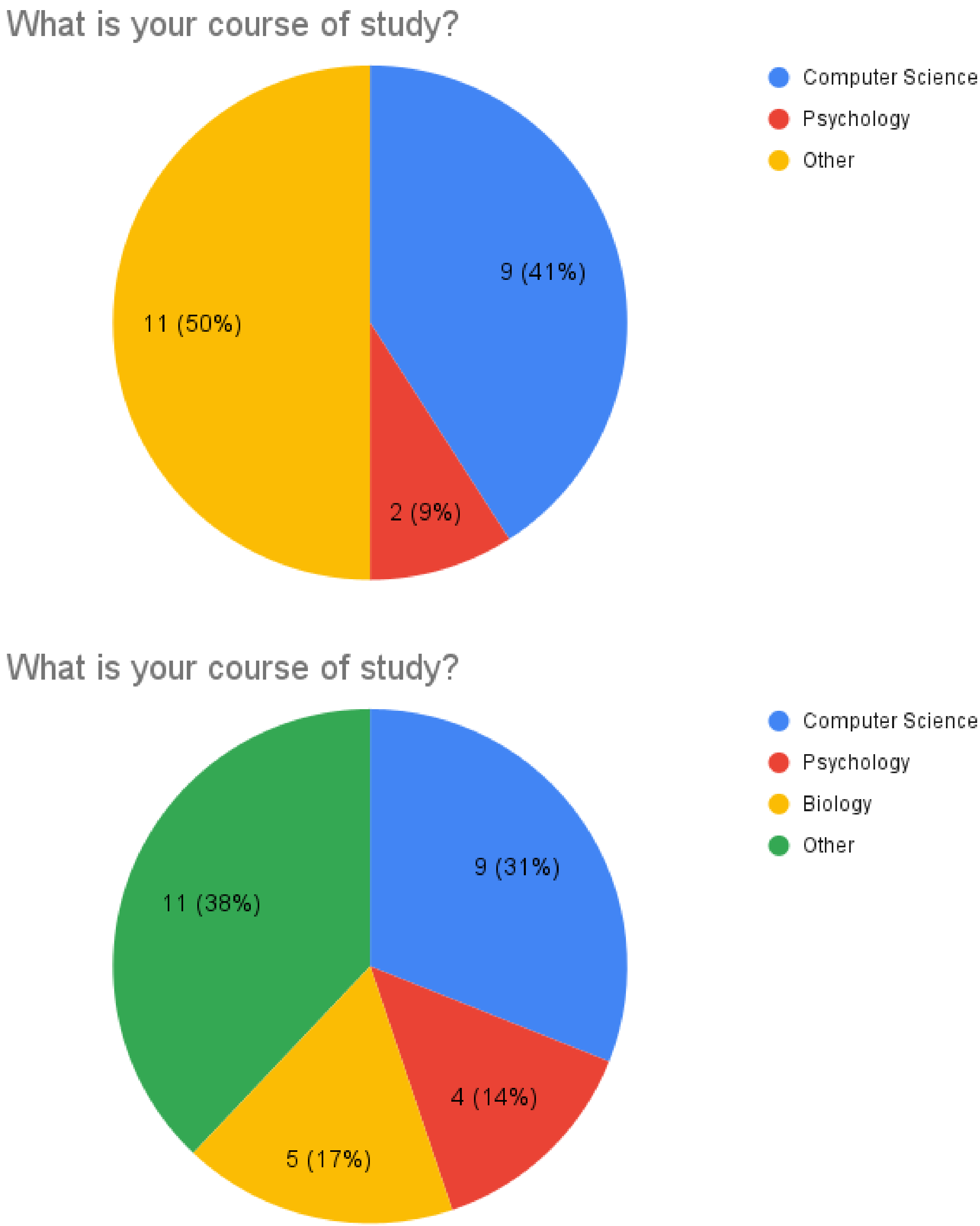
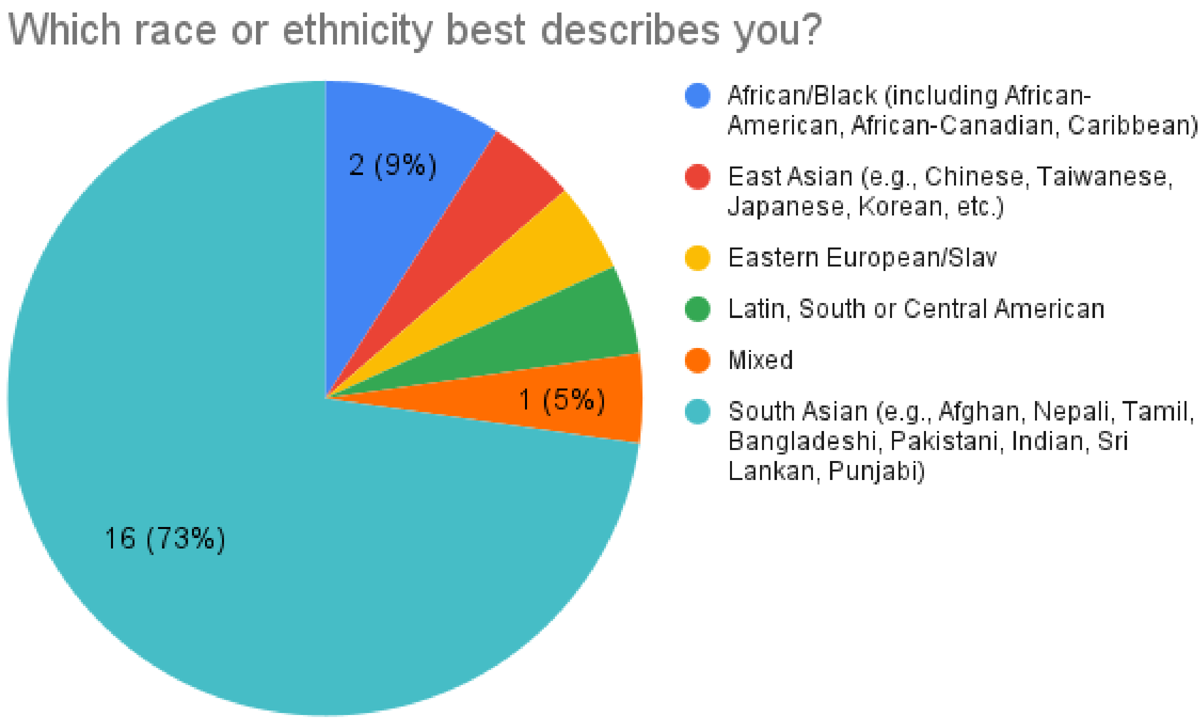


Figure B4: Pie charts showing distributions for course of study among the control group (top) and experimental group (bottom).



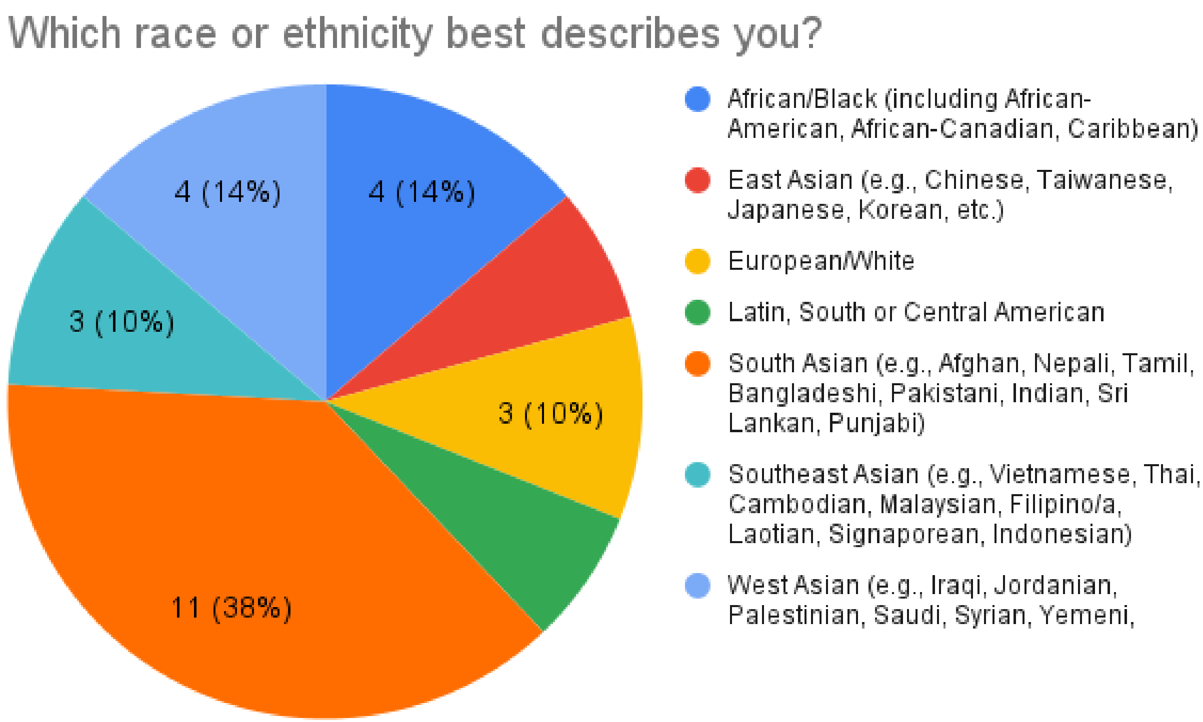
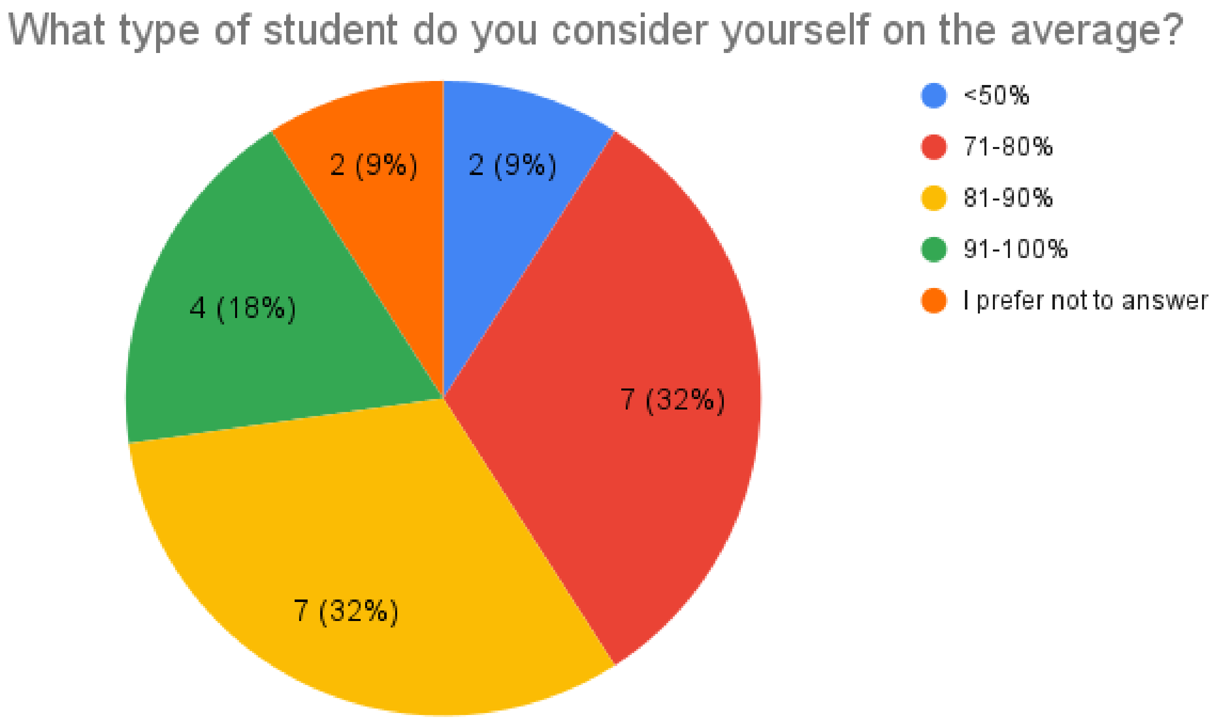


Figure B5: Pie charts showing distributions for race or ethnicity among the control group (top) and experimental group (bottom).



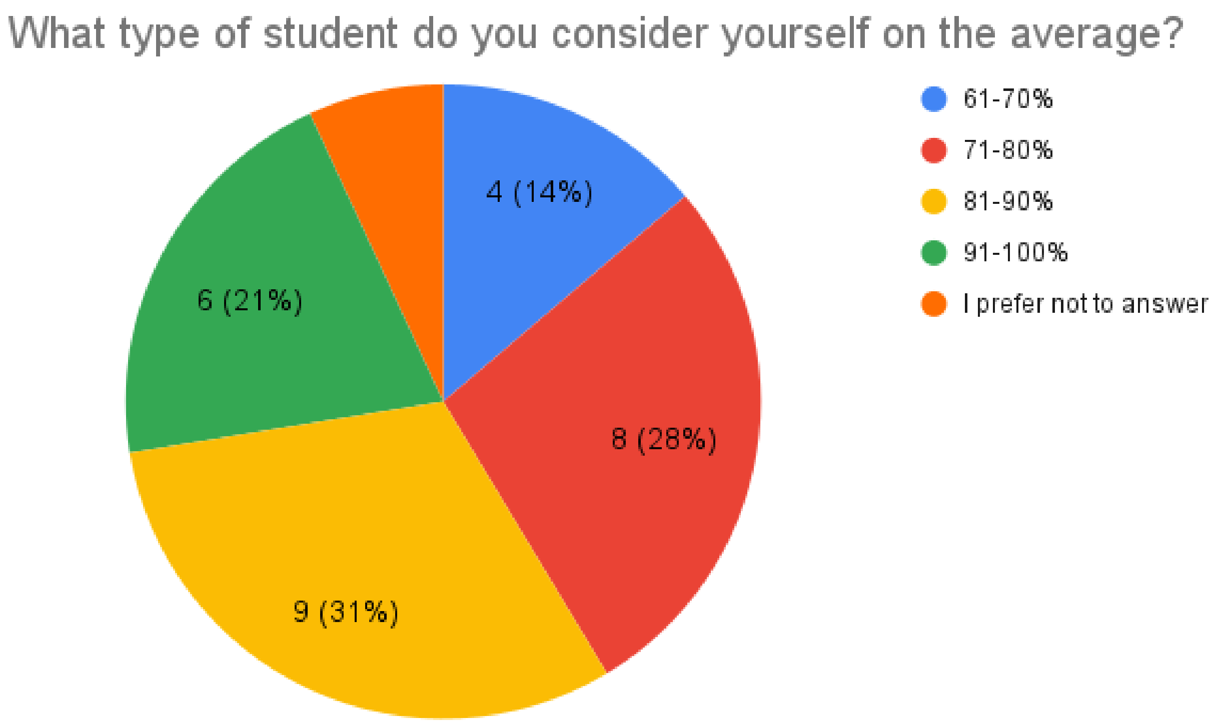


Figure B6: Pie charts showing distributions for average grades among the control group (top) and experimental group (bottom).

C Analysis of Test Data

C.1 Cohen's Kappa Tables

Table C1: Test averages and participant averages for Cohen's Kappa comparing 0 to other values (Left half: control, right half: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, PA: Participant Average, TA: Test Average).

	T1	T2	T3	PA	T1	T2	T3	PA
	1	1	0.9677	0.9892	0.8116	0.9677	1	0.9264
	1	0.9644	0.9040	0.9561	0.9382	1	0.9556	0.9646
	0.8908	0.9314	0.4882	0.7701	0.9673	1	1	0.9891
	0.9667	1	1	0.9889	0.6597	0.3158	1	0.6585
	1	1	1	1	0.9091	1	1	0.9697
	1	1	1	1	0.9685	0.9385	1	0.9690
	0.9020	0.9472	1	0.9497	1	1	1	1
	1	0.9673	1	0.9891	0.9091	0.8960	1	0.9350
	0.9247	0.9032	1	0.9426	0.9297	1	1	0.9766
	1	1	0.9351	0.9784	1	0.9384	0.9008	0.9464
	1	1	0.9691	0.9897	0.8901	0.9682	1	0.9528
	0.9622	0.5679	0.8390	0.7897	1	1	1	1
	0.8515	0.8267	1	0.8927	1	1	1	1
	0.8231	0.9297	0.9247	0.8925	0.9634	0.9074	1	0.9569
	1	1	1	1	1	0.9609	0.9384	0.9664
	0.9685	0.9360	0.8705	0.9250	0.9682	1	0.9677	0.9786
	1	0.9505	1	0.9835	1	1	1	1
	1	1	0.9691	0.9897	0.8465	0.9324	1	0.9263
	1	1	0.9653	0.9884	1	0.9324	1	0.9775
	1	0.9359	1	0.9786	0.9279	0.9384	0.7347	0.8670
	0.5476	0.9594	1	0.8357	1	0.9091	0.9247	0.9446
	1	0.9327	1	0.9776	0.9653	1	1	0.9884
	1	0.9385	1	0.9795	0.9653	1	1	0.9884
					1	0.9622	1	0.9874
					1	0.9682	0.9644	0.9775
					0.9533	0.9001	0.8575	0.9036
					1	1	0.9692	0.9897
					0.9660	1	1	0.9887
					0.9367	1	0.8807	0.9391
					0.9660	0.9377	1	0.9679
					1	1	0.9373	0.9791
					1	0.9634	1	0.9878
TA	0.9494	0.9431	0.9492		0.9513	0.9480	0.9697	

Table C2: Test averages and participant averages for Cohen's Kappa comparing 1 to other values (Left half: control, right half: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, PA: Participant Average, TA: Test Average).

	T1	T2	T3	PA	T1	T2	T3	PA
	1	1	1	1	1	1	1	1
	1	0	1	0.6667	1	1	1	1
	0.8980	0.8578	0.8598	0.8718	1	1	1	1
	1	1	0.8807	0.9602	1	1	1	1
	1	0	1	0.6667	1	1	1	1
	1	1	0.8164	0.9388	1	0.7174	1	0.9058
	0.7923	1	1	0.9308	0.6286	1	1	0.8762
	0	1	1	0.6667	1	1	1	1
	0.8807	0.5476	1	0.8094	1	1	1	1
	1	1	0.8403	0.9468	1	0.8492	1	0.9497
	1	1	1	1	0.7841	1	0.8492	0.8777
	0.8403	1	1	0.9468	1	1	1	1
	0.8297	0	1	0.6099	1	1	1	1
	1	1	0.8492	0.9497	0.7923	1	1	0.9308
	1	1	1	1	1	1	0	0.6667
	0.6448	0.2664	0	0.3037	0.8403	1	1	0.9468
	0	1	1	0.6667	1	1	1	1
	1	1	0.9667	0.9889	1	1	1	1
	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	0.6597	0.8866
	0.2073	0.6524	1	0.6199	1	1	1	1
	1	1	1	1	0.9008	1	1	0.9669
	0.6597	0.7923	1	0.8173	1	1	1	1
					1	1	1	1
					1	1	0.6597	0.8866
					0.9146	0.9008	1	0.9384
					1	1	1	1
					1	1	1	1
					1	1	1	1
					1	1	0.7841	0.9280
					1	1	1	1
					1	1	1	1
TA	0.8153	0.7877	0.9223		0.9644	0.9834	0.9360	

Table C3: Test averages and participant averages for Cohen's Kappa comparing blanks to other values (Left half: control, right half: RADAR, T1: Time 1, T2: Time 2, T3: Time 3, PA: Participant Average, TA: Test Average).

	T1	T2	T3	PA	T1	T2	T3	PA
	1	1	0.9667	0.9889	0.8267	0.9690	1	0.9319
	1	1	0.9172	0.9724	0.9382	1	0.9556	0.9646
	1	1	0.9577	0.9859	0.9673	1	1	0.9891
	0.9634	1	0.9673	0.9769	0.6597	0.5724	1	0.7440
	1	0.9673	1	0.9891	0.8980	1	1	0.9660
	1	1	0.9256	0.9752	0.9691	0.9351	1	0.9681
	0.9314	0.9472	1	0.9595	0.7385	1	1	0.9128
	0.9384	0.9673	1	0.9685	0.9233	0.8983	1	0.9405
	0.8980	0.8756	1	0.9245	0.9328	1	1	0.9776
	1	1	1	1	1	0.9685	0.9008	0.9564
	1	1	0.9691	0.9897	0.9673	0.9691	0.9594	0.9653
	0.9472	0.5646	0.8411	0.7843	1	1	1	1
	0.9327	0.8785	1	0.9371	1	1	1	1
	0.8274	0.9313	1	0.9196	0.9313	0.9069	1	0.9461
	1	1	1	1	1	0.9609	1	0.9870
	0.8656	0.8983	0.9360	0.8999	0.9634	1	0.9677	0.9770
	0.9634	0.9505	1	0.9713	1	1	1	1
	1	1	1	1	0.8344	0.9247	1	0.9197
	1	1	0.9667	0.9889	1	0.9505	1	0.9835
	1	0.9373	1	0.9791	0.9279	0.9378	0.9008	0.9221
	0.3455	0.8904	1	0.7453	1	0.9172	0.9324	0.9499
	1	0.9350	1	0.9783	1	1	1	1
	0.9594	0.9069	1	0.9554	0.9577	1	1	0.9859
					1	0.9609	1	0.9870
					1	0.9660	1	0.9887
					1	0.9377	0.8575	0.9317
					1	1	0.9692	0.9897
					0.9660	1	1	0.9887
					0.9328	1	0.8807	0.9378
					0.9653	0.9366	0.9297	0.9439
					1	1	0.9367	0.9789
					1	0.9533	1	0.9844
TA	0.9379	0.9413	0.9760		0.9469	0.9583	0.9747	

C.2 Mean Difference Graphs

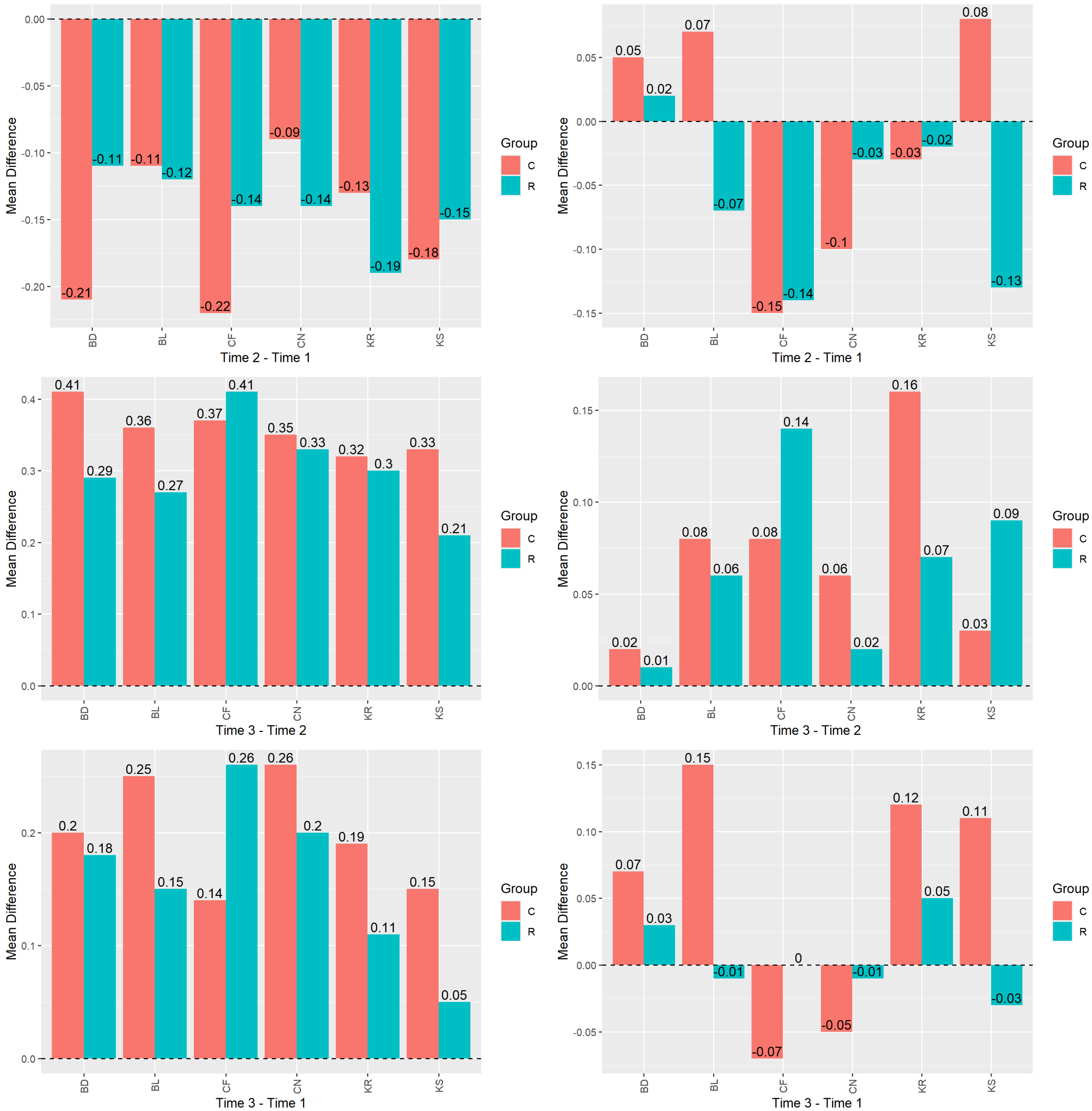


Figure C1: Topic-based performance mean difference between each pair of times in which participants were tested for recall (aka remembering - left) and recognition (aka understanding - right), with dotted line showing no difference in range -1 to 1. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

C.3 ANOVAs and Pairwise Comparisons

Table C4: Topic-based pairwise comparison of combined recall and recognition scores with Tukey method familywise correction due to the main effect of topic in Table 15. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
CN - CF	0.81	0.53	139.53	33.69	1750	4.14	0.0005
CN - BL	0.81	0.76	-222.58	33.69	1750	-6.61	0.0000
CN - BD	0.81	0.61	69.59	33.69	1750	2.07	0.3062
CN - KS	0.81	0.54	165.48	33.69	1750	4.91	0.0000
CN - KR	0.81	0.58	226.07	33.69	1750	6.71	0.0000
CF - BL	0.53	0.76	-362.12	33.69	1750	-10.75	0.0000
CF - BD	0.53	0.61	-69.95	33.69	1750	-2.08	0.3004
CF - KS	0.53	0.54	25.94	33.69	1750	0.77	0.9725
CF - KR	0.53	0.58	86.54	33.69	1750	2.57	0.1057
BL - BD	0.76	0.61	292.17	33.69	1750	8.67	0.0000
BL - KS	0.76	0.54	388.06	33.69	1750	11.52	0.0000
BL - KR	0.76	0.58	448.66	33.69	1750	13.32	0.0000
BD - KS	0.61	0.54	95.89	33.69	1750	2.85	0.0510
BD - KR	0.61	0.58	156.49	33.69	1750	4.64	0.0001
KS - KR	0.54	0.58	60.59	33.69	1750	1.80	0.4670

Table C5: Time-based pairwise comparison of combined recall and recognition scores with Tukey method familywise correction due to the main effect of time in Table 15. SE: Standard Error, Df: Degrees of freedom, T1: Time 1, T2: Time 2, T3: Time 3.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
T1 - T2	0.64	0.60	172.16	23.98	1750	7.18	0.0000
T1 - T3	0.64	0.67	-181.10	23.98	1750	-7.55	0.0000
T2 - T3	0.60	0.67	-353.26	23.98	1750	-14.73	0.0000

Table C6: One-way ANOVA based on task and topic due to interaction between them in Table 15. RCL: Recall, RCG: Recognition, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

One-way ANOVA at each level of task				
		RCL	RCG	Task Effect
One-way ANOVA at each level of topic	CN	0.42	0.81	$F(1,259) = 183.08$, $p = 0.0000$
	CF	0.53	0.53	$F(1,259) = 0.03$, $p = 0.8599$
	BL	0.71	0.76	$F(1,259) = 10.18$, $p = 0.0016$
	BD	0.53	0.61	$F(1,259) = 9.56$, $p = 0.0022$
	KS	0.48	0.54	$F(1,259) = 5.71$, $p = 0.0176$
	KR	0.39	0.58	$F(1,259) = 46.77$, $p = 0.0000$
	Topic Effect	$F(5,879) = 29.04$, $p = 0.0000$	$F(5,879) = 42.38$, $p = 0.0000$	

Table C7: Recall's and Recognition's topic-based pairwise comparison with Tukey method familywise correction due to the simple effect of topic in Table C6. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
CN - CF	0.42	0.53	-84.08	23.31	879	-3.61	0.0044	0.81	0.53	242.08	22.46	879	10.78	0.0000
CN - BL	0.42	0.71	-223.12	23.31	879	-9.57	0.0000	0.81	0.76	41.17	22.46	879	1.83	0.4448
CN - BD	0.42	0.53	-72.73	23.31	879	-3.12	0.0230	0.81	0.61	178.66	22.46	879	7.96	0.0000
CN - KS	0.42	0.48	-33.23	23.31	879	-1.43	0.7117	0.81	0.54	230.22	22.46	879	10.25	0.0000
CN - KR	0.42	0.39	28.87	23.31	879	1.24	0.8179	0.81	0.58	207.94	22.46	879	9.26	0.0000
CF - BL	0.53	0.71	-139.04	23.31	879	-5.96	0.0000	0.53	0.76	-200.90	22.46	879	-8.95	0.0000
CF - BD	0.53	0.53	11.36	23.31	879	0.49	0.9966	0.53	0.61	-63.41	22.46	879	-2.82	0.0547
CF - KS	0.53	0.48	50.86	23.31	879	2.18	0.2475	0.53	0.54	-11.86	22.46	879	-0.53	0.9951
CF - KR	0.53	0.39	112.96	23.31	879	4.84	0.0000	0.53	0.58	-34.13	22.46	879	-1.52	0.6515
BL - BD	0.71	0.53	150.39	23.31	879	6.45	0.0000	0.76	0.61	137.49	22.46	879	6.12	0.0000
BL - KS	0.71	0.48	189.89	23.31	879	8.14	0.0000	0.76	0.54	189.05	22.46	879	8.42	0.0000
BL - KR	0.71	0.39	251.99	23.31	879	10.81	0.0000	0.76	0.58	166.77	22.46	879	7.43	0.0000
BD - KS	0.53	0.48	39.50	23.31	879	1.69	0.5360	0.61	0.54	51.56	22.46	879	2.30	0.1968
BD - KR	0.53	0.39	101.6	23.31	879	4.36	0.0002	0.61	0.58	29.28	22.46	879	1.30	0.7831
KS - KR	0.48	0.39	62.10	23.31	879	2.66	0.0837	0.54	0.58	-22.28	22.46	879	-0.99	0.9205

Table C8: One-way ANOVA based on task and time due to interaction between them in Table 15. RCL: Recall, RCG: Recognition, T1: Time 1, T2: Time 2, T3: Time 3.

		One-way ANOVA at each level of task		
		RCL	RCG	Task Effect
One-way ANOVA at each level of time	T1	0.50	0.64	$F(1,571) = 50.03$, $p = 0.0000$
	T2	0.35	0.60	$F(1,571) = 149.77$, $p = 0.0000$
	T3	0.68	0.67	$F(1,571) = 0.45$, $p = 0.5036$
	Time Effect	$F(2,882) = 118.43$, $p = 0.0000$	$F(2,882) = 6.44$, $p = 0.0017$	

Table C9: Recall's and Recognition's time-based pairwise comparison with Tukey method familywise correction due to the simple effect of time in Table C8. SE: Standard Error, Df: Degrees of freedom, T1: Time 1, T2: Time 2, T3: Time 3.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
T1 - T2	0.50	0.35	114.10	15.77	882	7.23	0.0000	0.64	0.60	42.47	17.53	882	2.42	0.0413
T1 - T3	0.50	0.68	-128.51	15.77	882	-8.15	0.0000	0.64	0.67	-18.98	17.53	882	-1.08	0.5253
T2 - T3	0.35	0.68	-242.61	15.77	882	-15.38	0.0000	0.60	0.67	-61.45	17.53	882	-3.50	0.0014

Table C10: Recall's and Recognition's topic-based pairwise comparison with Tukey method familywise correction due to the main effect of topic in Table 16. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
CN - CF	0.42	0.53	-90.87	22.55	850	-4.03	0.0009	0.81	0.53	257.49	23.29	850	11.06	0.0000
CN - BL	0.42	0.71	-235.69	22.55	850	-10.45	0.0000	0.81	0.76	43.82	23.29	850	1.88	0.4141
CN - BD	0.42	0.53	-88.24	22.55	850	-3.91	0.0014	0.81	0.61	189.41	23.29	850	8.13	0.0000
CN - KS	0.42	0.48	-47.39	22.55	850	-2.10	0.2875	0.81	0.54	235.13	23.29	850	10.10	0.0000
CN - KR	0.42	0.39	24.28	22.55	850	1.08	0.8907	0.81	0.58	214.12	23.29	850	9.19	0.0000
CF - BL	0.53	0.71	-144.82	22.55	850	-6.42	0.0000	0.53	0.76	-213.67	23.29	850	-9.18	0.0000
CF - BD	0.53	0.53	2.63	22.55	850	0.12	1.0000	0.53	0.61	-68.08	23.29	850	-2.92	0.0413
CF - KS	0.53	0.48	43.48	22.55	850	1.93	0.3857	0.53	0.54	-22.35	23.29	850	-0.96	0.9303
CF - KR	0.53	0.39	115.14	22.55	850	5.11	0.0000	0.53	0.58	-43.37	23.29	850	-1.86	0.4262
BL - BD	0.71	0.53	147.45	22.55	850	6.54	0.0000	0.76	0.61	145.59	23.29	850	6.25	0.0000
BL - KS	0.71	0.48	188.30	22.55	850	8.35	0.0000	0.76	0.54	191.31	23.29	850	8.22	0.0000
BL - KR	0.71	0.39	259.96	22.55	850	11.53	0.0000	0.76	0.58	170.29	23.29	850	7.31	0.0000
BD - KS	0.53	0.48	40.85	22.55	850	1.81	0.4589	0.61	0.54	45.72	23.29	850	1.96	0.3644
BD - KR	0.53	0.39	112.51	22.55	850	4.99	0.0000	0.61	0.58	24.71	23.29	850	1.06	0.8966
KS - KR	0.48	0.39	71.66	22.55	850	3.18	0.0191	0.54	0.58	-21.02	23.29	850	-0.90	0.9459

Table C11: Recall's and Recognition's time-based pairwise comparison with Tukey method familywise correction due to the main effect of time in Table 16. SE: Standard Error, Df: Degrees of freedom, T1: Time 1, T2: Time 2, T3: Time 3.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
T1 - T2	0.50	0.35	120.78	15.50	850	7.79	0.0000	0.64	0.60	38.64	17.62	850	2.19	0.0730
T1 - T3	0.50	0.68	-137.44	15.50	850	-8.87	0.0000	0.64	0.67	-28.99	17.62	850	-1.65	0.2271
T2 - T3	0.35	0.68	-258.22	15.50	850	-16.66	0.0000	0.60	0.67	-67.63	17.62	850	-3.84	0.0004

Table C12: One-way ANOVA due to interaction between condition and topic in Table 16 for recall only [$F(5,850) = 4.34, p = 0.0007$]. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of topic	CN	0.35	0.48	$F(1,154) = 6.93$, $p = 0.0093$
	CF	0.47	0.57	$F(1,154) = 3.06$, $p = 0.0822$
	BL	0.72	0.70	$F(1,154) = 0.48$, $p = 0.4897$
	BD	0.53	0.53	$F(1,154) = 0.01$, $p = 0.9231$
	KS	0.36	0.56	$F(1,154) = 13.16$, $p = 0.0004$
	KR	0.32	0.44	$F(1,154) = 4.23$, $p = 0.0414$
	Topic Effect	$F(5,369) = 20.51$, $p = 0.0000$	$F(5,505) = 11.37$, $p = 0.0000$	

Table C13: Control's and RADAR's topic-based pairwise comparison for recall with Tukey method familywise correction due to the simple effect of topic in Table C12. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
CN - CF	0.35	0.47	-38.70	15.03	369	-2.58	0.1061	0.48	0.57	-45.95	17.82	505	-2.58	0.1045
CN - BL	0.35	0.72	-118.23	15.03	369	-7.87	0.0000	0.48	0.70	-103.23	17.82	505	-5.79	0.0000
CN - BD	0.35	0.53	-52.69	15.03	369	-3.51	0.0067	0.48	0.53	-20.04	17.82	505	-1.12	0.8710
CN - KS	0.35	0.36	0.25	15.03	369	0.02	1.0000	0.48	0.56	-32.74	17.82	505	-1.84	0.4425
CN - KR	0.35	0.32	8.29	15.03	369	0.55	0.9939	0.48	0.44	19.01	17.82	505	1.07	0.8945
CF - BL	0.47	0.72	-79.53	15.03	369	-5.29	0.0000	0.57	0.70	-57.28	17.82	505	-3.21	0.0174
CF - BD	0.47	0.53	-13.98	15.03	369	-0.93	0.9385	0.57	0.53	25.91	17.82	505	1.45	0.6939
CF - KS	0.47	0.36	38.95	15.03	369	2.59	0.1019	0.57	0.56	13.21	17.82	505	0.74	0.9767
CF - KR	0.47	0.32	46.99	15.03	369	3.13	0.0233	0.57	0.44	64.96	17.82	505	3.65	0.0040
BL - BD	0.72	0.53	65.55	15.03	369	4.36	0.0002	0.70	0.53	83.19	17.82	505	4.67	0.0001
BL - KS	0.72	0.36	118.48	15.03	369	7.88	0.0000	0.70	0.56	70.49	17.82	505	3.96	0.0012
BL - KR	0.72	0.32	126.52	15.03	369	8.42	0.0000	0.70	0.44	122.24	17.82	505	6.86	0.0000
BD - KS	0.53	0.36	52.94	15.03	369	3.52	0.0064	0.53	0.56	-12.70	17.82	505	-0.71	0.9804
BD - KR	0.53	0.32	60.98	15.03	369	4.06	0.0009	0.53	0.44	39.05	17.82	505	2.19	0.2436
KS - KR	0.36	0.32	8.04	15.03	369	0.53	0.9947	0.56	0.44	51.75	17.82	505	2.90	0.0443

Table C14: One-way ANOVA due to interaction between condition and topic in Table 16 for recognition only. C: Control, R: RADAR, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

		One-way ANOVA at each level of condition		
		C	R	Condition Effect
One-way ANOVA at each level of topic	CN	0.84	0.79	$F(1,154) = 3.01$, $p = 0.0850$
	CF	0.54	0.52	$F(1,154) = 0.13$, $p = 0.7225$
	BL	0.78	0.74	$F(1,154) = 0.96$, $p = 0.3284$
	BD	0.62	0.60	$F(1,154) = 0.18$, $p = 0.6723$
	KS	0.48	0.57	$F(1,154) = 2.88$, $p = 0.0917$
	KR	0.56	0.58	$F(1,154) = 0.50$, $p = 0.4794$
	Topic Effect	$F(5,369) = 23.42$, $p = 0.0000$	$F(5,505) = 20.51$, $p = 0.0000$	

Table C15: Control's and RADAR's topic-based pairwise comparison for recognition with Tukey method familywise correction due to the simple effect of topic in Table C14. SE: Standard Error, Df: Degrees of freedom, BD: Biological Organization Disciplines, CF: Cranial Nerves Functions, CN: Cranial Nerves, KR: Krebs Cycle Reactions, KS: Krebs Cycle Substrates.

Contrast	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value	Score 1	Score 2	Estimate	SE	Df	t-ratio	p-value
CN - CF	0.84	0.54	103.83	14.27	369	7.28	0.0000	0.79	0.52	139.34	17.30	505	8.06	0.0000
CN - BL	0.84	0.78	18.45	14.27	369	1.29	0.7887	0.79	0.74	23.42	17.30	505	1.35	0.7546
CN - BD	0.84	0.62	76.92	14.27	369	5.39	0.0000	0.79	0.60	102.17	17.30	505	5.91	0.0000
CN - KS	0.84	0.48	118.34	14.27	369	8.29	0.0000	0.79	0.57	110.61	17.30	505	6.39	0.0000
CN - KR	0.84	0.56	99.01	14.27	369	6.94	0.0000	0.79	0.58	108.23	17.30	505	6.26	0.0000
CF - BL	0.54	0.78	-85.37	14.27	369	-5.98	0.0000	0.52	0.74	-115.92	17.30	505	-6.70	0.0000
CF - BD	0.54	0.62	-26.91	14.27	369	-1.89	0.4126	0.52	0.60	-37.17	17.30	505	-2.15	0.2642
CF - KS	0.54	0.48	14.52	14.27	369	1.02	0.9121	0.52	0.57	-28.73	17.30	505	-1.66	0.5585
CF - KR	0.54	0.56	-4.82	14.27	369	-0.34	0.9994	0.52	0.58	-31.11	17.30	505	-1.80	0.4676
BL - BD	0.78	0.62	58.46	14.27	369	4.10	0.0007	0.74	0.60	78.76	17.30	505	4.55	0.0001
BL - KS	0.78	0.48	99.89	14.27	369	7.00	0.0000	0.74	0.57	87.19	17.30	505	5.04	0.0000
BL - KR	0.78	0.56	80.55	14.27	369	5.64	0.0000	0.74	0.58	84.81	17.30	505	4.90	0.0000
BD - KS	0.62	0.48	41.42	14.27	369	2.90	0.0450	0.60	0.57	8.44	17.30	505	0.49	0.9966
BD - KR	0.62	0.56	22.09	14.27	369	1.55	0.6333	0.60	0.58	6.06	17.30	505	0.35	0.9993
KS - KR	0.48	0.56	-19.33	14.27	369	-1.35	0.7539	0.57	0.58	-2.38	17.30	505	-0.14	1.0000