Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Svein-Kåre Bjørnsen

Teaching for Tomorrow

Exploring the viability of applying methodology from Direct Instruction and Precision Teaching as foundation for digital educational gamification

Master's thesis in Master in Applied Computer Science
Supervisor: Deepti Mishra
June 2023

NTNU
Norwegian University of
Science and Technology

Svein-Kåre Bjørnsen

Teaching for Tomorrow

Exploring the viability of applying methodology from Direct Instruction and Precision Teaching as foundation for digital educational gamification

Master's thesis in Master in Applied Computer Science
Supervisor: Deepti Mishra
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Serious games, gamification and adaptive learning systems have in recent decades been exploding fields of research. However, high-quality research on their efficacy is rare, and the results are highly variable, while concrete evidence-based educational methodology is lacking.

The discipline of behavior analysis has generated evidence-based educational methods that have long histories of good results, however these are not commonly known, and have not seen mainstream adoption in educational systems around the world.

In this thesis, I present a prototype of a lightly gamified system for teaching facts that was aimed at adapting to the learner's personal learning aptitude and typing speed. The prototype applied principles from the evidence-based methods Direct Instruction and Precision Teaching to train the participants using frequency-building of correctly typed responses in a flashcard-like web-based application.

A mixed-methods study was conducted in an attempt to gauge the prototype's efficacy, the user's reactions to it, and the viability of applying these methods without considerable assistance from experts in them. The study had a quasi-experimental part that utilised a pretest, posttest and a retention test; following the experiment, a questionnaire containing likert items as well as free-text responses was issued. A thematic analysis was conducted on the qualitative data from the questionnaire. The prototype was tested in conjunction with a cloud technology course in Norway with students at the bachelor level in university.

There was statistically significant increase in score for the experimental group between the pretest and posttest, with a slight statistically significant decrease between posttest and retention test. The percent-wise increase in score had a mean of 95% from pretest to posttest, and -9% from posttest to retention test. Comparison with the control group, and calculation of effect size, was infeasible due to sampling issues. The reactions from the participants were mixed, and several technical deficiencies with the prototype were identified.

The prototype showed promise when tentatively compared to results from the literature on serious games, gamification and adaptive learning systems, and a future replication seems worthwhile. However, I recommend that future work in this direction be interdisciplinary, since properly implementing the methods is not trivial without expert guidance.

Sammendrag

Serious games, gamifisering og adaptive læringsystemer har de siste tiårene vært eksplosive forskningsområder. Imidlertid er det sjeldent høykvalitetsforskning når det kommer til deres effektivitet, og resultatene er svært variable, samtidig som det mangler konkrete evidensbaserte pedagogiske metoder.

Forskningsdisiplinen atferdsanalyse har generert evidensbaserte pedagogiske metoder som har langvarig historikk med gode resultater. Disse er imidlertid lite kjent og har ikke sett bred implementering i utdanningssystemer rundt om i verden.

I denne avhandlingen presenterer jeg en prototype av et lett gamifisert system for å undervise fakta som var rettet mot tilpasning til brukerens personlige læringsevne og skrivehastighet. Prototypen brukte prinsipper fra de evidensbaserte metodene *Direct Instruction* og Presisjonsopplæring for å trene deltakerne ved frekvensbygging av korrekt skrevne svar i en flashcard-lignende webapplikasjon.

Det ble gjennomført en studie med kombinerte metoder i et forsøk på å vurdere prototypens effektivitet, brukernes reaksjoner på den og levedyktigheten i å anvende disse metodene uten betydelig hjelp fra eksperter på området. Studien hadde en kvasi-eksperimentell del som benyttet en fortest, ettertest og en tilbakekallstest. Etter eksperimentet ble det utført en spørreundersøkelse med Likert-skala samt åpne tekstbaserte svar. En tematisk analyse ble gjennomført på kvalitative data fra spørreundersøkelsen. Prototypen ble testet i forbindelse med et skyteknologikurs i Norge med studenter på bachelornivå ved universitet.

Det var en statistisk signifikant økning i poengsum for forsøksgruppen mellom fortest og ettertest, med en liten, statistisk signifikant nedgang mellom ettertest og tilbakekallstest. Prosentvis økning i poengsum hadde et gjennomsnitt på 95% fra fortest til ettertest, og -9% fra ettertest til tilbakekallstest. Sammenligning med kontrollgruppen og beregning av effektstørrelse var ikke fornuftig på grunn av utvalgsproblemer. Reaksjonene fra deltakerne var blandede, og flere tekniske mangler med prototypen ble identifisert.

Prototypen viste lovende resultater når den ble tentativt sammenlignet med resultater fra litteraturen om serious games, gamifisering og adaptive læringsystemer, og en fremtidig replikasjon virker verdt å gjennomføre. Imidlertid anbefaler jeg at fremtidig arbeid i denne retningen blir tverrfaglig, ettersom implementering av metodene på riktig måte ikke er enkelt uten ekspertveiledning.

Acknowledgements

A master thesis is not written in a vacuum without help and advice from others. I wish to acknowledge the help and support I have received from people in connection with this thesis and degree.

First my partner, Beatrice, deserves my gratitude. She has been supportive and as patient as she could be throughout my struggles with the degree and thesis, especially considering the narrow confines of our house in the latter year of the pandemic. Thank you.

I am also grateful for the support of several competent academics in the making of this thesis. Deepti Mishra deserves thanks for excellent supervision and advice throughout the master thesis work. Christopher Frantz should also be thanked, since he was instrumental during the early phases of my ideation and planning for this project, as well as support and advice throughout the last year — not to mention for also allowing me to run my experiment in conjunction with his course. This thesis would not exist without the help of these two people. Rasmi Krippendorf also deserves a mention here, for his advice on the behavior analytic aspects of the prototype and helping me present behavior analysis as a discipline in an accurate light.

Contents

Abstract . . . . . **iii**
Sammendrag . . . . . **v**
Acknowledgements . . . . . **vii**
Contents . . . . . **ix**
Figures . . . . . **xi**
Tables . . . . . **xiii**
Acronyms . . . . . **xv**
Preface . . . . . **xvii**
1 Introduction **1**
1.1 What behavior analysis is 2
1.2 Evidence-based educational methods 2
1.2.1 Direct Instruction 3
1.2.2 Precision Teaching 3
1.3 Study objective and findings 4
1.4 Structure of the thesis 5
2 Background **7**
2.1 Systematic literature reviews 7
2.2 Related work from behavior analysis 9
2.3 Research questions 10
3 Implementation **13**
3.1 Curriculum material 13
3.2 Users' interaction with the system 13
3.3 Procedural details 27
3.4 System architecture 30
3.4.1 Software architecture 30
3.4.2 Infrastructure 31
4 Method **37**
4.1 Participants and setting 37
4.2 Design 37
4.2.1 Sampling change and grade credit incentive 39
4.3 Data collection 40
4.3.1 Experimental tests 40
4.3.2 Questionnaire 40
5 Data analysis and results **43**

5.1 Data analysis . . . 43
5.1.1 Experimental metrics . . . 43
5.1.2 Statistical methods . . . 44
5.1.3 Thematic analysis . . . 44
5.2 Quantitative results . . . 45
5.2.1 Experimental tests . . . 45
5.2.2 Questionnaire . . . 53
5.2.3 Test diligence . . . 54
5.3 Thematic analysis of free-text responses . . . 59
5.3.1 Themes . . . 59
5.3.2 Coding . . . 61
5.4 Time series . . . 68
6 Discussion . . . 73
6.1 Answering the research questions . . . 73
6.1.1 RQ 1: Efficacy . . . 73
6.1.2 RQ 2: Student reactions . . . 75
6.1.3 RQ 3: Viability without method expert integration . . . 76
6.2 Other implications of the results . . . 77
6.3 Deficiencies of the current prototype . . . 78
6.4 Limitations of the study . . . 78
6.5 Future work . . . 79
7 Conclusion . . . 81
Bibliography . . . 83
A Curriculum template . . . 87
B Questionnaire . . . 99
C Complete coding for the thematic analysis . . . 105
C.1 Chapter system . . . 107
C.2 Cheating/wanting to cheat . . . 107
C.3 Enjoyed using it . . . 108
C.4 Forgot early codes . . . 108
C.5 Looked up codes . . . 109
C.6 Other feature requests . . . 109
C.7 Pretest was bad . . . 110
C.8 Random selection of next item backfired . . . 111
C.9 Rigid scoring system in PT phase made typos more punishing . . . 111
C.10 Sees speed training as a test . . . 112
C.11 Speed mismatch . . . 113
C.12 Stuck on 4xx modules . . . 114
C.13 Thinking break . . . 115
D Verbatim free-text responses from questionnaire . . . 117
E Serious games / gamification review . . . 123
F Adaptive learning systems review . . . 135

Figures

3.1 Main view with only an experimental test available 14
3.2 Example of multiple choice prompt during experimental testing . . 15
3.3 Dialog shown on completion of the pretest 16
3.4 Dialog shown before starting training on a new module 16
3.5 Overview of training process 17
3.6 Sample of DI prompt presentation 18
3.7 DI trial without hint 18
3.8 DI trial after user has made an error 19
3.9 Dialog shown when a DI session has been completed 19
3.10 Dialog shown before measuring typing speed for the first time . . . 20
3.11 Warm up high score dialog 20
3.12 Main view of frontend 21
3.13 Dialog shown before beginning a 10-second learning speed measuring PT session 22
3.14 Visual countdown before start of PT session 22
3.15 Normal PT trial 23
3.16 PT trial after a user has entered a wrong response 23
3.17 Dialog after PT session where the user beat their high score 24
3.18 Dialog after PT session where the user didn't beat their high score . 24
3.19 Overview of PT process 25
3.20 Dialog shown when trials to criterion has been established 25
3.21 Dialog shown after 9th unsuccessful try on the learning speed measurement phase 26
3.22 The server executable and its modules 33
3.23 Overview of the frontend architecture 34
3.24 Overview of the system infrastructure 35

4.1 Participant fall off during the study 39

5.1 Q-Q plot of pretest data 46
5.2 Score comparison, pre- vs posttest 47
5.3 Individual score differences, pre- vs posttest 47
5.4 Individual score increase from pre- to posttest 48
5.5 Duration comparison, pre- vs posttest 48

5.6 Individual duration differences, pre- vs posttest 49
5.7 Score rate comparison, pre- vs posttest 49
5.8 Score comparison, pre- vs posttest vs retention test 50
5.9 Individual score differences, pre- vs posttest vs retention test 50
5.10 Individual score increases, pre- to posttest vs post- to retention test 51
5.11 Duration comparison, pre- vs posttest vs retention test 51
5.12 Score rates, between pre- vs post- vs retention test 52
5.13 Engagement question results 55
5.14 Phase 1 sentiment question results 55
5.15 Phase 2 sentiment question results 56
5.16 Overall enjoyment question results 56
5.17 Desire of future use results 56
5.18 Usage convenience question results 57
5.19 Too many typos question results 57
5.20 External factors question results 57
5.21 Pretest seriousness question results 58
5.22 Posttest seriousness question results 58
5.23 Retention test seriousness question results 58
5.24 Thematic structure of free-text responses 59
5.25 Example of typical time series around the start of 4xx status code training 69
5.26 Aggregate time series of speed training of all participants 70
5.27 Time series for an individual who procrastinated 71

B.1 Page 1 of questionnaire 101
B.2 Page 2 of questionnaire 101
B.3 Page 3 of questionnaire 102
B.4 Page 4 of questionnaire 102
B.5 Page 5 of questionnaire 103
B.6 Page 6 of questionnaire 103

Tables

2.1 Mean learning gains from serious games/gamification and ALS literature 9

3.1 DI sequence with no errors 27
3.2 DI sequence with a leading chain of errors 28
3.3 DI sequence with intermixed error 28
3.4 Advised max number of new items to learn at a time based on trials-to-criterion 29

Acronyms

ALS Adaptive learning system. xiii, 1, 2, 8, 9, 74, 75

DI Direct Instruction. 2–4, 14, 15, 17, 21, 22, 27, 29, 75, 76, 78, 81

LG Learning Gain. xv, 8, 9

LGC Learning Gain (LG) for a control group. 9

LGE Learning Gain (LG) for an experimental group. 9, 74

LGI Learning Gain Improvement. 9

PT Precision Teaching. 2–4, 9, 14, 17, 20, 29, 74–76, 81

SG Serious Game(s). 1, 3, 9, 74, 75

TTC Trials-to-criterion. 78

Preface

This thesis was written as part of a Master's degree in Applied Computer Science at NTNU Gjøvik. The topic was conceived by myself. It was a consequence of having dived into the literature on serious games and gamification as part of course work. I was astonished at the low volume of high-quality studies, and the variance of the results of those I found. With me having personally experienced, and witnessed, the effectiveness of evidence-based educational methods from behavior analysis, I wanted to see how the incorporation of such would compare to what I had seen of serious games and gamification.

The thesis is aimed at a computer science audience. Familiarity with educational approaches can be useful, but is not necessary. However, being familiar with the React frontend framework should be helpful in understanding the prototype implementation.

I had planned this study to be a proper experiment with randomised groups, but this plan did not survive contact with the students for long. I was not prepared for the generally low level of motivation among the students, and my initial assumptions didn't hold. Together with the course responsible I was able to amend the situation, and what looked to become a meaningless experiment with less than 10 participants turned out OK.

In retrospect, I wish I had anticipated this in the planning stage, so as not to have to pivot like this. I believe the experience of the participants would have been a better one if that had been the case. Nevertheless, I was able to gather enough data to draw some conclusions from the study.

All in all it has been a valuable experience for me, both technicaly and academically. I hope this work will also provide some value to the reader.

30 May 2023

Chapter 1

Introduction

Serious games (SG) and gamification have been exploding research trends for the last couple of decades [1–5]. The former are games that have at least one other goal besides entertainment [6]. The latter on the other hand is any use, in non-game contexts, of design elements that are normally characteristic of games [7]. In this thesis, I am only concerned with SGs and gamification in the context of education.

Although the volume of research has exploded, the existing research on serious games and gamification has considerable limitations. There is low focus on investigating the efficacy of the solutions that are made, and when it is done, the quality of the studies is generally low, evidenced by an extremely low percentage of studies with moderately-to-high levels of rigor [8]. The application of known effective methodology is near non-existent, and application of known educational theory is low [8].

Adaptive learning systems (ALS) is a similarily growing research field, having grown steadily for the last two decades [9, 10]. ALSes have been defined as any system which can be considered either '...a specific platform that provides structured learning activities or sequenced learning paths; or for the purpose of targeting a specific learning population' [11, p. 1920].

The state of the research on ALSes seem to be better than with SGs and gamification. The studies that I have reviewed on ALSes still suffer from some of the same methodological issues as the former, although to a lesser degree, but a unique weakness is the thorough theoretical bias. On the other side, serious games and gamification mostly lack application of concrete theory, the theory that ALSes are based on is cognitive, and only so. The research is also usually attempting to break new ground in terms of metrics and methodology, rarely comparing the use of the ALS with normal classroom education [12].

There has also historically been a strong emphasis on adapting to individual learning styles in the ALS literature [12], which is problematic, since adapting to learning styles have long since been shown to not have sufficient evidence of positive effects on learning [13–18], and there even exists an argument for the identification of such preferences to be harmful to the learner [19].

The literature on serious games, gamification and ALSes also seem largely ignorant of the long tradition of evidence-based educational methods within the field of behavior analysis. Good and well evidenced methods for adaptive learning as an approach already exist within this field, and have existed for half a century, but they have been mostly ignored in formal education [20], which seems to also be the case in the research on these emerging modern dynamic systems [12, 21].

1.1 What behavior analysis is

'Behavior analysis is a comprehensive, natural-science approach to the study of the behavior of organisms' [22, p. 3]. The scientific discipline was pioneered by the psychologist B. F. Skinner early in the 20th century, and was built on the basic assumption that the behavior of organisms is lawful [23]. The emergent first principle of behavior, seemingly for all organisms, is that behavior is selected by its consequences, rather than being initiated by internal mental agents [22, 24].

Most behavior is governed by *operant conditioning*. Such conditioning occurs when the behavior is being continuously *reinforced* by the environment of the organism, which is all stimuli both inside and outside of it. By manipulating antecedent stimuli and the reinforcing consequences of a given behavioral response from an organism, one may alter the likelihood of similar responses in similar situations [22]. A stimulus is in this context defined as 'an energy change that affects an organism through its receptor cells' [25, p. 45]. *Reinforcement* is the increase in rate of a behavioral response, and the responsible consequence is called a *reinforcer*, which can be any stimulus [25]. Operant conditioning is the underlying principle for the evidence-based educational methods that are included in this thesis [20].

1.2 Evidence-based educational methods

The term 'evidence-based educational methods' means to educational approaches what 'evidence-based medicine' means to clinical treatments, and the term seems to have emerged as an answer to the 2002 US 'No Child Left Behind Act'. This legislation outlined a strict challenge to US educators on the degree of empirical rigor needed for the science behind educational approaches. A challenge which the field of behavior analysis was already well positioned to meet [26]. Moran and Malott [20] presented a collection of such evidence-based methods, some of which already had multiple decades history of consistently good effects compared to the normal classroom situation. Two such methods are Direct Instruction (DI) and Precision Teaching (PT).

Perhaps the best real-world example of the use of these two methods comes from the experimental school *Morningside Academy* in Seattle. Using both methods as integral parts of their foundational work with students, they have since their founding in 1980 guaranteed that their students will progress at least two grade

levels in their most deficient academic skill within one year, or they refund the tuition, a guarantee they at least up to 2012[1] have had to fulfill for less than 1% of their students [27].

Given the problems with current efforts within SG, gamification and adaptive learning systems mentioned above, an interesting question is: Can the incorporation of already proven educational methods, like DI and PT, as a foundation to serious games, gamification, or other type of adaptive learning system, be as effective or better than the existing approaches?

1.2.1 Direct Instruction

Direct Instruction (DI) is a methodology that is based on teaching generalizable strategies using clear and explicit instruction and correction on the parts of teachers, taking the guesswork out of answering questions for the students. The curriculum is carefully sequenced to maximise efficiency, and lessons are scripted so that teachers have a repeatable procedure to follow in each instance. Students are carefully grouped by current skill level so that they work on what they need to, not what they have already mastered or do not yet have the necessary foundation to learn [28].

An example of such sequencing may be to teach students the sounds necessary to pronounce 1000 phonetically regular words first, and how to blend the sounds together, rather than starting on the words themselves.

A fundamental tenet of DI is that students learn the most when they are actively engaged with the instructional content, the method is therefore built upon rapid overt responding on the part of students. DI doesn't happen in the privacy of a student's mind, they are responding to prompts in small groups. The responding happens simultaneously on a cue so that the teacher can reliably observe accuracy of each individual and, be able to determine if the group is ready to continue to the next stage [28].

1.2.2 Precision Teaching

Precision Teaching (PT) is a foundation of a standardised method of measuring learning as well as a general way to work with any sort of curriculum that has spawned many specific techniques and strategies. Central to this technology of teaching is three main points: It is that which the learner *does* that can guide the teacher, learning is measured as *celeration* and this can be charted and determined by a glance on a semi-logarithmic chart [29].

Celeration is either the *acceleration* of a frequency of a unit of behavior or the *deceleration* of it. In practice, this means that one increases the frequency of correct responses, and observe a decrease in frequency of incorrect responses.

[1] This is the most recent publication that I've been able to find on the ratio of refunds, however it is a 32-year long trend.

The chart has a semi-logarithmic scale on the vertical axis, and counts time on the horizontal axis. The frequencies of correct per minute are marked as dots and the frequencies of incorrect can be marked as crosses. Fitting a straight line through a series of points gives an estimate as to when a given skill will be mastered.

The logarithmic nature of the chart, as opposed to a regular scale, lets the teacher simply fit a ruler to the 'learning curve' of a student to determine if the *celeration* is steep enough that the student will reach the learning outcome within the designated time.

As such, this method is a decision tool for the teacher, and highly data-driven. If some approach does not give results, it is immediately visible in the chart.

Typical to all the different techniques within Precision Teaching is that the student is beating their own score in a given skill in terms of number of correct per unit of time, without having to compete with others but progress at their own fastest speed. In cases where something doesn't work, or not well enough, the data shows it, and the teacher can adjust variables for that student, or the whole class if the problem is systematic—this is arguably a level of distinction when it comes to troubleshooting that is not available to teachers within traditional circumstances.

Originating from the 1970s, the technique and semi-logarithmic standardised chart, along with strategies and rules, were developed by Ogden Lindsley in co-operation with teachers in real classrooms [30], as opposed to laboratory schools. For quite some time now, PT schools have been capable of reliably raising the performance of students by an entire grade level with 20 hours or less of instruction [29].

1.3 Study objective and findings

The objective of the present study was to investigate the viability of using principles from evidence-based educational methods from behavior analysis as a foundation for gamification. The literature surrounding gamification, serious games and adaptive learning systems didn't seem to have attempted incorporating such methods at all, so the present thesis is potentially a first attempt at this from a computer science perspective.

To this end, I implemented a web-based prototype that could import curriculum data from files made with a generic CSV template and train users in that curriculum using principles from the two methods above. DI was used as a primer for each part that the user would train, and PT was used to increase the user's fluency in each part before continuing to the next.

This prototype was evaluated through a mixed-methods study, with an experimental part and a questionnaire, yielding both quantitative and qualitative data for analysis.

Due to issues with participant recruitment and sampling, between-groups comparison of the experimental data was infeasible, however the partial results from the experimental group indicate that the system *may* have performed as well or better than the averages found for gamification, serious games and adaptive

learning systems studies. It is however impossible to determine this conclusively without repeating the experiment with better sampling.

The questionnaire, along with time series data, revealed several technical deficiencies with the prototype, that should be addressed in any future replication.

In total, this study cannot say much conclusively in relation to the efficacy of the prototype, other than there seems to be potential for future work in this direction. However, such future work should be interdisciplinary, including researchers from behavior analysis to make sure that the curriculum and programmed procedure follow the methods accurately.

1.4 Structure of the thesis

The rest of this thesis is structured as follows. Chapter 2 details the literature and reasoning behind the present study. Chapter 3 presents the software implementation of the prototype used in the experiment. Chapter 4 describes the study methodology. Chapter 5 presents the choices of analytic methods and the study results. Chapter 6 discusses the implications of the results, deficiencies of the prototype, limitations of the study and avenues for future work. Chapter 7 summarises my conclusions drawn from the study.

Chapter 2

Background

In coursework leading up to the present thesis, I investigated to what degree serious games or gamification for education, and adaptive learning systems could be counted as such, and if they incorporated any of the evidence-based methods from behavior analysis. This work was done through two separate systematic literature reviews, the first covering serious games and gamification, the second covering adaptive learning systems.

The reviews showed high variance in the learning gains achieved, and there were few studies of high quality found in either review. Additionally it didn't seem like evidence-based educational methods from behavior analysis had been attempted incorporated into these kinds of systems at all. Knowing this, I wanted to attempt implementing a system that did this and could serve as a backbone for educational gamification.

In the following sections, the results of the two systematic reviews are summarised, related work from behavior analysis where evidence-based educational methods have been used digitally is explored, and finally the research questions for the present thesis are presented.

2.1 Systematic literature reviews

I did a systematic review [8] on the empirical evidence on the efficacy of serious games and gamification for education. In this review, I filtered over 7000 papers for efficacy studies that used a controlled pre-post-test design, be it quasi-experimental or randomly controlled, that didn't have obvious statistical errors and at least *some* findings with a significantly low p-value. I found and read 14 papers.

The results from the review indicated that serious games may be more effective for learning than gamification, however both categories showed a standard deviation higher than the mean in this regard. Interestingly, most papers showed rather high relative mean score increase from pretest to posttest, the mean for control groups across the papers was 62.3%, while the mean for experimental groups

was 85.3%. In the review, I also calculated the percent-wise increase between control and experiment conditions, which often exceeded 100%, and in the extreme case exceeded 1000%.

There was a dearth of the application of concrete educational methods, theory was to a little degree applied as well. It is also worth noting the sheer minority of studies that are methodologically sound in this area. None of the papers read mentioned applied behavior analysis, nor evidence-based methods of any kind. Testing retention a period after the experiments was rare, and the same was true for the reporting of effect sizes.

Following the above review, I did a second systematic literature review, covering the evidence on adaptive learning systems in a similar fashion as the first review [12]. Adaptive learning systems are similar to gamified systems for education, but work the problem of dynamic digital teaching from a different angle, often involving adaptation to the user's emotional state or assumed position in relation to their zone of proximal development[1].

In this review, the selection was smaller, starting at almost 1600 papers and filtering down to 13 for final reading. However, contrary to the case with serious games and gamification, it was possible to derive pretest/posttest metrics only from a subset of the papers reviewed in the latter review, 5 to be exact.

The data that was possible to extract mirrored the high variance of the former review, but also the high relative increases, although the extremes weren't as big as with serious games. The mean increase for the control groups across the subset of papers was 30.8%, and the same for the experimental groups was 47.5%. The mean relative increase from the control to the experiment conditions was 130.4%.

The quality of evidence found on the efficacy of ALSes in the review was generally not high — again mirroring the former review. Studies often looked at novel methods or approaches for incorporation in these systems, but usually did not look at the format of an ALS itself as a unit for measuring efficacy. I had to extract the efficacy metrics from the quite varied studies on particular configurations of such systems, with a few exceptions.

The papers reviewed had a strong emphasis on cognitive theory as basis for the systems built or used, and the adaptive focus was often emotive—meaning that the system attempted to measure emotional states and adapt accordingly. There was no uniform way that such a measurement or classification of emotional state was accomplished. Some systems, however, did only adapt to learner achievement without accounting for emotional state. No instance of evidence-based educational methods from behavior analysis was observed.

The mean percent-wise increases mentioned for both reviews above are reproduced in Table 2.1. The values in parentheses are the standard deviations related to the neighboring value. The formula used to calculate the percentages is given in Equation (2.1). In the reviews I called this metric Learning Gain (LG). The in-

[1]The zone of proximal development is a much cited pedagogical theory, originating with Vygotsky and Cole [31], it involves keeping the student in a zone of challenge where they aren't quite able to solve the problem at hand alone, but is so with a little help form a senior

Table 2.1: Mean learning gains from serious games/gamification and ALS literature

	ALS	SG/Gamification
Mean LGC	30.8% (31.6)	62.3% (114.9)
Mean LGE	47.5% (40.4)	85.3% (153.9)
Mean LGI	130.4% (200.5)	157.8% (274.1)

crease in LG from control to experiment was calculated using the same formula, substituting the pretest mean for the LG of the control group (LGC) and the posttest mean with the LG of the experimental group (LGE), the resulting metric was called Learning Gain Improvement (LGI).

$$LG = \frac{post\ mean - pre\ mean}{pre\ mean} \cdot 100 \tag{2.1}$$

2.2 Related work from behavior analysis

Although I have not been able to find any instance of a combination of Direct Instruction or Precision Teaching in the literature on gamification or adaptive learning systems, there *have* been at least one attempt at digitizing manual methods from the literature on applied behavior analysis in a way that is similar to the system that I built in the work on this thesis.

Lovitz *et al.* [32] made a digital PT strategy called TAFMEDS, which is an acronym describing how the strategy works: 'Type All Fast Minute Every Day Shuffled'. This strategy is a variation on a well known PT strategy with almost the same acronym: SAFMEDS, '*Say* All Fast Minute Every Day Shuffled', which is a strategy that uses traditional physical flashcards in a deck, and then the deck is shuffled before doing a minute-long timed session daily to practice the deck. Correct and incorrect responses are marked as dots and crosses on a semi-logarithmic celeration chart respectively for each session, which guides the practice by showing student response to adjustments of the strategy if learning isn't fast enough, and yields a pragmatic way to predict completion time of the deck at current speed by a simple manual fitting of a straight line [29].

What the TAFMEDS variation essentially changes is only the *learning channel* from *see-say* to *see-type* (how the learner receives the prompt, and how they deliver the response). The authors created a computer system where the students were typing the responses instead of vocalising them, because, according to the authors, speech-recognition technology was neither fast nor reliable enough at the time. Ideally such a system should be able to correctly detect and classify verbal responses with a frequency of at least once per second [32].

Among other findings, the authors reported that there was a significant positive correlation between the frequency of correct responses on the TAFMEDS program and pen-and-paper application checks on the same material. There was also

a statistically significant positive correlation between the frequency of correct responses in the program and results on fill-in-blank generativity tests, where the participants had to use the trained knowledge in novel contexts. These two findings indicate that the learning channel has merit.

However, the system they created was not further adaptive in any regard, beyond automating charting and session timing, as well as digitizing the flashcard deck. Any adjustments to the learning material or learning goals was done manually by staff. No extra efforts were made to make the practice entertaining either, and daily usage—even during weekends—was *mandated* as part of course work, and social validity was understandably low. In fact, 44 out of 58 participants who responded to the social validity survey reported the method as their *least favourite* part of the course work, only a single student reported it as their favourite.

2.3 Research questions

The systematic literature reviews indicated that evidence-based educational methods may not have been employed in the making of neither adaptive learning systems, gamification nor serious games — at least not to an observable degree — in a computer science academic context.

With this thesis I wanted to attempt to investigate how well the incorporation of such methods, or at least the principles of them, would work, considered as a foundation for educational gamification, introducing concrete evidence-based methodology into that realm, while attempting to remove the need for human intervention in the on-going learning of the student. University students taking a course that I have taken previously seemed like a good candidate for the recruitment of participants, as well as ensuring that I would be able to supply some curricular material for the training.

Ideally, I would have the continuous cooperation of experts on behavior analysis throughout the development of this system, however this was not possible due to time constraints. It would require a considerable amount of time on the part of such experts, and additionally there was little time available for many iterations of the system, given that it needed to be ready for the experiment early in the spring semester for the learning of fundamental facts and definitions to make sense with course schedules. Recruitment was also a concern, since the later the experiment would run, the higher the risk of participants prioritizing exams instead of the experiment.

Given these gaps and constraints, the following are the research questions to be answered:

1. How effective is the application of DI and PT principles as foundation for educational gamification in teaching facts in a university computer science context?
2. How do university students react to the use of such a system?
3. How viable is such an application of these methods without tight project-

integration of experts on these methods?

Chapter 3

Implementation

3.1 Curriculum material

The curriculum in the experiment consisted of the definitions of basic terms related to the HTTP protocol, like the HTTP methods and response codes, and was manually created from the Mozilla Developer Network HTTP documentation[1], and put into a Google Sheet template, each topical module on its separate sub-sheet. This template, for the entire curriculum, is reproduced in Appendix A.

The idea was that the rough sequencing of the material in the program would be of the topical modules. Meaning that module 2 would follow after module 1 once the latter had been fully mastered.

Each sub-sheet was then downloaded as a CSV file, uploaded to the database server and imported into the database through a subcommand of the server executable. The system only had support for a single sequence of modules per deployment for the purposes of this evaluation.

The system was designed to be generic in this way because at the time of design it was unclear which, and how many, courses I would have access to for the purposes of the experiment. It also allowed for deployments for entirely different subjects if necessary. Using a google sheet for the templating also gave convenient access to it for those responsible for the course to review the material before the experiment.

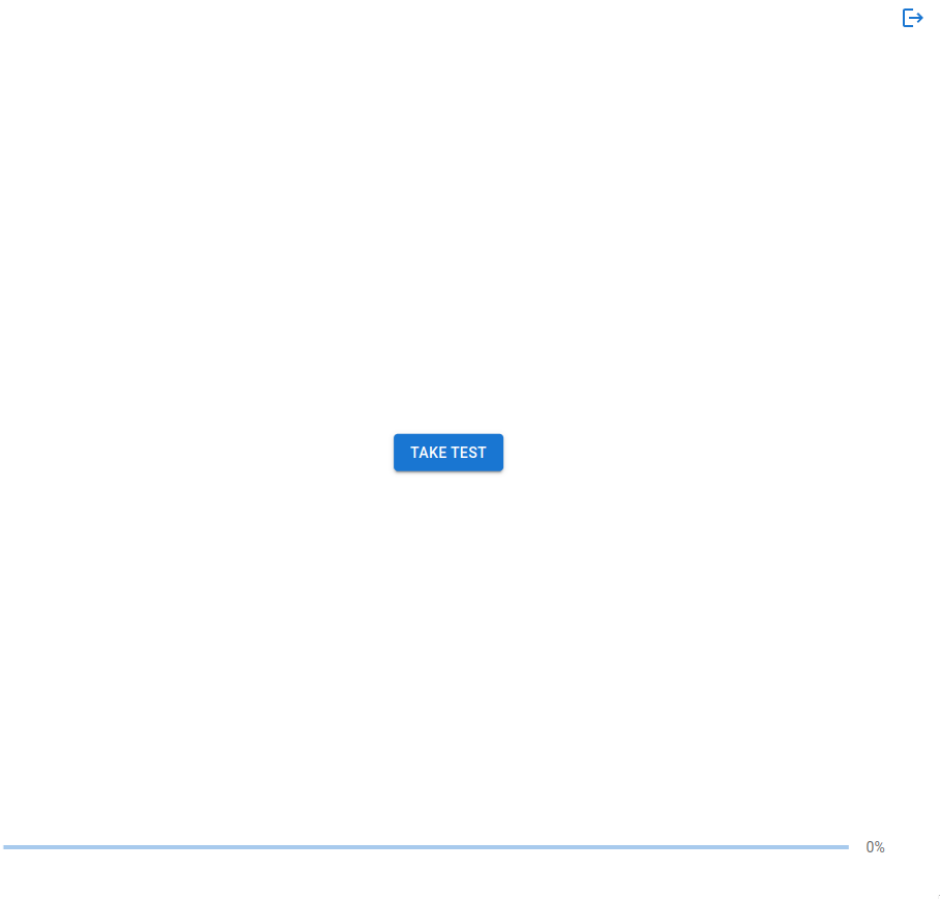
Besides the overarching sequencing established by the separation of the curriculum into modules, the system would programmatically split each module into smaller partitions for training. The size of these partitions would be adapted to the individual.

3.2 Users' interaction with the system

Besides a simple authentication system, the frontend was composed of four user-facing components based on function. These are presented below in the same

[1] `https://developer.mozilla.org/en-US/docs/Web/HTTP`

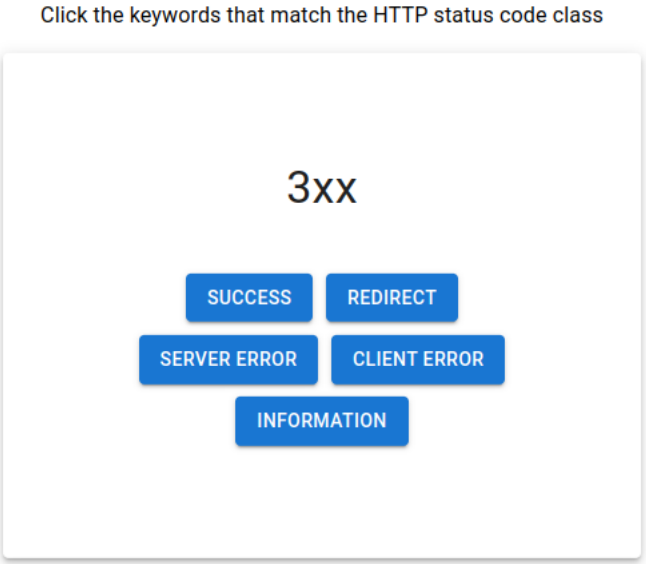
Figure 3.1: Main view with only an experimental test available



order as the user would incrementally encounter them: experimental test, DI-training, typing speed measurement and PT-training. Baked into the first few sessions of PT was a differentiating test of learning ability to determine how many pieces of information to train at a time for each individual user, assuming that this material would be unfamiliar for the users.

Once having registered an account in the system, the user was presented with a view of a single button, labeled 'Take test', shown in Figure 3.1. Upon clicking this button, the user was presented with the pretest, which covered all of the training curriculum except for reversals and other duplications. A card view with a question prompt and a set of buttons labeled with the different alternatives was presented for each prompt in the test, Figure 3.2 shows such a card view. The task of the user was to click the correct alternative corresponding to the prompt. The phrasing of the prompts and answers were the same as in the training material, however the sequence in which the user encounters the prompts were randomised, and the placement of the buttons was also randomised. Each test given was

Figure 3.2: Example of multiple choice prompt during experimental testing



randomised in this manner individually, so no two tests were completely identical, neither between students in the pretest or posttest nor between pretest and posttest for the same student. The same material in total was however covered in every test.

After having completed the pretest, the dialog shown in Figure 3.3 was presented, and the 'Take test' button was swapped out for a 'Train' button that otherwise looked the same. Clicking this started the first DI-training session for the user, introducing the first three terms in the training material. This, as with the first partition of any curricular module, was introduced with an information dialog like the one shown in Figure 3.4. The information in these dialogs were gathered from the template data supplied to the server. A flowchart covering the differentiation of training for new and returning users is shown in Figure 3.5.

The DI mode was always presented when the current curricular module partition was new to the user, and would present the questions and correct answers simultaneously, and prompt the user to type the correct answer to the question as well, as shown in Figure 3.6. When the user typed the correct answer, a bell sound effect would play and the next stage of the sequence would appear, while no sound would play on incorrect responses. The hints were present by default only for the first presentation of new prompts, and subsequent prompts would look like the one in Figure 3.7. Each time an incorrect answer was given the correct one was shown as an input hint, as can be seen in Figure 3.8. The user had to answer correctly at each stage of the sequence to continue, and the session was complete once the user had correctly answered the prompts without hints three times each. This process was deterministic. Once done, a congratulatory sound ef-

Figure 3.3: Dialog shown on completion of the pretest

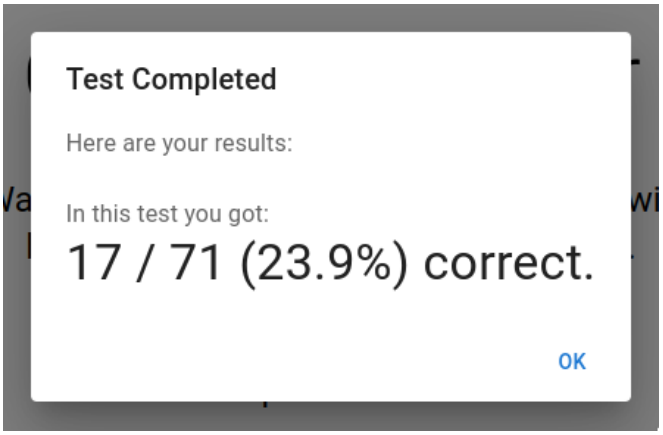


Figure 3.4: Dialog shown before starting training on a new module

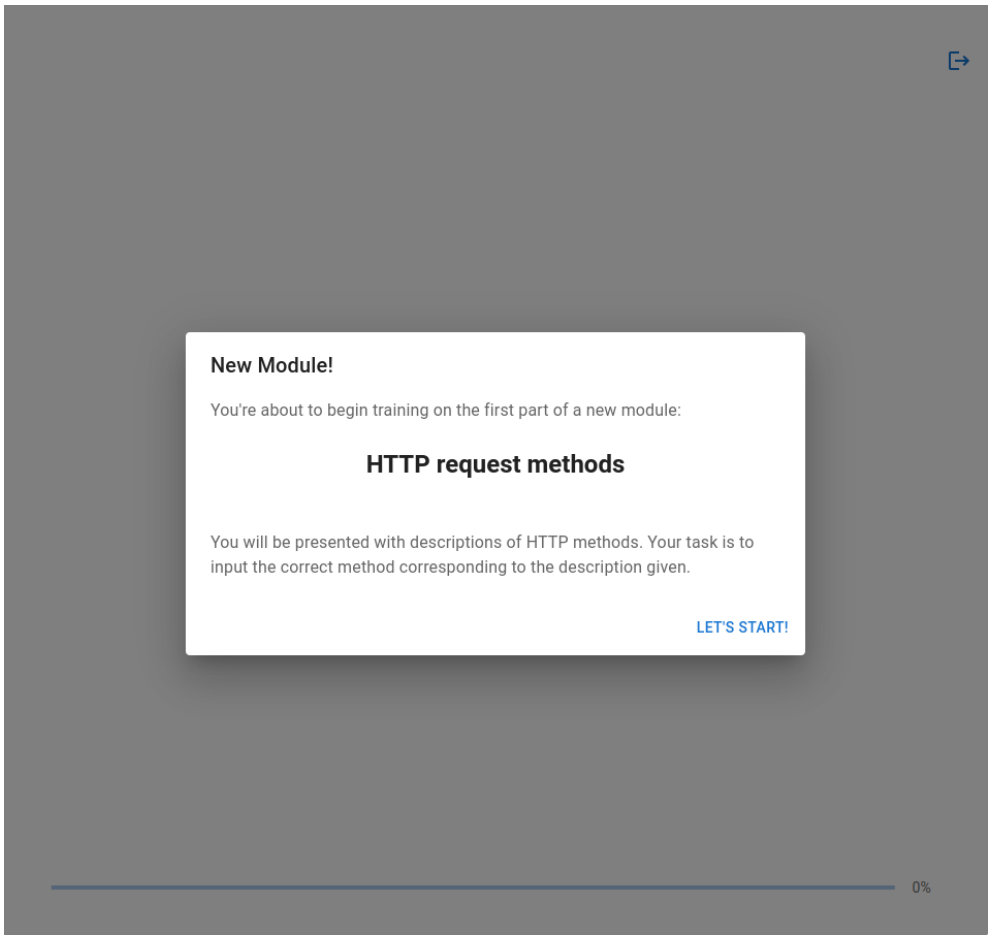
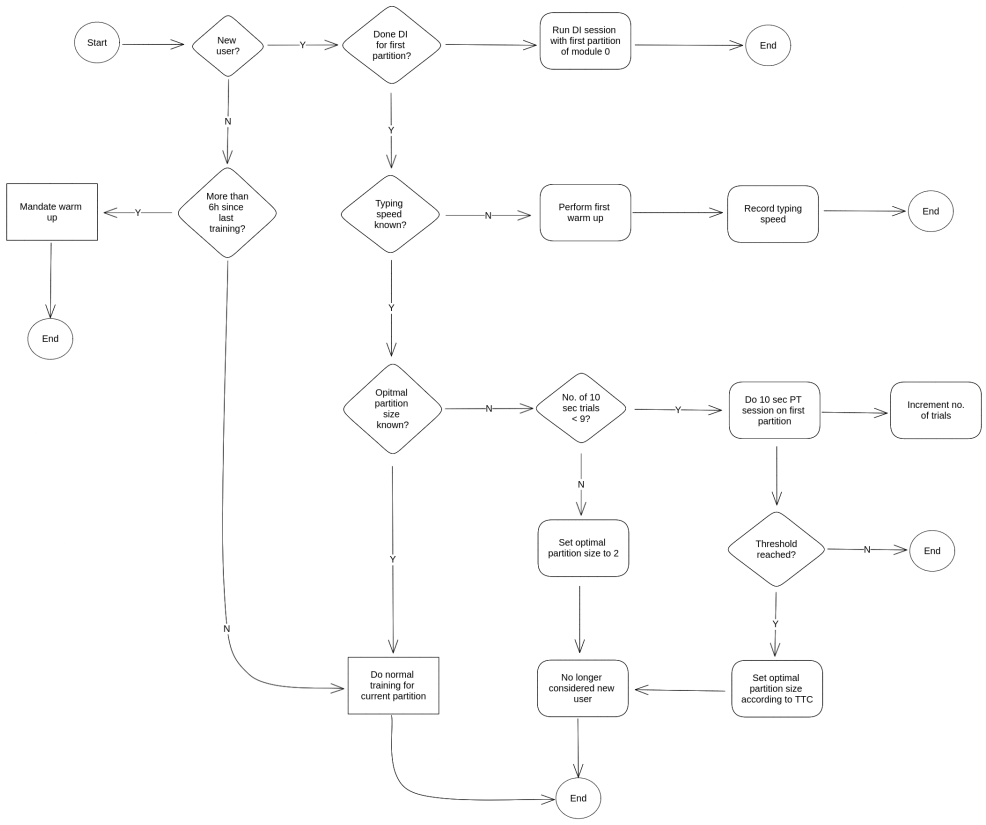


Figure 3.5: Overview of training process



fect and modal dialog window, like the one shown in Figure 3.9, were presented.

After successfully passing the first introductory DI-session, the first three terms of the learning material was considered primed for the PT-phase, called ‘speed training’ within the context of the application. However, at this point the system needed to know how fast the user was able to type, to be able to set success thresholds to values that could physically be achieved by the user. On next clicking the ‘Train’ button, the user was notified of the need to measure their typing speed via a dialog, shown in Figure 3.10, and on closing that dialog the typing speed measurement would begin.

On entering this mode, a dramatic sound effect was played until a visual countdown finished and the first word or words was shown along with a text box for entry with the challenge to type as fast as possible still visible. The countdown was similar to the one shown in Figure 3.14, albeit that figure relates to the start of PT-sessions which will be described below, the only difference was the text.

During the typing speed training, each correct entry was awarded the same bell sound as in the DI-sessions. When 30 seconds have passed, a trumpet fanfare sound was played and a dialog modal window appeared to present the user with their now first and highest score in terms of correct words typed per minute, this

Figure 3.6: Sample of DI prompt presentation

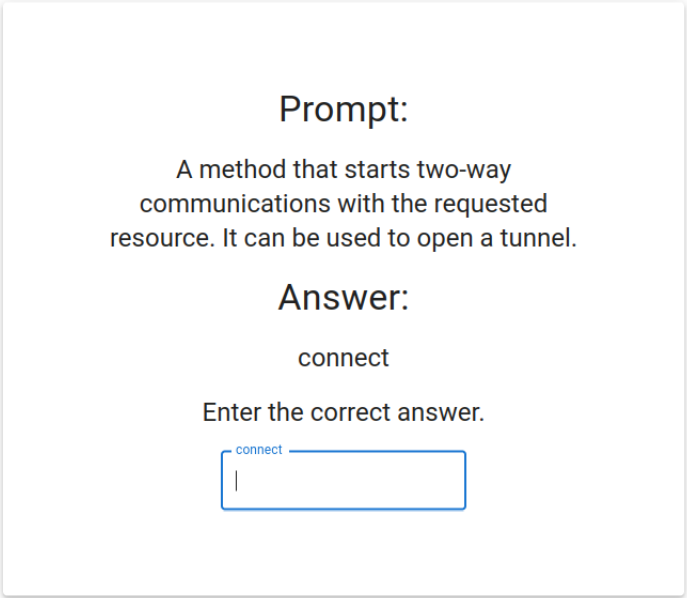


Figure 3.7: DI trial without hint

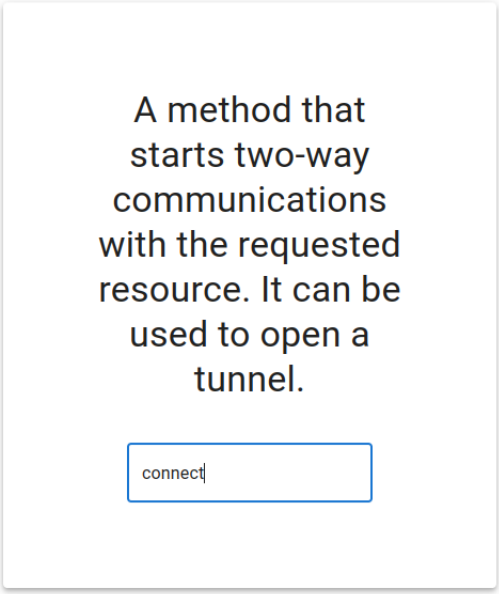


Figure 3.8: DI trial after user has made an error

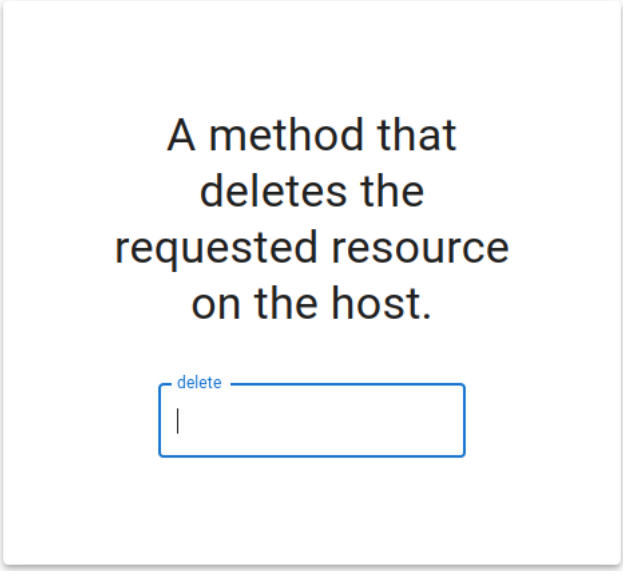
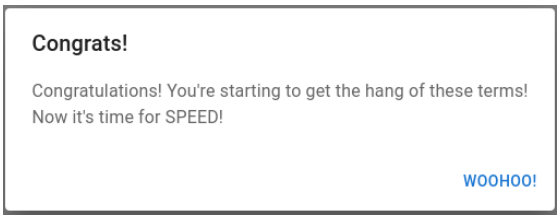


Figure 3.9: Dialog shown when a DI session has been completed



dialog is shown in Figure 3.11.

After this first encounter with this mode, a new second button was visible on the main view of the application, labeled 'Warm up'. Which started such a combined typing speed training and measurement of their typing speed. For the remainder of the experimental period the main view of the application would look like Figure 3.12.

This typing speed high-score was kept in the database, and future 'warm ups' that the user did would reflect and update their current high score in words typed per minute.

When the user had finished the first warm up, letting the system have a first

Figure 3.10: Dialog shown before measuring typing speed for the first time

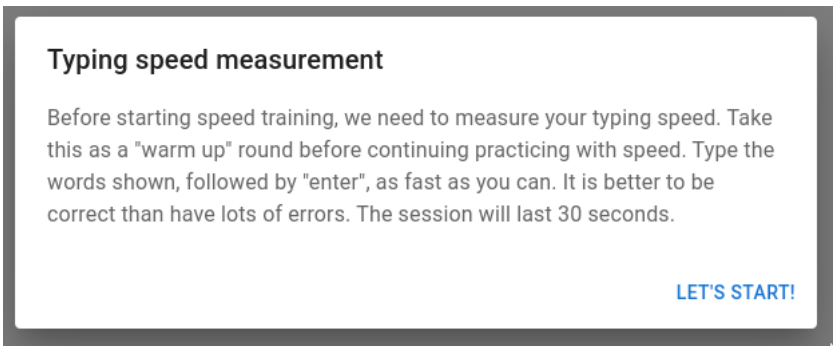
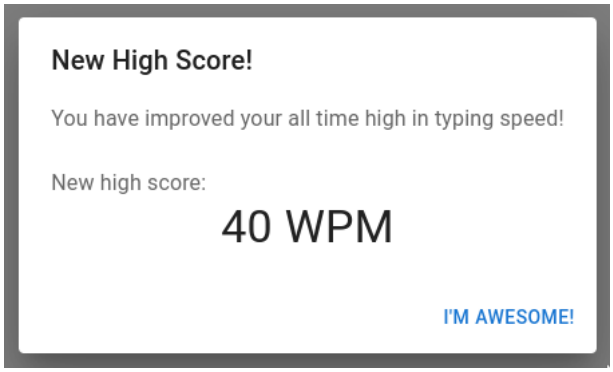


Figure 3.11: Warm up high score dialog

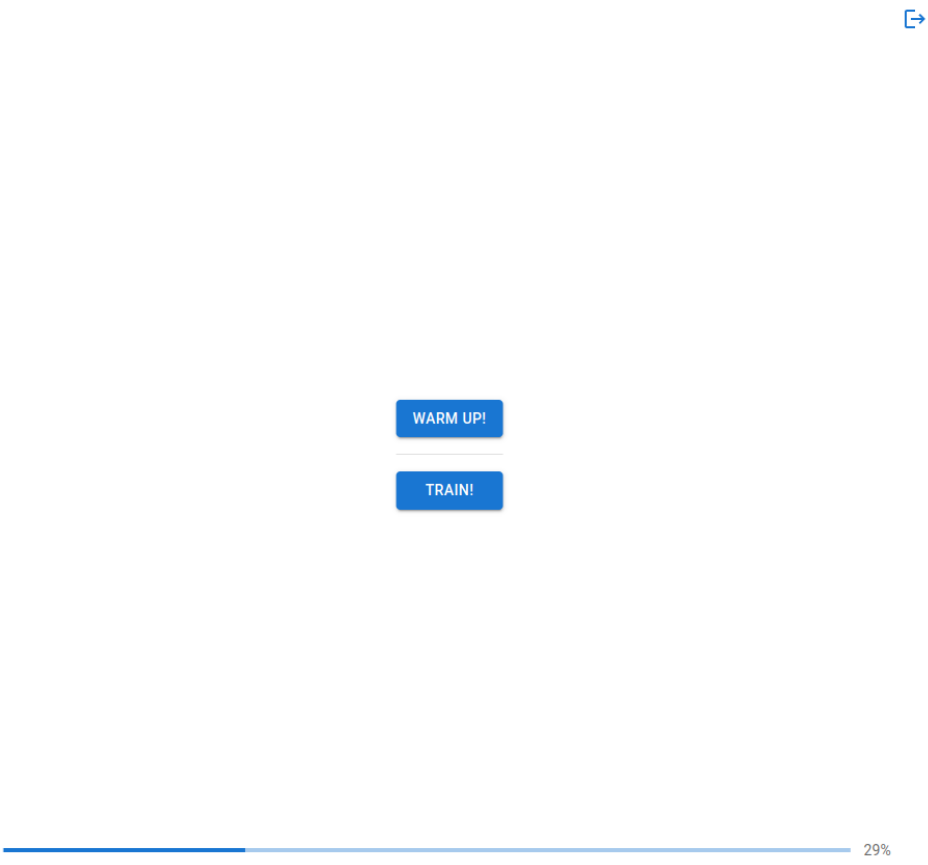


data point of their typing speed, the 'Train' would allow PT training. Up to the first 9 such sessions would however be a calibration stage, where the system assessed the user's general learning ability based on the number of trials needed to reach the mastery threshold. These first sessions lasted 10 seconds each, contrary to the later normal of 30 seconds, and would continue being offered until the user either reached their threshold or had done 9 unsuccessful attempts. The user had to click the 'Train' button between attempts, and so could take breaks. Before each such calibrating session, the explanatory dialog shown in Figure 3.13 would be presented to the user. This would adjust how much new material that the user would encounter at a time.

The speed training itself was visually similar to the warm ups, having a countdown, like the one shown in Figure 3.14 and a high score. Instead of random words, the questions in the partition would be presented, and the user would have to enter the correct answer. The normal card view of such prompts would look like Figure 3.15, and if the user made a mistake.

On a correct response, the bell sound would play, and on incorrect responses no sound would play, but the correct answer would be shown as an input hint, as shown in Figure 3.16. In this mode the system would track incorrect and correct responses, and present the results to the user at the end of a session, each session

Figure 3.12: Main view of frontend



was 30 seconds long. On a high score a fanfare sound effect would play, and a congratulatory dialog like in Figure 3.17 would show. The mastery threshold was also shown, telling the user how high their score needed to go to be able to continue to new material in the curriculum. If a high score was not achieved, a comical dissonant fanfare sound effect would play as the results were presented in a dialog like the one shown in Figure 3.18. To be able to continue to the next stage in the curriculum, the user would have to meet their individually calculated fluency criterion, and have no errors in a single session. Upon the next click of the 'Train' button, the system would train the next partition of the module with the DI-mode. A flowchart of this process is shown in Figure 3.19.

In the case of the initial calibrating sessions, if the user reached their threshold within the 9 sessions, they would be congratulated with the dialog shown in Figure 3.20. In the case that they did not, on the 9th attempt, the dialog in Figure 3.21 would be shown, and their training would be reconfigured to only work with the

Figure 3.13: Dialog shown before beginning a 10-second learning speed measuring PT session

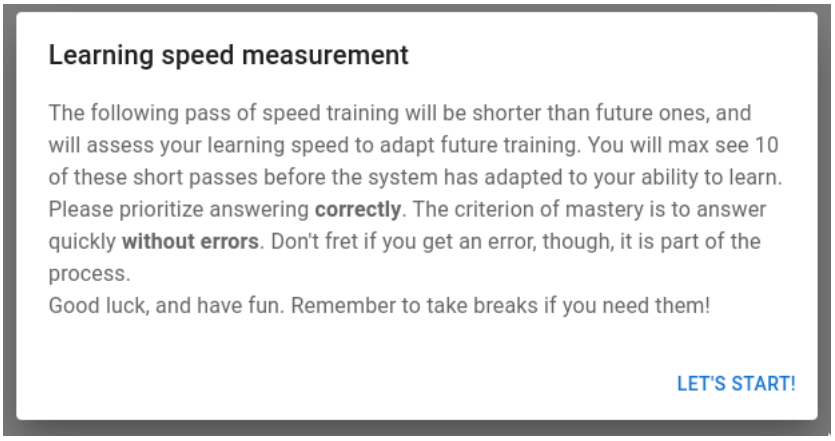
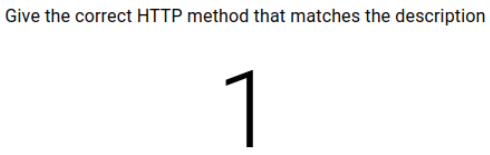


Figure 3.14: Visual countdown before start of PT session



first two terms of the material, and on their next training session they would be going through the DI session again for these two terms only.

For each achievement in the application, a progress bar at the bottom of the screen would gradually fill up, and upon completing the entire curriculum it would be at 100%. Completing the learning speed test and initial typing speed measurement each counted as separate achievements, and would bump the progress by a relatively large amount initially. Each completed curricular module then counted as an achievement for the progress bar after this.

Figure 3.15: Normal PT trial

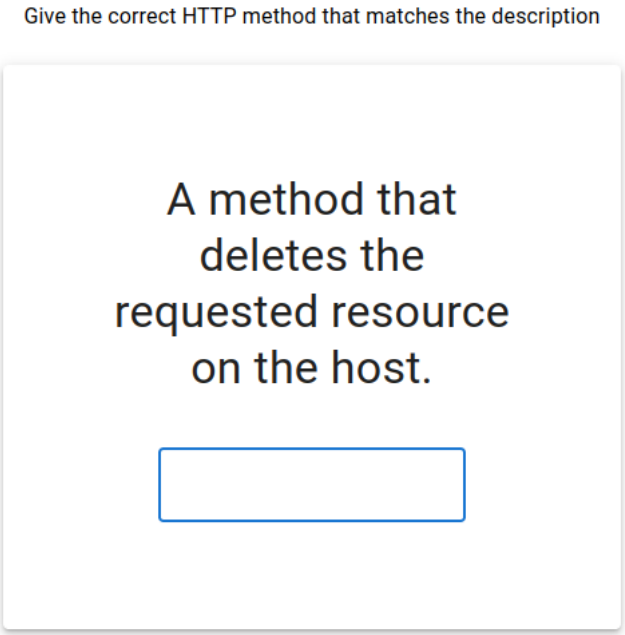


Figure 3.16: PT trial after a user has entered a wrong response

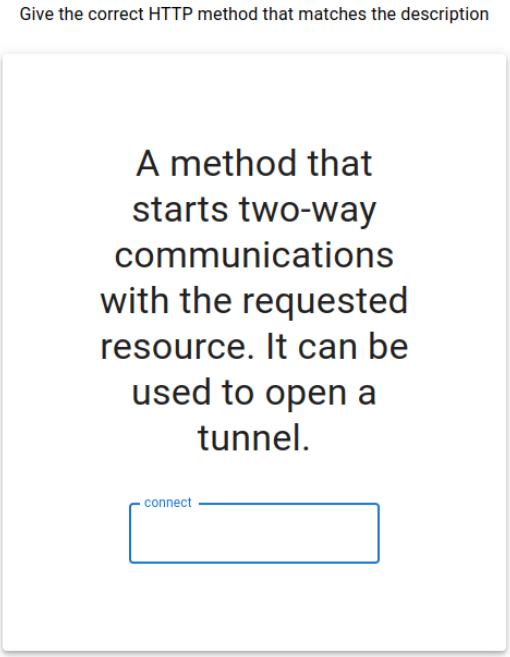


Figure 3.17: Dialog after PT session where the user beat their high score

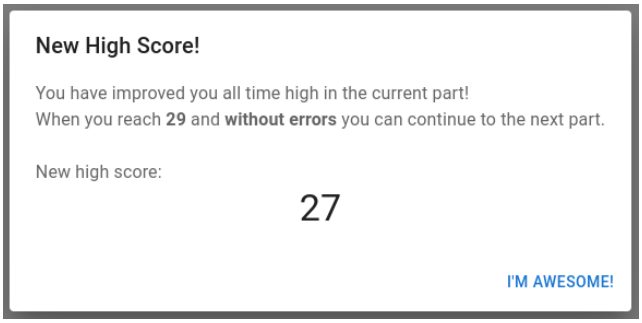


Figure 3.18: Dialog after PT session where the user didn't beat their high score

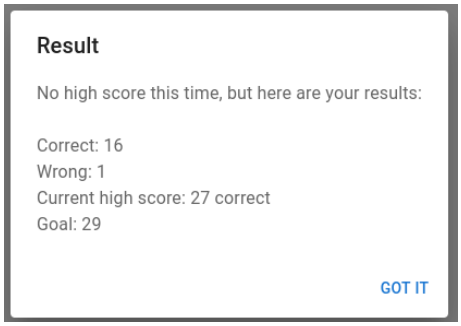


Figure 3.19: Overview of PT process

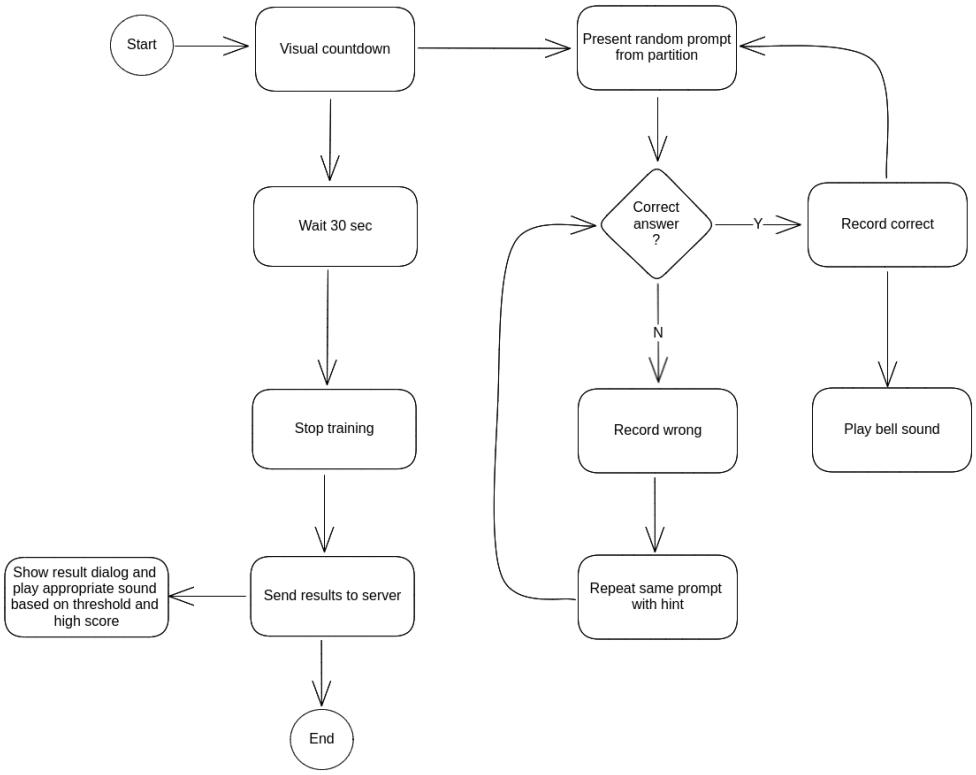


Figure 3.20: Dialog shown when trials to criterion has been established

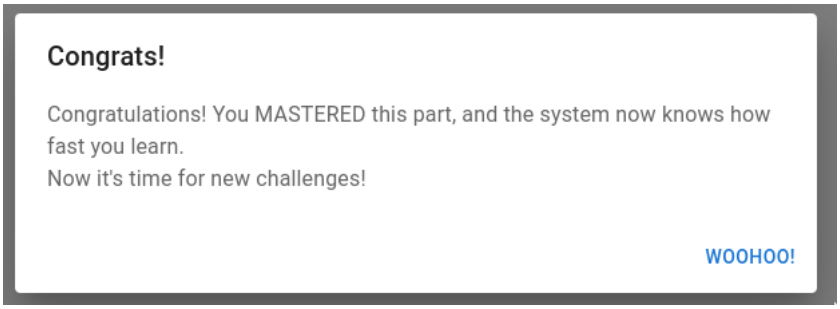


Figure 3.21: Dialog shown after 9th unsuccessful try on the learning speed measurement phase

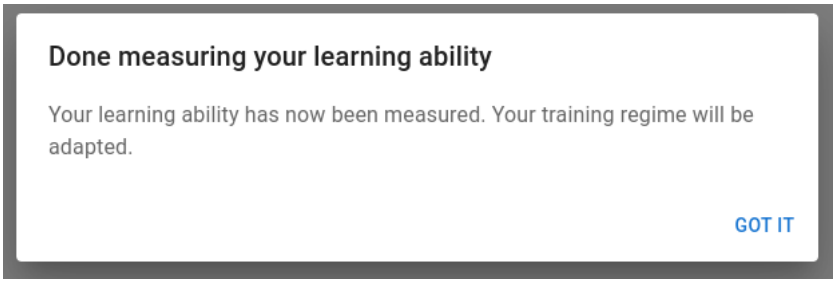


Table 3.1: DI sequence with no errors

Step	P_1	P_2	P_3
1	c*		
2		c*	
3	c		
4		c	
5			c*
6	c		
7		c	
8			c
9			c
10			c
11	c		
12		c	
13			c

3.3 Procedural details

The sequencing of the prompts/questions within the DI-sessions was implemented based on a queue of counts for each prompt, the counts being the number of unassisted correct responses the user had given to a particular prompt. When not presenting a new term, or if the user just made a mistake, the term with least current unassisted correct responses would be picked. Prompts that had been correctly answered to the needed degree minus one, would be removed from the overall queue, so as to cycle through the partition of prompts, before making a final pass to achieve at least 3 unassisted correct responses for each prompt. The ideal scenario where the user makes no mistakes during this phase is shown in Table 3.1. In the table, asterisks refer to the first presentation of a term, which would also contain the correct response and an accompanying input hint, like shown in Figure 3.6. The character *c* stands for 'correct'.

In cases where the user made mistakes, the application would still mandate at least three *unassisted* responses, but in the case of wrong responses, would show assisted prompts, like the one shown in Figure 3.8, and repeat it until the user answered correctly, then move to the next prompt in the queue before revisiting it later with another unassisted trial like shown in Figure 3.7. This was the case both when the user has a string of leading errors before starting to answer correctly, as shown in Table 3.2; and when the user has errors intermixed with correct responses, like shown in Table 3.3. The character *w* stands for a wrong response, and the subscript *a* stands for assisted prompt in both the latter tables.

The words used in the warm ups were picked at random from a list of the 100 most common English words, except for single-letter words, and up to two words were displayed at a time to mimic the possibility of an answer within the learning material having more than a single word in it. For purposes of adjusting

Table 3.2: DI sequence with a leading chain of errors

Step	P_1	P_2	P_3
1	c*		
2		c*	
3	w		
4	w_a		
5	c_a		
6		c	
7	w		
8	c_a		
9			c*
10		c	
11	c		
12			c
13	c		
14			c
15			c
16		c	
17	c		
18			c

Table 3.3: DI sequence with intermixed error

Step	P_1	P_2	P_3
1	c*		
2		c*	
3	c		
4		c	
5			c*
6	w		
7	c_a		
8		c	
9	c		
10			c
11			c
12		c	
13	c		
14			c

Table 3.4: Advised max number of new items to learn at a time based on trials-to-criterion

TTC	No. of items
2	7
3	6
4	5
5	4
6	3
7	3
8	3
9	3
> 9	2

the mastery threshold in the PT-sessions, the 3-long moving average in characters typed per minute was used to account for regression toward the mean.

The formula for this adjustment of the success threshold is shown in Equation (3.1). c is the moving average of the user's typing speed in characters per minute; a is the mean character length of answers in the currently trained partition; d is the duration of the session in seconds; u and k are constant tuning parameters used to further dampen the threshold to accommodate the user's assumed unfamiliarity with the material and reaction time, respectively. The final values for the latter two after tuning were: $u = 0.7$ and $k = 0.8$. The tuning was done by me training unfamiliar parts of the material myself and finding values that gave a challenge while not making the threshold seem unattainable.

$$
\begin{aligned}
t &= \frac{cu}{a}\frac{d}{60}k \\
T &= \left\lfloor \begin{cases} t, \frac{t}{d60} \leq 60 \\ \frac{60}{d}, \frac{t}{d60} > 60 \end{cases} \right\rfloor
\end{aligned}
\tag{3.1}
$$

The purpose of the calibrating stage was to adapt to slow and fast learners by presenting less new material at a time for slow learners and more to fast learners. An expert in DI and PT was consulted who gave the rules of thumb, reproduced in Table 3.4, regarding the max number of new items to learn in one go with the method. Since 3 such trials to criterion was considered by the expert as the normal for most people, this was set as the default for the calibration stage. If the user failed to reach mastery within the 9 sessions, their optimal number of items would be set to 2, since this is the absolute minimum. The absolute max would be 7. An instant success would be regarded as a bug or extensive preexisting skill, and the system would fall back on the default 3 new items at a time.

3.4 System architecture

3.4.1 Software architecture

The architecture of the system was designed to be scalable, but still be simple to deploy on a single machine. The backend was written in go and utilised the *gin*[2] framework for simpler HTTP handling than default. The frontend was made in *React*[3] for the user interface, however it was embedded in the go backend executable, which served it on a subset of frontend-facing HTTP endpoints. The database used was *MongoDB*[4], spun up in a container on either the same machine as the backend or a separate machine, and could potentially be exchanged for the sharded MongoDB cloud service if need be.

Figure 3.22 shows an overview of the architecture of the go server executable, and how the interaction with the client and the database flows through the modules of the backend.

The server executable was responsible for handling requests from the client. In the case that such a request was of the method GET against the root URL endpoint, the server would reply with the frontend code as the payload. All requests for endpoints starting with `/api/` would be routed to an appropriate handler by the `api` package.

The system utilised eight API endpoints:

- Utility endpoints
  - `/api/register`
    - Handled the registration of new users
  - `/api/auth`
    - Handled login, logout and periodic authentication status checks
  - `/api/recover`
    - Handled user account recovery requests
  - `/api/status`
    - Handled periodic state requests
- Business logic endpoints
  - `/api/testing`
    - Returned randomised items for an experimental test on GET requests
    - Handled experimental test results from POST requests
  - `/api/training`

[2] `https://gin-gonic.com/`
[3] `https://react.dev/`
[4] `https://www.mongodb.com/`

  - Returned relevant training data for upcoming training session on GET requests
  - Handled training session results on POST requests
- `/api/warmup`
  - Handled warm up session requests. Storing results on POST
- `/api/lst`
  - Stands for Learning Speed Test (an artefact of an earlier iteration of the prototype)
  - Handled POST requests for storing the initial 10 sec session results, and updating the learning profile of the user when their TTC had been established.

The frontend in turn was a React single-page application. Meaning that it controlled any client-side routing of URLs that didn't map to the backend API. A diagram of its architecture is shown in Figure 3.23.

The main component of the frontend was the client-side router, routing the user to the different pages of the application. The `Main` page handled everything not related to registration, recovery and authentication. The authentication state was handled by the router, while all other client-side application state was handled by the main page root component. Each sub-component for the different functions contained its own logic, but the root component would communicate the results to the backend API.

State changes from the user's interactions with the application would toggle the view between the different subcomponents. There was one for each of the major functions of the application, and several separate dialog components to handle the subtle differences in their content.

The remaining pages were pretty similar to each other, and had only minor differences in terms of visible text inputs buttons and labels. They were effectively web forms for handling authentication.

3.4.2 Infrastructure

For the purposes of the present experiment, the system was deployed as an arrangement of virtual machines in the university private cloud. One database node running the MongoDB container and several worker nodes serving the React frontend and REST API for the application from behind an *OpenStack*[5] load balancer. A graphical overview of this setup can be seen in Figure 3.24.

This can in hindsight be said to be too much, and a single machine could most likely have run the experiment without special issues. The choice of a constellation like this was made due to my lack of experience in running production systems with real users — I did not know how demanding it would be, so I wanted as good a guarantee as I could get that the service would not be unresponsive for

[5] `https://www.openstack.org/`

the users, and I had the resources available. This choice did however give me the ability to fix a couple of early issues without taking down the service since I could do incremental deployment of the new code to each of the workers in turn.

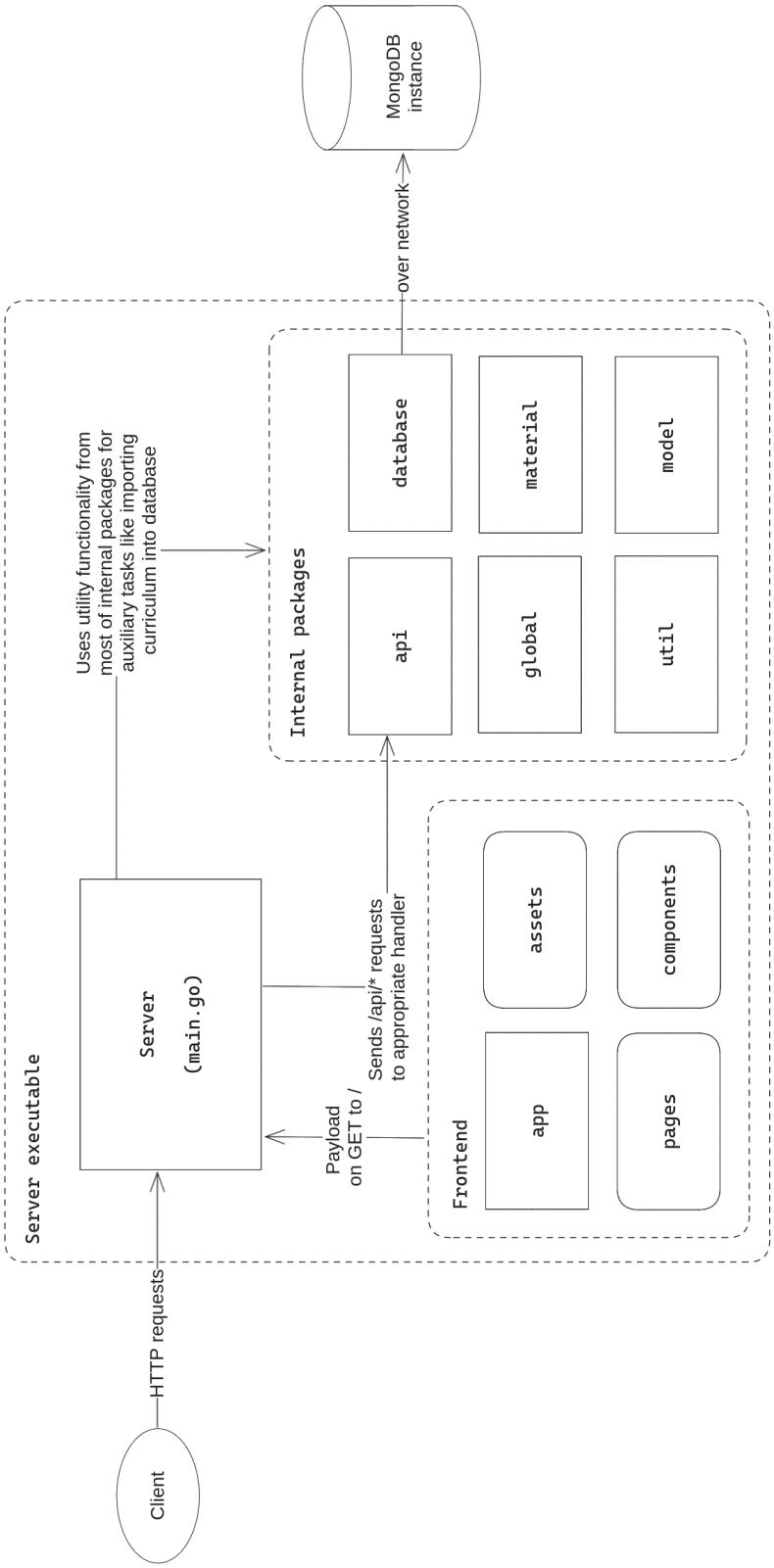


Figure 3.22: The server executable and its modules

Figure 3.23: Overview of the frontend architecture

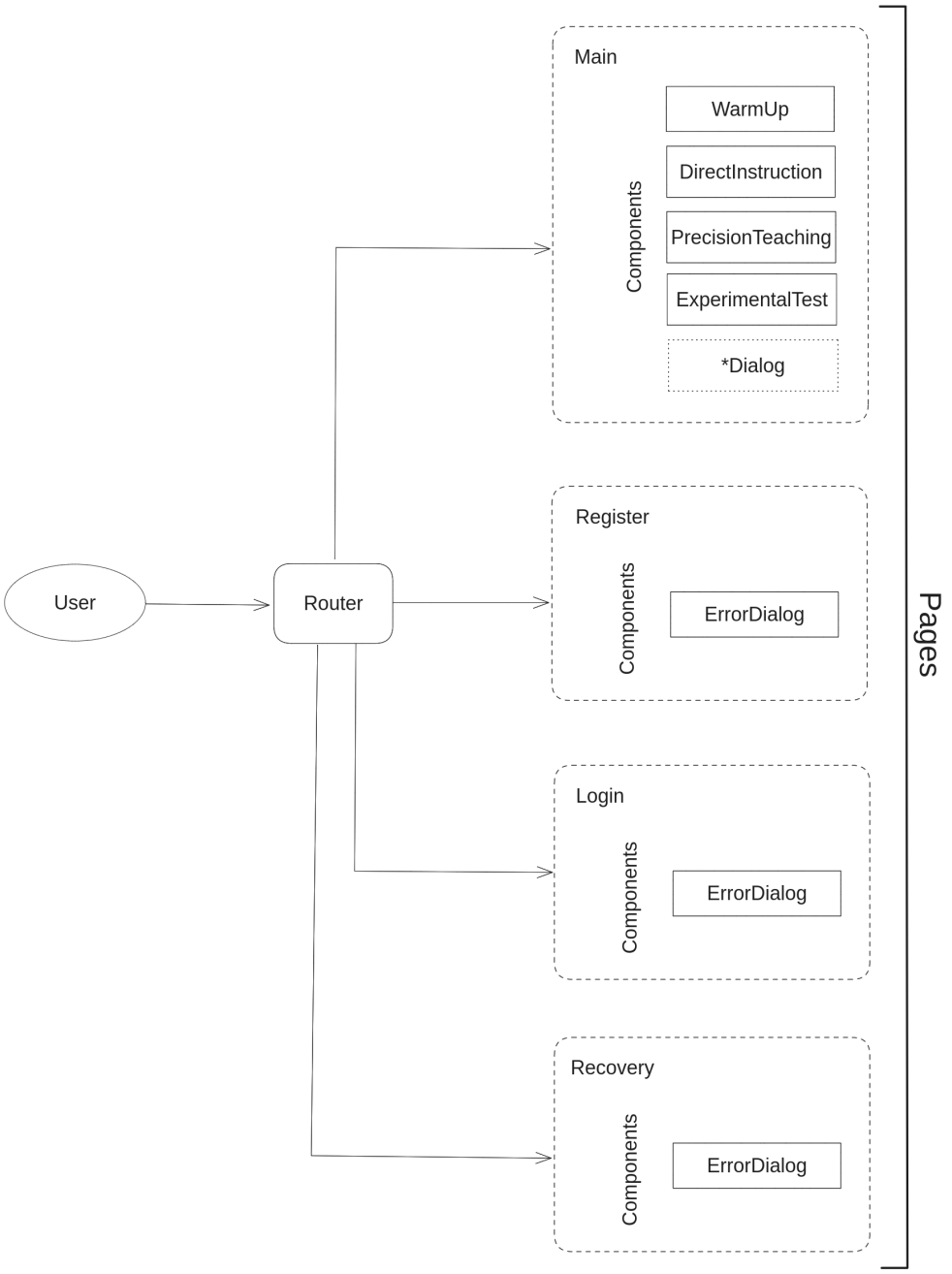
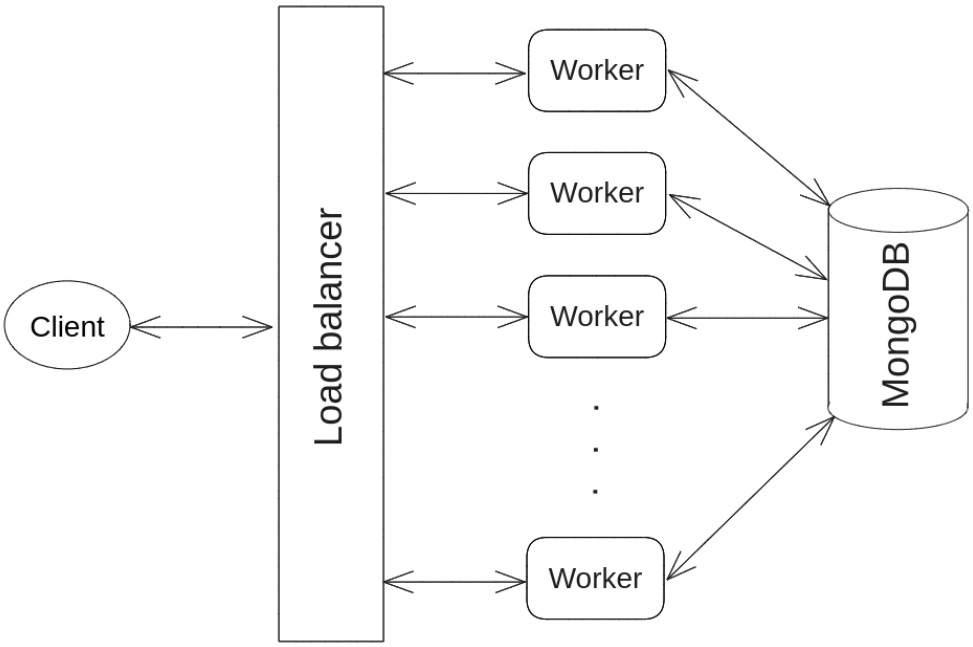


Figure 3.24: Overview of the system infrastructure



Chapter 4

Method

4.1 Participants and setting

The participants in this study were university students at the bachelor level in Norway. The study was conducted in conjunction with a cloud technology course with a total of 86 students coming from four different study programmes, three of which had the course as a mandatory part of their programme. Of these, up to 81 took the pretest[1], but only 29 participants completed all the tests as well as the concluding questionnaire.

Participation was anonymous and voluntary, but it was incentivised with course credits making up a potential 4% of the portfolio part of the course grade, which itself constituted 60% of the total course grade.

The experiment and prototype was presented in an early lecture of the course, just before the course was to start covering the material that was included in the training within the prototype. After this first presentation, time was allocated for taking the pretest before the lecture was continued. Use of the system after this was left to the participants' own management: they were free to use it as much as they liked in their own time for the next four weeks. However, I was forced to make a methodological change to the experiment a week into this period, so for the first week, only a subset of the participant had the training available. This is detailed further below, and in Section 4.2.1 especially.

4.2 Design

This study used an explanatory sequential mixed-methods design [33] with two parts. The first, a quasi-experimental evaluation of the effectiveness of the system in teaching facts relating to the course in which the experiment was ran. The second, a questionnaire combining likert-scale and free-text responses to judge

[1]The anonymous nature of the participation makes it impossible to account for duplicate accounts

the participants' reactions to the system and shed light on any common issues encountered.

This mixed-methods design was chosen for the sake of completeness, complementarity and triangulation. Specifically, to exploit the combination of quantitative and qualitative data to more fully address the research questions, compensate for low sampling and discover areas where the qualitative data converges with the quantitative to strengthen conclusions [33].

Initially, it was assumed that I would be able to recruit volunteers for randomised groups for the experimental part without additional incentives, however the initial participation was so low that the sampling was changed and course grade credits were announced for participation a week into the experiment. This obviously affected the validity of the study, which will be further discussed in Section 4.2.1 and Chapter 6, however it ensured enough data for meaningful analysis.

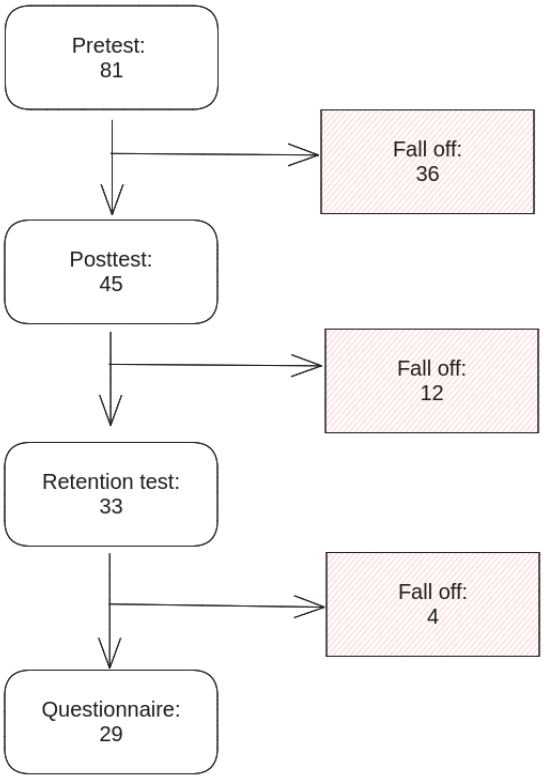
The experimental part was a pre-/posttest design with an added retention test. The experimental and control groups were convenience sampled based on usage of the training system. The control group was the students who didn't use the system, but still completed the experimental tests. Except for the comparison of the pre- and posttests, any participants who didn't complete all tests were excluded from the rest of the data analysis.

In total, there were 81 accounts in the system that performed the pretest. From this number there was a considerable fall off during the study. Of the original 81, 45 completed the posttest, and of these 33 completed the retention test. The 33 that completed the retention test were offered the questionnaire, which 29 completed, Figure 4.1 shows the detailed fall off between each of these points.

The experiment was divided into two phases: an experimental phase and a retention phase. During the first phase, participants were freely able to register user accounts in the system and were presented with, and had to complete, the pretest before being allowed to train. When the first phase was over, the system locked down. No new accounts could be made, and existing users were not allowed to train, but were presented with the posttest. The first phase lasted in total 4 weeks, however, due to the sampling change detailed in Section 4.2.1, only a few of the participants trained during the initial week. The second phase was a retention phase were the system remained locked down for another three weeks before another experimental test was issued, measuring how well the participants remembered the knowledge gained from the training. When the students had completed the retention test, they were prompted to complete the online questionnaire evaluating the participants' reactions to the system and experiment. After this the system was opened up again for anyone to register and train, although no course credits were offered for training after the conclusion of the experiment, and no data was included in the analysis for any training after the experimental phase.

The experimental tests measured the participants' knowledge of HTTP methods and status codes in the form of multiple choice without time limit. The degree of correctness and time spent was recorded in each instance. These tests were conducted within the system itself.

Figure 4.1: Participant fall off during the study



Statistical analysis was conducted on the results of the experimental tests, comparing the gains between pretest and posttest, and loss between posttest and retention test. The intention was also to compare the results of those who used the system and those who didn't for determining efficacy of the system as an educational tool, however this was not feasible due to the small size of the control group.

Concerning the questionnaire, confidence intervals were calculated for the likert-scale responses and a thematic analysis was conducted for the free-text responses.

4.2.1 Sampling change and grade credit incentive

This study was initially planned being a randomly controlled experiment, based on voluntary participation without external incentives. However, the participation of those in the experimental group within the first week of the original experimental period was so low that meaningful data analysis would have been impossible. Many students had created user accounts and completed the pretest the first day of the experiment, but after a week only a few had started using the system. The system was thus opened up to everyone for the remainder of the ex-

perimental period, and credits toward the course grade was offered as incentive for participation. The credits were given in equal parts for simply completing the experimental tests and concluding questionnaire, and proportionally for completion of the training material. Anonymity in the experimental context was upheld by the course responsible being supplied with a list of usernames from the system, paired with credit scores, and having to pair the usernames with actual people on a voluntary basis by collecting this information separately from the students.

4.3 Data collection

4.3.1 Experimental tests

Each experimental test was a multiple-choice test without time limit containing the same 71 items, but in a random order. The content of the tests used the same phrasing as the material covered in the training, although duplications like reversals were removed from the testing material. E.g. in training, a participant would encounter $A = B$ as well as $B = A$, but in testing would only encounter one of either. Furthermore, the channel of response was different in testing than in training. During training, the participants had to type out the correct responses, while in testing they had to click the correct alternative among a selection. As mentioned in Section 3.2, the tests were also randomised so that the sequence of topics, as well as the within-topic sequence of items, was different for every participant, and for the same participant between tests.

The purpose of these tests were to gather data that could be used to say something about the efficacy of the tool for teaching the facts involved. Accuracy on the tests would give a measure of how much of the curriculum that a participant was able to learn from the system, and any difference in time spent was thought to be able to show any change in fluency, or mastery of the material.

4.3.2 Questionnaire

Screenshots of the questionnaire, as it would have been seen by the participants, are shown in Appendix B, but an overview of its contents is provided in the present section.

The questionnaire had likert-scale responses for specific questions around the topics of motivating factors, usability factors, external factors that could affect the experimental results and the diligence with which the participants completed the experimental tests. These were mandatory and divided into topical sections. A free-text field was available for an optional elaborative response at the end of each such section. Concluding the questionnaire was a final optional free-text field that prompted any other feedback that the respondent may have on the system.

The questionnaire itself was divided into 6 pages, with one topic on each. The below list is ordered by the pages. Their contents are described as bullets underneath each page.

1. Questionnaire code
    - The participants were asked to supply a numeric ID that they would have received on completing the retention test. This was used to pair the questionnaire submission to the training data for the purpose of granting course credits for participation,
2. Measuring the motivating factors of the system
    - This was a set of likert-scale prompts with the following common description:

        > Please rate the statements in terms of how they apply to you.
        > "Phase 1" is the first walkthrough of each new part of the material, where new terms are introduced with "Prompt" and "Answer" headings.
        > "Phase 2" is the speed training.

    - The likert-scale was 'Strongly disagree', 'Disagree', 'Indifferent', 'Agree' and 'Strongly agree'. This scale remained the same for such prompts on all subsequent pages, except page 5.
    - The statements were as follows:
        - I found the system engaging
        - The Phase 1 of training was not boring
        - The phase 2 of training was repetitive
        - I enjoyed the system overall
        - I would like to use a system like this for other subjects
    - There was also a concluding free-text field for additional feedback relating to motivating factors
3. Measuring the usability factors of the system
    - This page had two likert-scale statemets and a concluding free-text field for additional feedback. The statements were as follows:
        - I think that the system is convenient to use
        - I made too many typos
4. External factors
    - This page had only one likert statement, and the usual free-text field for additional information. The likert statement was:
    - I used memorization techniques outside the DIPT system to practice the material
5. Measuring seriousness of the experimental tests
    - This page was an attempt at measuring to which degree the participant was diligent in doing the experimental tests. The description on the page detailed this, and gave three likert-scale propmts for seriousness, one for each test.

- The likert-scale was different than previous pages: ‘Quite serious’, ‘Serious’, ‘Indifferent’, ‘Not serious’ and ‘Skipped as fast as possible’ were the alternatives.
- There was no free-text field on this page

6. Other feedback
    - Contained a free-text field and an admonition of giving any other feedback not covered earlier in the questionnaire.

The questionnaire was offered to all participants who had completed all experimental tests.

Chapter 5

Data analysis and results

5.1 Data analysis

With regard to the data from the experimental tests, participants' achievement scores for, as well as the time they spent on, the pre-, post- and retention tests were analysed. The comparison between pre- and posttest was done for all participants who had completed both tests, regardless if they subsequently dropped out of the experiment or not. A second comparison between pre- and posttest, as well as between posttest and retention test, was done on the subset of the first grouping that had also completed the retention test.

Regarding the questionnaire, the analysis was two-fold. The likert scale responses were interpreted to find where the majority of respondents were positioned. Confidence intervals were calculated for the proportion of affirmative, negative and neutral responses for each likert statement or question, so as to determine overlaps which would make the result inconclusive. The responses on the free-text fields were not clearly divided among the different sections, often commenting on themes covered by future sections or previous sections, therefore the free-text responses are covered collectively, rather than per-section. The analysis for the latter was done using a qualitative thematic analysis to discover details not covered by the likert-responses and possible explanations for them. This is described in Section 5.1.3.

The following subsections will detail what specific quantitative metrics were statistically analysed, and how. The qualitative analysis method will be described in the final subsection.

5.1.1 Experimental metrics

There were three main quantitative metrics used for analysing the experimental data. The first was the test score, measured as percentage correct on each experimental test, which was used to compare the median of individual change within group between the pretest and posttest as well as between posttest and retention test. This was the chief efficacy metric for the study.

Second, the time spent in seconds on each test was also recorded to measure any change in the participants' speed between them. The medians of these durations were also compared within groups. This was the secondary efficacy metric for the study.

Third, score rate was an aggregate metric, translates to test score percentage points per second spent on the test. This was used to determine any connection between score and duration results.

Lastly, to give some grounds for comparison with the reviewed literature on serious games, gamification and adaptive learning systems, the individual percentwise increase in score between tests was also calculated where the change in score for the group was determined to be statistically significant.

5.1.2 Statistical methods

What assumptions can safely be made about the data colors the choice of statistical methods employed. The pretest score data was early determined to not be normally distributed using a Q-Q plot along with histograms, shown in Figure 5.1. Furthermore, the histograms of the pre- vs posttest scores shown in Figure 5.2 indicated that the skewness was flipped between the two tests, which excluded nearly all non-parametric tests. Since there were such significant outliers in terms of pre-existing knowledge on the pretest, I regarded the data of the posttest as dependent on the pretest, and so wanted a paired test. The paired sign test was chosen since it makes no assumptions about the data [34]. This was used, with an $\alpha = 0.05$, for all significance tests on the data from the experimental tests. Python was used to filter the data and perform the tests.

For the questionnaire, confidence intervals, at 95% confidence level, were calculated using a python implementation of the Clopper-Pearson approach [35], which is a well known and much used method of calculating confidence intervals for binomial distributions [36] and I know it has been used with success in a specific situation very close to mine regarding likert response proportions and the size of n [37]. This was done to determine general applicability of the results on a similar population.

Beyond the statistical tests, the data was visually and numerically inspected for possible further implications.

5.1.3 Thematic analysis

Thematic analysis is a method of analysing qualitative data, which comes from psychology. However the method is not tightly tied to only that field, it is generally applicable to qualitative data like interviews, free-text survey responses and so on [38]. Clarke *et al.* [38] outline 6 phases, starting with familiarisation of the data, followed by coding, theme search, theme review, theme definition and report production.

The method is best suited to large bodies of data, while the 52 responses in this thesis did not constitute a very large body. It was however enough that it

was difficult to grasp themes that are not completely obvious by simple perusal. Adding a layer of formality to the analysis also contributes a level of rigor to it.

Some adaptation was needed, however. Since in this case the analysis is only part of the thesis, and not a report in itself, the last phase is not applicable in itself, but the rest of the phases were followed as described by Clarke *et al.* [38].

5.2 Quantitative results

5.2.1 Experimental tests

The participants who had completed experimental tests but had not completed any module in the learning system were separated out into a cursory control group. No serious comparison between this control and the experimental group was possible, due to the low number of participants in the control, as well as the problematic sampling method; however, these data were tentatively explored nevertheless. Below are the results.

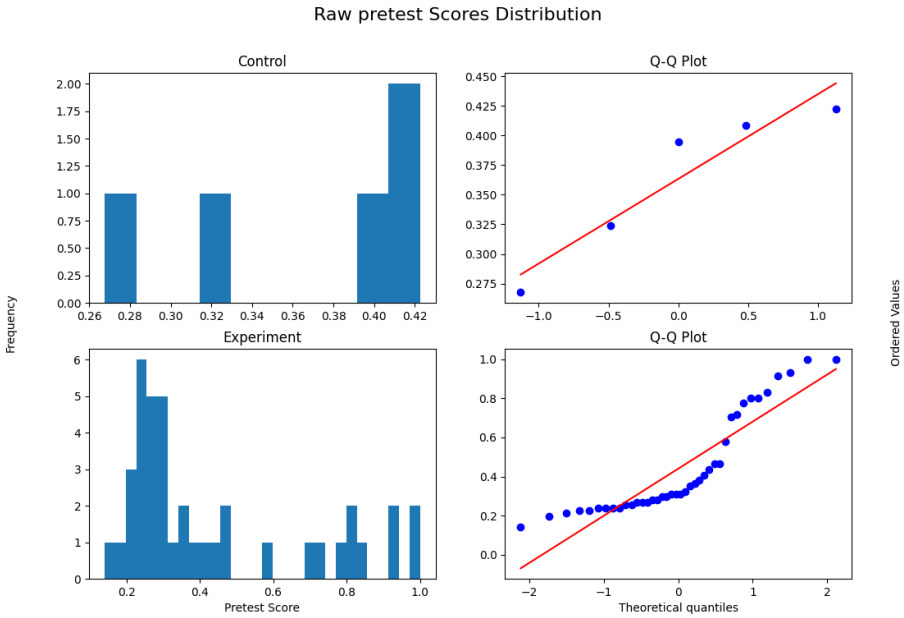
Unsurprisingly, after looking at the movement of the median score in Figure 5.2, the changes in the experimental group were statistically significant ($p = 0.00000003$). Although there is an observable shift in median in the control group, the paired sign test found no significant change in that group. I do however not exclude the possibility that an effect could exist, due to the low number in the control and the relatively low power of the sign test, in addition to a pattern of background increase emerges in the pre-post-retention comparisons further below. The distribution of individual score differences for the experimental group is shown in Figure 5.3, the clear majority scored better on the posttest vs the pretest, with a median percentage point difference of 22.5. The mean percent-wise individual score increase for these tests was 94.5%, while the median was 64.9%. Figure 5.4 shows the distribution for this metric.

Looking at the comparison of durations for the pre- and posttests in Figure 5.5, there seems to be a consolidation towards the median between the two tests, except for a couple of extreme outliers in the experimental group. There was no significant change between the pre- and posttest duration for the control group, but there was a significant *increase* in duration for the experimental group ($p = 0.003$). The individual differences for the experimental group in duration between pre- and posttest is plotted as a distribution in Figure 5.6. The mean duration being 206 seconds (3.5 minutes).

The score rate distributions for pre- vs posttest is shown in Figure 5.7. No significant change in this metric was detected in either direction for either group by the paired sign test.

On towards the comparison including retention. The comparison of test score results can be seen in Figure 5.8. There was still a significant increase in score for the experimental group between pre- and posttest after the filtration for the retention test ($p = 0.000004$). There was also a significant decrease in score between posttest and retention test for the same group ($p = 0.01$) The control, although

Figure 5.1: Q-Q plot of pretest data



now only two participants, is not omitted for sake of completeness. Figure 5.9 shows the distributions for the individual differences, and Figure 5.10 shows the distributions of the individual percent-wise increases for the two sets of tests. The median percentage point difference between pre- and posttest in this sample was 25.4. The mean percent-wise increase was 95.1%, with a median of 69.2%. The median percentage point difference between posttest and retention test was -5.63. The percent-wise increase between posttest and retention test had a mean of -9.27% and a median of -7.84%.

The latter shows a slight increase in score across the posttest and retention test that is mirrored in the histogram of the experimental group, where the floor of the distribution belonging to the experimental group is elevating throughout the tests, even though the median recedes a bit in the last test.

In terms of durations, there was still a significant increase in duration between pre- and posttest for the experimental group ($p = 0.005$). There was however no significant change in duration for the same group between posttest and retention test. The distributions and group-level medians for this comparison is shown in Figure 5.11. Despite no significant increase or decrease in a statistical sense, the distribution seems to converge on a median that stands more or less in place.

For score rates, shown in Figure 5.12, no significant change was found. Although there are signs of the same type of convergence as mentioned previously.

Figure 5.2: Score comparison, pre- vs posttest

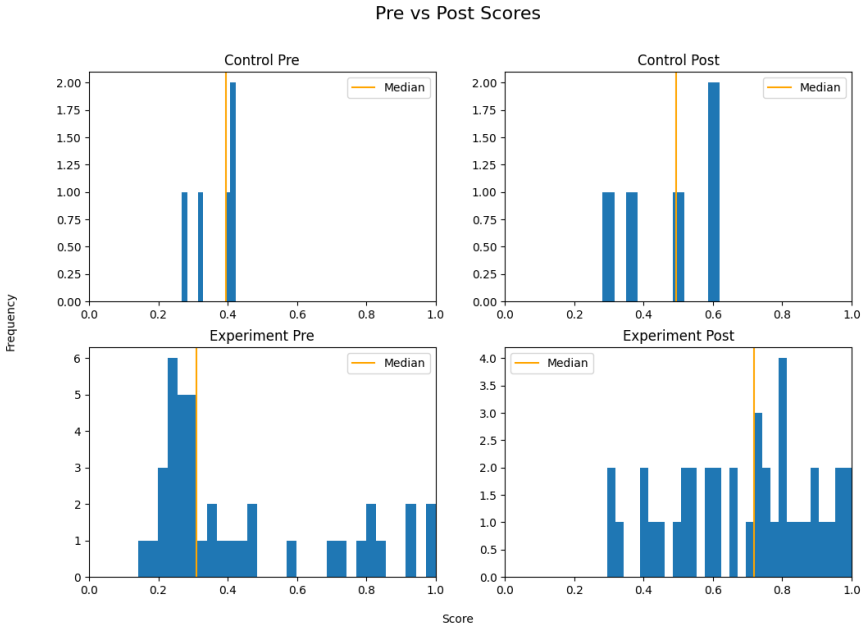


Figure 5.3: Individual score differences, pre- vs posttest

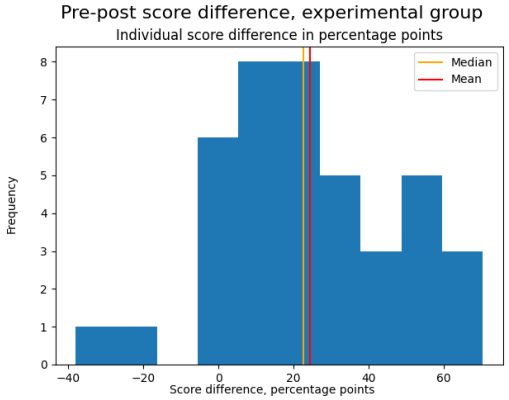


Figure 5.4: Individual score increase from pre- to posttest

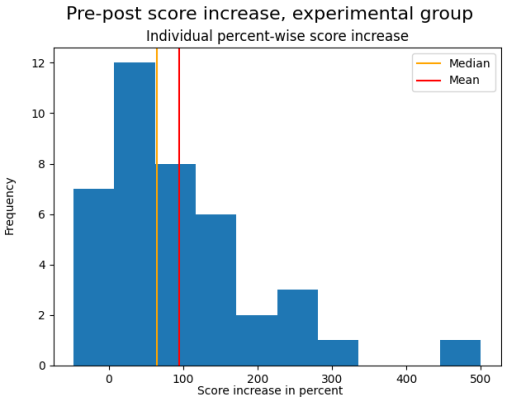


Figure 5.5: Duration comparison, pre- vs posttest

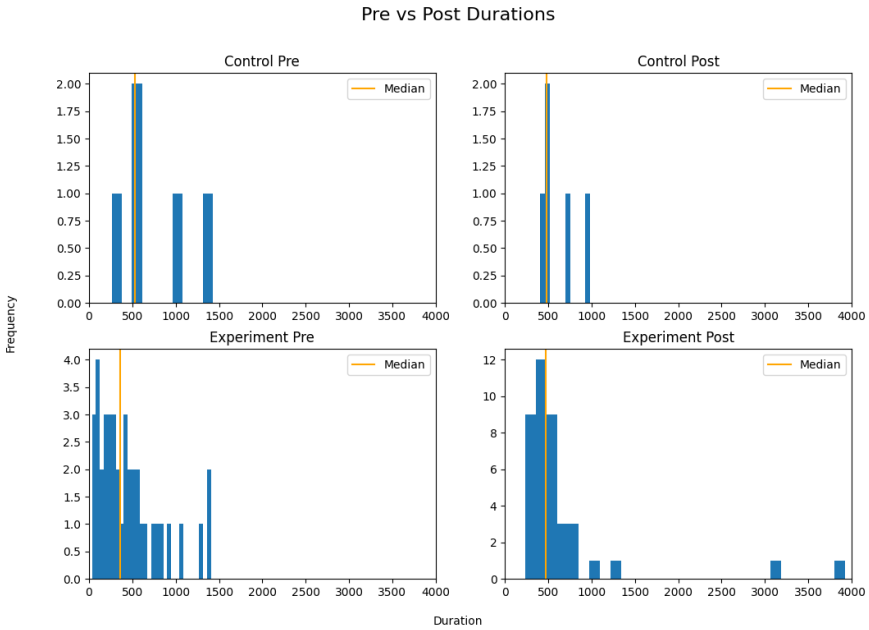


Figure 5.6: Individual duration differences, pre- vs posttest

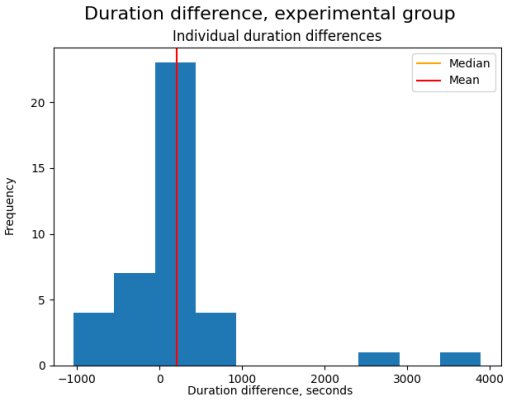


Figure 5.7: Score rate comparison, pre- vs posttest

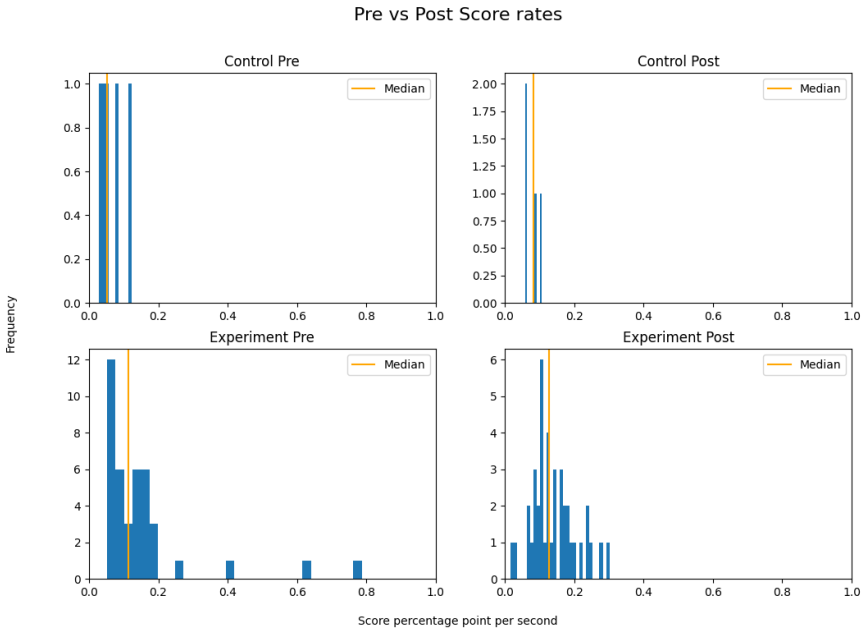


Figure 5.8: Score comparison, pre- vs posttest vs retention test

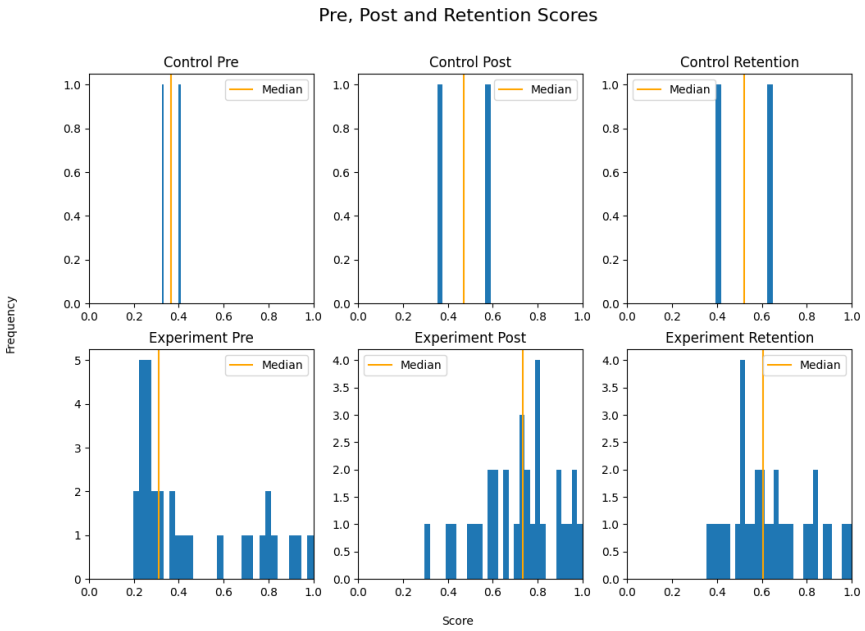


Figure 5.9: Individual score differences, pre- vs posttest vs retention test

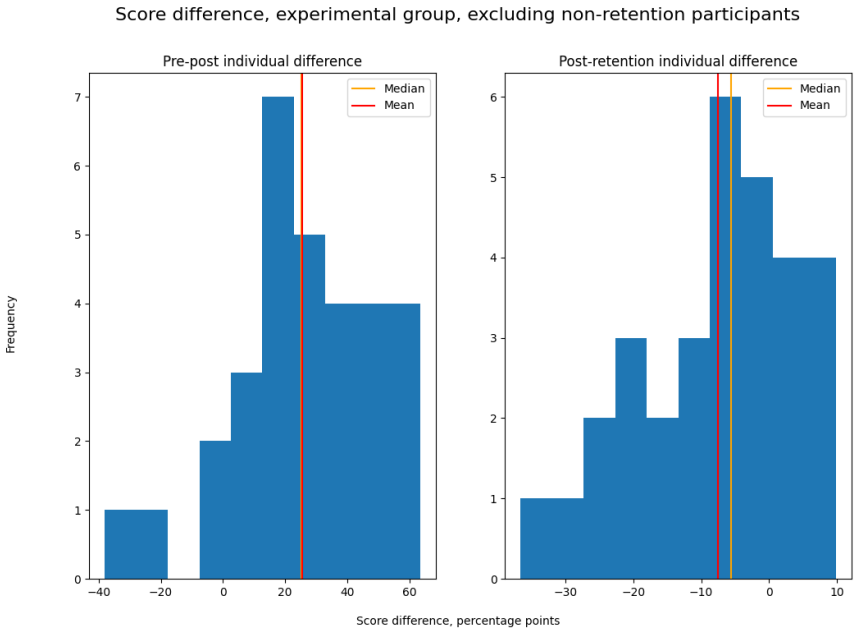


Figure 5.10: Individual score increases, pre- to posttest vs post- to retention test

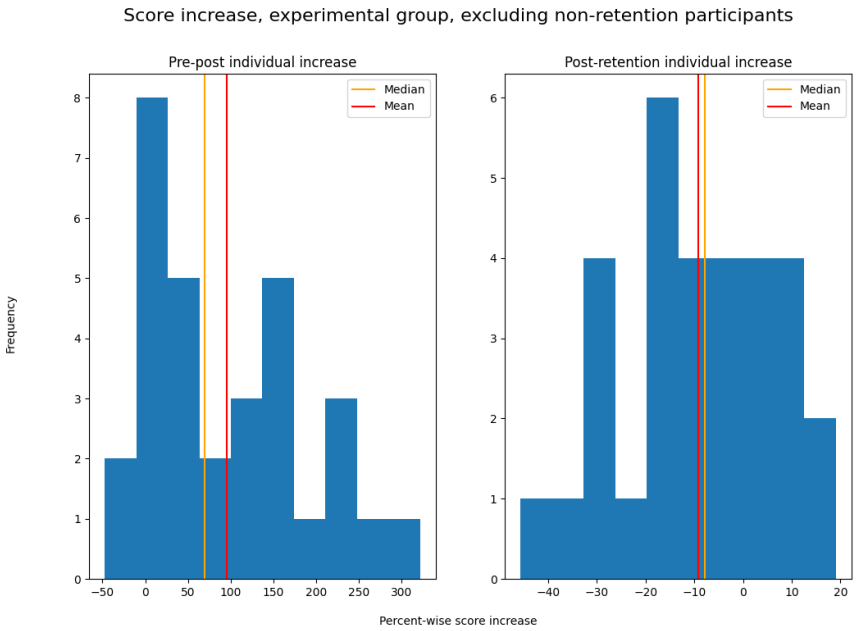


Figure 5.11: Duration comparison, pre- vs posttest vs retention test

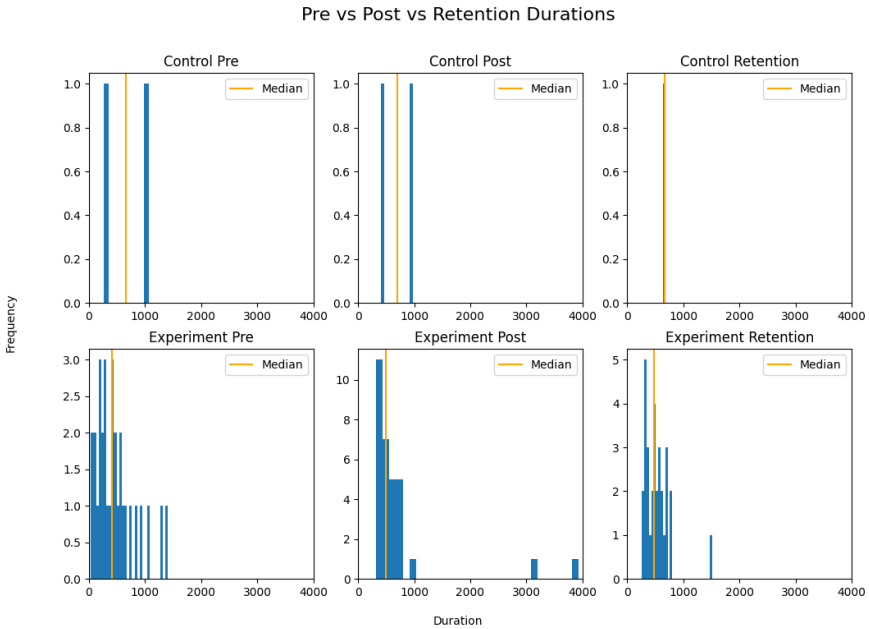
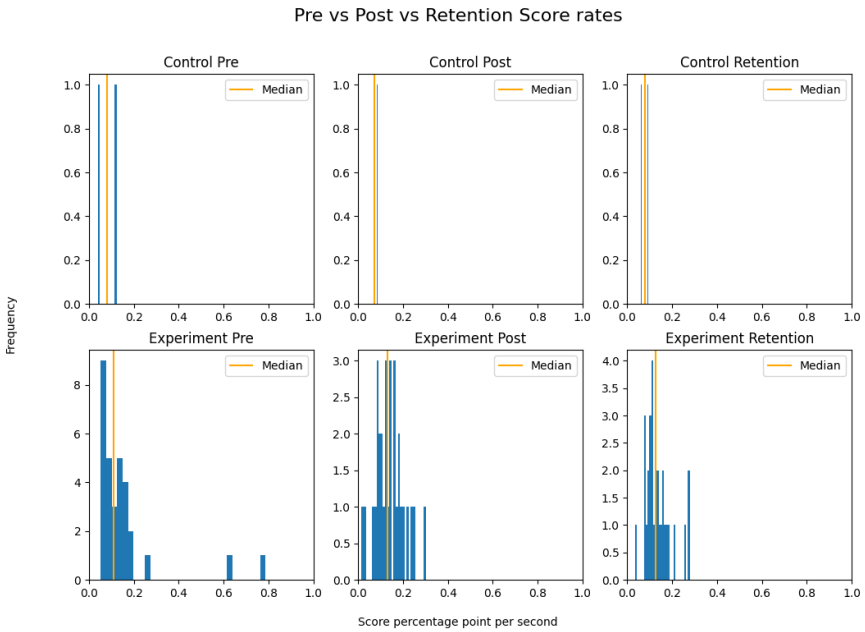


Figure 5.12: Score rates, between pre- vs post- vs retention test



5.2.2 Questionnaire

I need to preface the results from the questionnaire with the fact that only those participants who completed all the experimental tests (pre- posttest and retention test) were given the questionnaire. The amount of drop off, and low number of participants who didn't use the system represents views that this questionnaire cannot account for.

Below are the most important findings from the questionnaire.

Motivating factors

Beginning with the motivation factors, a majority (65.5%) of the respondents found the system engaging, meaning that they agreed or strongly agreed, as can be seen in Figure 5.13. The 95% confidence interval for this observation is [45.7%, 82.1%], showing that in a wider study it may have dipped below 50%. However, the confidence interval for the inverse case, that the majority did not find the system engaging, was [8.0%, 39.7%], with the indifferent response having [3.9%, 31.7%], indicating that in the worst case there would still be more people who were engaged than not, assuming a similar population.

When answering the more specific questions about the two phases of the system, 20.7%, a $\frac{1}{5}$, of the respondents report the neutral, indifferent, position. The confidence interval for this was [8.0%, 39.7%]. Comparing only agreeable and disagreeable responses, there was a majority of respondents who reported that the first phase was not boring (48.3%, [29.4%, 67.5%], against 31.0%, [15.3%, 50.8%]) but a majority also reported that the second phase of training was repetitive (62.1%, [42.3%, 79.3%], against 17.2%, [5.8%, 35.8%]). Looking at the confidence intervals of the former, one can see that there exists a significant possibility of those majorities being flipped if this study had more participants, meaning that the majority could be finding the system boring. I therefore regard that finding as inconclusive. The results of these two questions are plotted in Figures 5.14 and 5.15.

On the question of general enjoyment, there was still 20.7% neutral responses, but the rest were in a majority agreeing that the system was generally enjoyable (55.2%, [35.7%, 73.6%], against 24.1%, [10.3%, 43.5%]), shown in Figure 5.16. However, the confidence intervals also here has an overlap allowing for the majority being flipped in a larger study.

When asked if they would like to use a similar system for other subjects, a majority of respondents agreed (65.5%, [45.7%, 82.1%]), while nearly a quarter of them reported indifference (24.1%, [10.3%, 43.5%]). Only 3 (10.3%, [2.2%, 27.4%]) disagreed outright. The confidence intervals show that even in the worst case, more respondents would like to use a similar system for other subjects than not, as well as indifferent respondents, given a larger study. The plot is shown in Figure 5.17.

Usability factors

In terms of usability factors, two statements were posed to the respondents. The first was that the system was convenient to use, and as can be seen in Figure 5.18, responses were divided between disagreement and agreement, with the indifferent response at 13.8%, [3.9%, 31.7%], although there were a majority (55.22%, [35.7%, 73.6%], against 31.0%, [15.3%, 50.8%]) for agreement. The confidence intervals show that this majority could have been flipped in a larger study, and so these answers are not conclusive.

The second measured usability factor was that the respondent felt they made too many typos, shown in Figure 5.19, which had a clearer majority of agreement than the former at 69.0%, [49.2%, 84.7%]. This was against 13.8%, [3.9%, 31.7%], for disagreement; with 17.2%, [5.8%, 35.8%], for the neutral response.

External factors

Only a single external factor was conceived for inclusion in the questionnaire, that being the use of any memorization techniques outside of what the system provided. In this respondents largely disagreed that they had used such techniques (44.8%, [26.4%, 64.3%]) with 31.0%, [15.3%, 50.8%], responding indifference. The proportion of respondents who agreed that they had used memorization techniques casts doubt on this, however. These responses amounted to 24.1%, with the confidence interval [10.3%, 43.5%], meaning that yet again the majority could have been the opposite in a larger study. In addition the proportion of neutral responses to this question is also so high that in a larger study it may have been the dominating response, as can be seen from its confidence interval. Figure 5.20 gives a graphical view of these data.

5.2.3 Test diligence

Three prompts were posed to the respondents for measuring test diligence. The section was prefaced with a description of the context of describing how serious the respondent was in taking the different tests, with an admonition of being as honest as possible. The options were *Quite serious*, *Serious*, *Indifferent*, *Not serious* and *Skipped as fast as possible*.

The first prompt was the pretest, the results of which are shown in Figure 5.21, and resulted in a pretty symmetric distribution, with a slightly higher ratio of *skipped as fast as possible*, the most extremely non-serious response, than the mirrored extreme serious alternative. The majority, if measuring the individual options, reported indifference (27.6%, [12.7%, 47.2%]). However, when pooling the serious and non-serious alternatives one gets a slight majority for the non-serious side at 37.9%, [20.7%, 57.7%], against 34.5%, [17.9%, 54.3%], for the serious side. There is such big overlap in the confidence intervals of all three sides that the result is inconclusive in a wider context.

Figure 5.13: Engagement question results

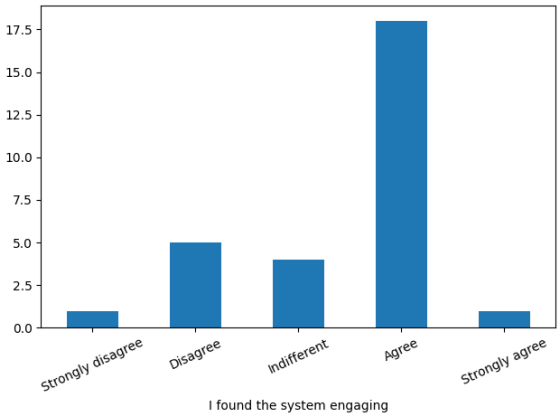
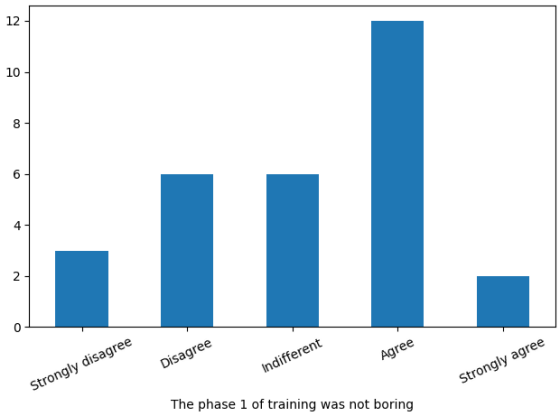


Figure 5.14: Phase 1 sentiment question results



The second prompt, the posttest, resulted in a clear majority for seriousness at 79.3%, with confidence interval [60.3%, 92.0%]. Against only 2 respondents reporting not being serious (6.9%, [0.8%, 22.8%]). The neutral option had 13.8% of the responses, with a confidence interval of [3.9%, 31.7%]. See Figure 5.22 for the plot.

The third prompt, for the retention test, still had a clear majority of seriousness, at 72.4%, [52.8%, 87.3%], and still only 2 respondents reporting not being serious. Indifference was reported a bit higher for this last prompt, at 20.7%, [8.0%, 39.7%]. The plot is shown in Figure 5.23.

Figure 5.15: Phase 2 sentiment question results

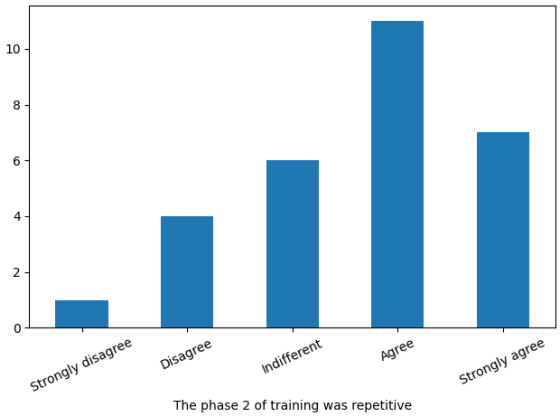


Figure 5.16: Overall enjoyment question results

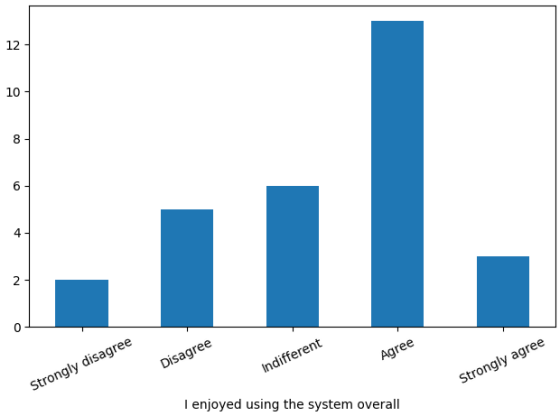


Figure 5.17: Desire of future use results

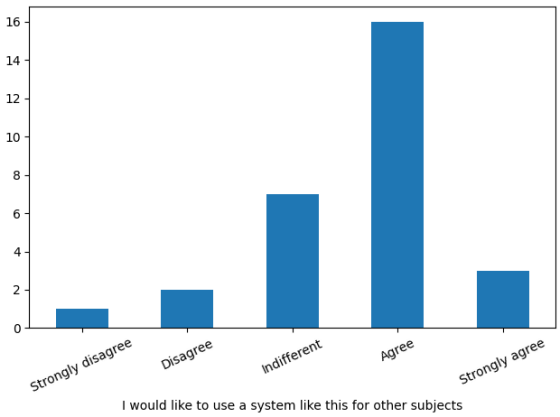


Figure 5.18: Usage convenience question results

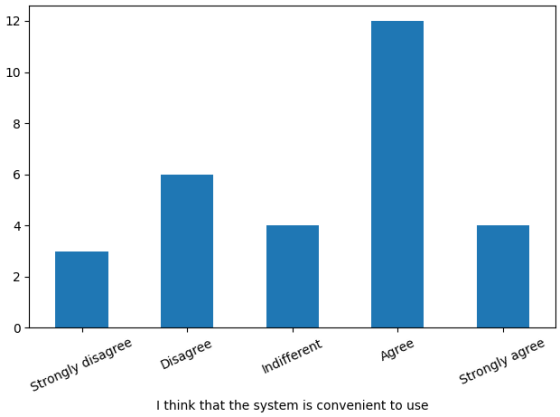


Figure 5.19: Too many typos question results

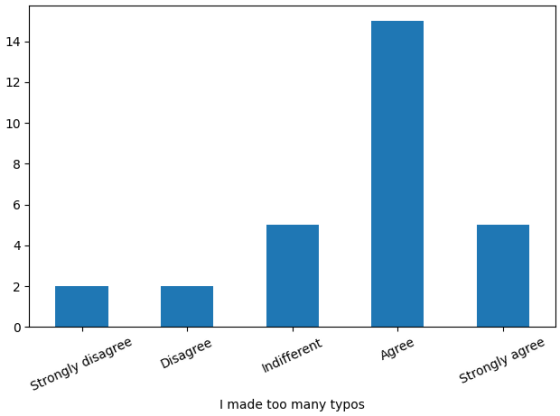


Figure 5.20: External factors question results

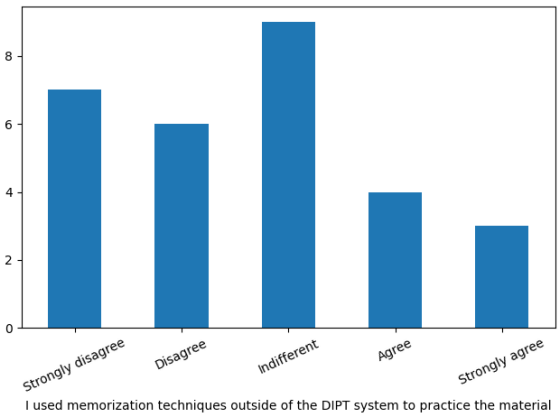


Figure 5.21: Pretest seriousness question results

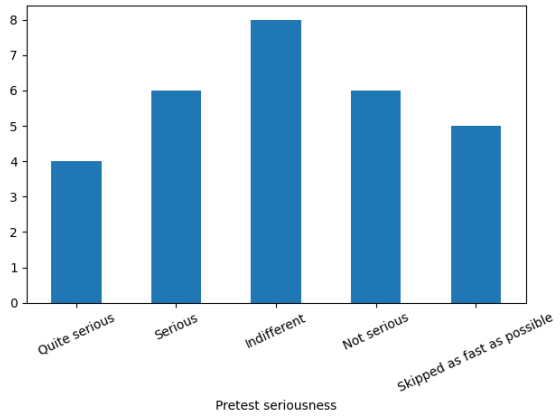


Figure 5.22: Posttest seriousness question results

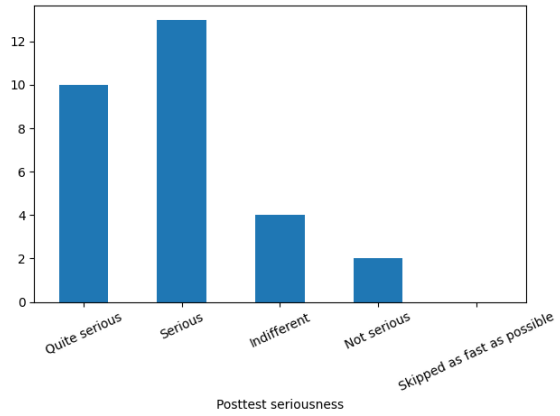


Figure 5.23: Retention test seriousness question results

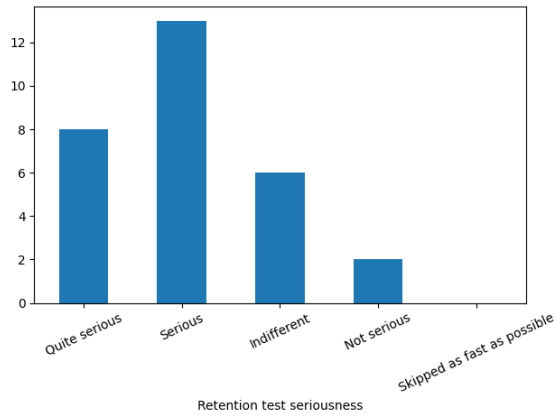
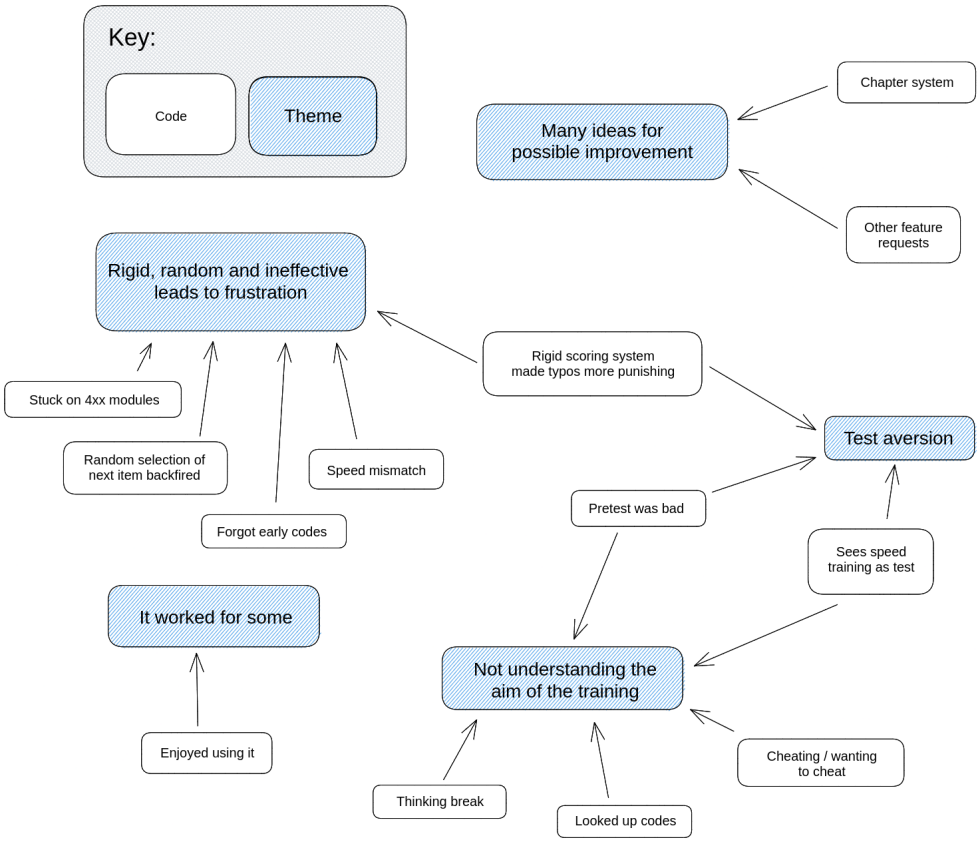


Figure 5.24: Thematic structure of free-text responses



5.3 Thematic analysis of free-text responses

The themes and associated codes will be presented here, and discussed further in Chapter 6.

5.3.1 Themes

In this section the identified themes are described. The codes that support them are detailed in Section 5.3.2. The overall thematic structure is also shown as a diagram in Figure 5.24.

Rigid, random and ineffective leads to frustration

The system seems to have had some flaws that have resulted in frustration with the respondents, besides the repetitiveness uncovered by the likert response in Figure 5.15. Which is indicated by the codes supporting this theme:

1. Rigid scoring system in PT phase made typos more punishing
2. Random selection of next item backfired

3. Speed mismatch
4. Forgot early codes
5. Stuck on 4xx modules

Test aversion

A problem that emerged as a possible pre-existing factor that I didn't account for when designing this experiment was the participants' aversion towards testing, supported by the following codes.

1. Pretest was bad
2. Sees speed training as a test
3. Rigid scoring system in PT phase made typos more punishing

Not understanding the aim of the training

The misunderstanding mentioned above is one symptom of a more general issue, supported by the following codes.

1. Thinking break
2. Pretest was bad
3. Sees speed training as a test
4. Looked up codes
5. Cheating/wanting to cheat

Many ideas for possible improvement

The responses were not just pointing out the pain points, however, many responses also detailed ideas for possible improvement of the system, supported by the following codes.

1. Chapter system:
2. Other feature requests

The main improvement, which was common to several respondents, was the introduction of a chapter system, or some other way to pick and choose to train on specific modules that one had unlocked. Other ideas ranged from making the progress bar more granular to procedural improvements like melding speed training of earlier completed material with current material upon completing a new module.

It worked for some

Finally, the system seems to have worked more or less as intended for some of the respondents, who as shown by the below code enjoyed it.

1. Enjoyed using it

5.3.2 Coding

The following were the identified codes in the material:

- Chapter system
- Cheating/wanting to cheat
- Enjoyed using it
- Forgot early codes
- Looked up codes
- Other feature requests
- Pretest was bad
- Random selection of next item backfired
- Rigid scoring system in PT phase made typos more punishing
- Sees speed training as a test
- Speed mismatch
- Stuck on 4xx modules
- Thinking breaks

Below, each code is described in more detail, with text extracts from the free-text responses in the questionnaire. As mentioned earlier, the respondents often crossed over into previous or future categories when giving their free-text responses, so for analysis I pooled all these responses. The entirety of the responses can bee seen verbatim, annotated with the part of the questionnaire they come from, in Appendix D.

Code: Chapter system

A recurring wish was that the respondents wanted a way to go back or forth between the different modules, to practice what they felt they needed to. One respondent even went as far as outlining a possible improvement in the form of a chapter system:

> From response 38:
>
> [...] getting stuck on a specific set of codes will hinder progress both in learning other codes, and will keep you from practicing old ones. I suggest creating a "chapter" system where you can choose any old ones you have done, but new ones are not unlocked before you have completet or done a certain number of (proper) tries.

I identified in total 10 such wishes within the responses, not always as well articulated as the above quote. A few samples of the rest are given below:

> From response 6:
>
> For further improvements of the system, I think it should be possible to train for specific status codes. For instance, if I feel unconfident on the 1xx status codes, it would be great if there was an option to train for these specifically.

> From response 7:
>
> [...] if there was a way for users to repeat sections at will I think the system and users would benefit a lot.
>
> From response 13:
>
> [...] Or even better, to revisit each module when you want.
>
> From response 21:
>
> I would like a way of going back to earlier codes too ensure i remember them.

Code: Cheating/wanting to cheat

The concept of cheating came up just a couple of times, but it fits well with the code *sees speed training as test* and a general trend of conflation of measuring and testing that I will revisit when describing the themes.

The first instance of this code came from a response that seems to talk about the speed training as 'tests', and found them too long. I don't think the respondent meant the experimental tests due to the mentioning of questions repeating — which was impossible for the experimental tests.

> From response 11:
>
> The tests were too long sould be more shorter test. For example a test for five and five questions, then a test for 10 (two of the earlier five questions) and 10 questions, then 20 and 20 questions and so on. At the end you can have a big test of questions you have repeted many times so you would know the answers. Sometimes you got the same question you just answered. It was very tempting to cheet and google the answers.

The second instance talks specifically about the pretest, and the respondent admits cheating on it, but also that they had misunderstood the purpose of it.

> From response 47:
>
> I fluked on the pre-test and cheated, I looked up quite a few codes I had misundersood that this was supposed to be a memory test [...]

It seems to me that both had a notion that they were being personally evaluated, instead of it being a measurement as part of training and an aggregate experimental pretest to *evaluate the tool*.

Code: Enjoyed using it

Some repondents expressed enjoyment in using the system. In total I identified six extracts that fit with this code, with a few reporoduced below:

> From response 3:
>
> Overall i really liked the system and I feel like this is a great way to learn factoids that would be pretty boring just to root learn by reading.

> From response 10:
>
> Having speedtests made it more engaging, where I felt rewarded after completing each module. The sound effects were also nice.

> From response 42:
>
> It was kind of addicting. Keyboard dopamine.

> From response 51:
>
> Overall it was a good system. Having multiple forms of input for the same concepts made it easier to memorize, like having to remember both the status code and later the description.

Code: Forgot early codes

It seems that the forgetting of material from early parts of the curricular material in the system was a recurring problem, with a total of five extracts that mention some variation of this. A couple of examples:

> From response 2:
>
> [...] getting stuck on 4xx codes for too long made me start to forget about the 1xx, 2xx and 3xx codes.

> From response 48:
>
> At the end of the training set when we had to do prompts from all segments I barely remembered the 1xx status codes.

Code: Looked up codes

Especially when asked about external factors, a common response was that the respondent had looked up codes outside of the system and read about them. Below are a couple of examples.

> From response 32:
>
> [...] i read through the 4xx status codes before the big test there to remember them all as training was looooooong, and i forgot the early ones halfway

> From response 37:
>
> [...] especially when it came to 4xx and 5xx status codes I read up on the codes trying to understand them and differentiate them better, as those modules were more challenging than the previous ones.

Code: Other feature requests

Besides the ideas for some type of chapter system above, the respondents also offered a diverse set of other suggestions for improvements or features they would like to see in the tool.

These are diverse, and drawing out specific examples do not make much sense. For the complete list, see Appendix C.

Code: Pretest was bad

I identified in total four passages where the respondents criticized, or otherwise shed light on some weakness of, the experimental pretest:

> From response 28:
>
> The beginning of the first part was to challenging in my opinion, because when you only know the status codes 200 and 404 it feels pointless to sit and guess new status codes.

This quote, along with the next, shows that at least these participants found the experimental pretest bothersome.

> From response 46:
>
> I take full responsibility for not doing the pre-test properly, however putting 71 status codes in front of me before we had even started to learn properly what each one meant was brutal and a horrible way to do things.

It seems that in these cases the pretest was taken with a considerable amount of negative emotions. It might be that this ties into my suspicion, mentioned earlier, that they saw the test as personal evaluation rather than a baseline.

The next two instances relates to slightly different misunderstandings of what the pretest was, but aren't abviously connected with negative emotions.

> From response 45:
>
> After the very first test I thought I would get to learn various status codes right away, but instead I was faced with lots of training on GET, PUT and POST requests.

> From response 47:
>
> I fluked on the pre-test and cheated, I looked up quite a few codes I had misundersood that this was supposed to be a memory test, so that is why the result is way worse at the post-test.

Code: Random selection of next item backfired

In total four responses mentioned that they repeatedly encountered the same prompt during the speed training (phase 2, PT), indicating that the random number generator without moderation may be a bit too repetitive in some cases.

> From response 11:
>
> Sometimes you got the same question you just answered.

In the case of a single repetition, it isn't too bad, but the following two passages indicate that the situation may have got far worse in some cases.

> From response 2:
>
> [...] a single code could get shown multiple times, then you barely get to learn it, and then it does not show up for another 10 tries because there are so many 4xx codes. And then you forget that one code you just learned.

> From response 46:
>
> Through luck you could get 3-4 of the same status code or the easy ones in a row(!!!!!) which made it largely unrepresentative in my mind.

I also read into the following that the random nature of the choice of next prompt may not have given even distribution of training across the material.

> From response 7:
>
> Some things were over repeated, other not repeated enough, [...]

Code: Rigid scoring system in PT phase made typos more punishing

The respondents consistently reported annoyances and grievances with making a lot of typos during the speed training. The making of typos connected with needless difficulty or negative emotions is captured in this code. A total of nine responses had passages fitting this. Below are a few example extracts:

> From reponse 1:
>
> [...] after a typo, it would be less fun to type more answers as you know you failed anyway.

> From response 15:
>
> Questions that required a very long answer to be written perfectly time after time were very frustrating.

> From response 18:
>
> The typo fails were really annoying. You could get a lot more correct than what was needed for the test, but still fail because you accidentally hit enter twice in a row once

> From response 43:
>
> I was focused on beating the high score in time, which was difficult in itself, but the fact that you could not misspell once or else the speed test was invalid made it way harder.

Code: Sees speed training as a test

A common notion within the free-text responses was to equate the speed training (PT) with testing. A total of 12 responses were identified to mention the speed training as tests, which this code encapsulates. Examples given below, with the terms bolded by me.

> From response 8:
>
> It was extremely demotivating during the **speed tests** when i was typing as fast as i could with no mistakes [...]

Especially striking is this example, where the respondent has felt the need to practice outside the system for the 'tests'. It defeats the purpose of the system, and indicates that something has gone wrong with the phase 2 training procedure.

> From response 34:
>
> I had to make a txt file with all the promtps for many of the **speed tests** to study them between tries [...]

Similarly, the following response points to the fact that one had to get everything right before moving on. This also ties into the typo code mentioned earlier.

> From response 43:
>
> Overall a good way to learn but it expected you to get everything right on a speed test before you could move on.

Following this thread, the next response notes that there is 'no opportunity' to refresh one's memory before taking the 'test'. Even though the correct response is presented when answering incorrectly on a prompt during the speed training.

> From response 13:
>
> If you stop between the parts referred to as phase 1 and phase 2 in the first question, and take it up again some days later, then you have no opportunity to refresh your memory using the system; **only the test is available.**

Code: Speed mismatch

The questionnaire resulted in a total of 6 responses indicating that something may have gone wrong with the threshold adaptation, resulting in a mismatch between the speed the system expected of the respondent and the speed they were capable of. A few examples are given below.

> From response 3:
>
> Sometimes the number of correct answers seemed high and really challenging, and other times very low.

> From response 4:
>
> The speed training was really difficult if you had fast typing speed because I felt like it expects you to get alot more rights answers in a shorter time and if you made 1 mistake you could not continue until you had everything right with no mistakes.

> From response 50:
>
> Should be able to lower difficulty. I needed 13 correct to continue, which is fine for 1xx status codes, but super hard for 4xx status codes.

Code: Stuck on 4xx modules

The respondent struggling or getting stuck with the specific module that trained HTTP status codes in the 4xx range was a occurred in 5 entries within the free-text responses. Below are a couple examples.

> From response 3:
>
> I talked to other people who said the section with 4xx codes was hard to pass.

> From response 46:
>
> Furthermore the 4xx status codes were a PAIN to go through, training should at least have been split into 4 parts, and the final test of that module was a pain in and of itself.

Code: Thinking breaks

Rather than the system not being able to adapt, as with the speed mismatch code, the following passages indicate that the respondent has not understood that the point of the training was to automate the responses, so that a several seconds long thinking break is not what you want. As such it indicates that I failed in communicating the aim of the system, and that this aim may have to be conveyed by the system itself.

> From response 8:
>
> It was extremely demotivating during the speed tests when i was typing as fast as i could with no mistakes, but i was thinking for less than 3s on some prompts and that caused me to not be quick enough.
>
> From response 34:
>
> I had to make a txt file with all the promtps for many of the speed tests to study them between tries when i simply wasn't fast enough to think and type at the same time
>
> From response 40:
>
> Some of the speed tests require too many words to complete. 15 words in 30 seconds when one has to think for a second what to answer can be a little bit too much

5.4 Time series

The participants seemed, from the thematic analysis, to have struggled especially with the 4xx series of status codes, and far more so than other ones. To corroborate these responses in the questionnaire, I plotted the time series of the speed training for each individual for inspection. Although reproducing all these here is too much, Figure 5.25 is a representative example.

In the figure, the dashed grey vertical lines signify the first speed training of a new partition of a module. And the usual case was, like in the figure, that the participants used at most two to three trials to get to the criterion of each partition, but once they hit the final stage of the training on the first 4xx module, something happens that make them struggle for a long time.

Touching on the misunderstandings mentioned in the analysis, I also plotted the aggregate time series for all participants in the entire experiment to see how well my admonition that they should train a little bit every day rather than going for hours at a time. This is shown in Figure 5.26. The vertical red streaks are low-opacity vertical lines drawn for each recorded training session, as the color darkens so does the density of training. There is clear darkening near the end of the experiment. The participants likely treated the training as any other course assignment. The extreme case, shown in Figure 5.27, is a participant who started and completed the entire curriculum of the system within the single day before the end of the experiment. Although that participant was en extreme example, the pattern of training for hours at a time, with days in between was common among the participants.

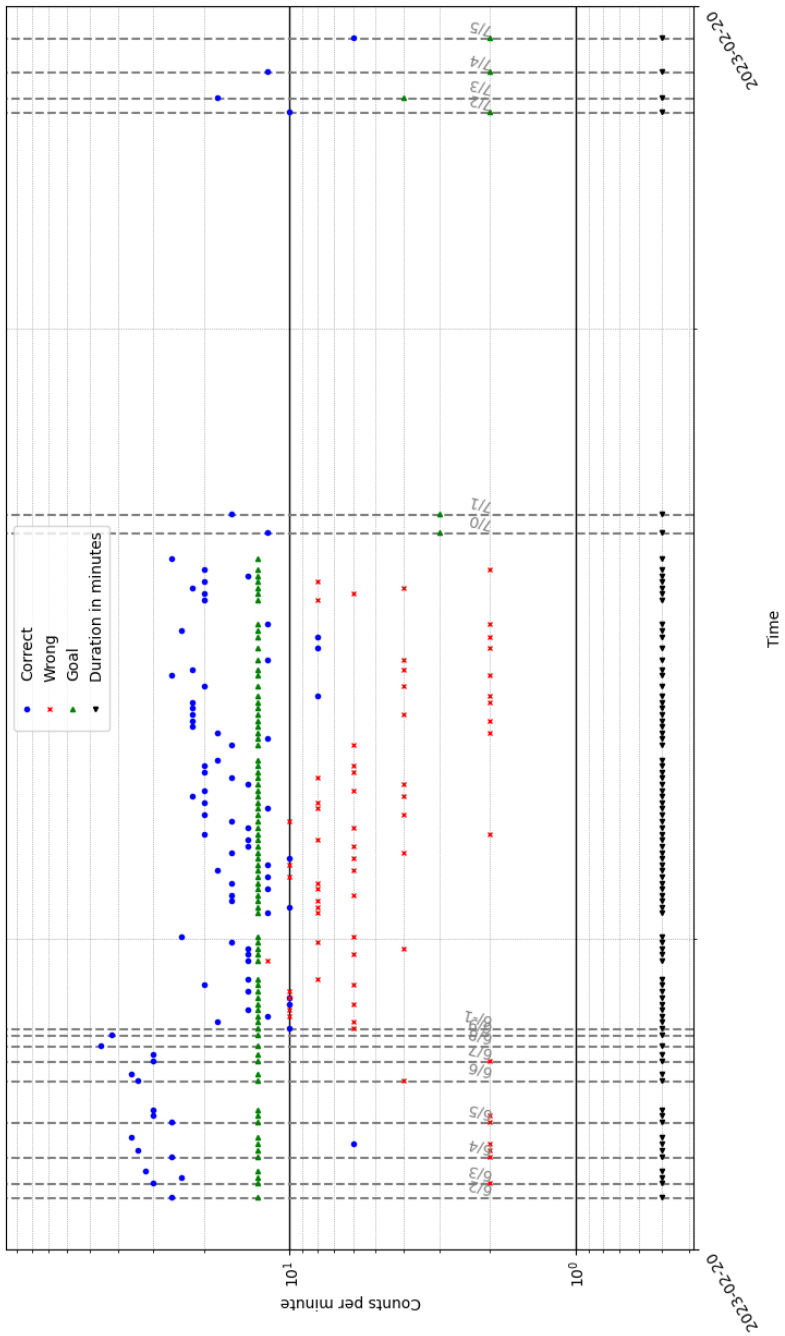


Figure 5.25: Example of typical time series around the start of 4xx status code training

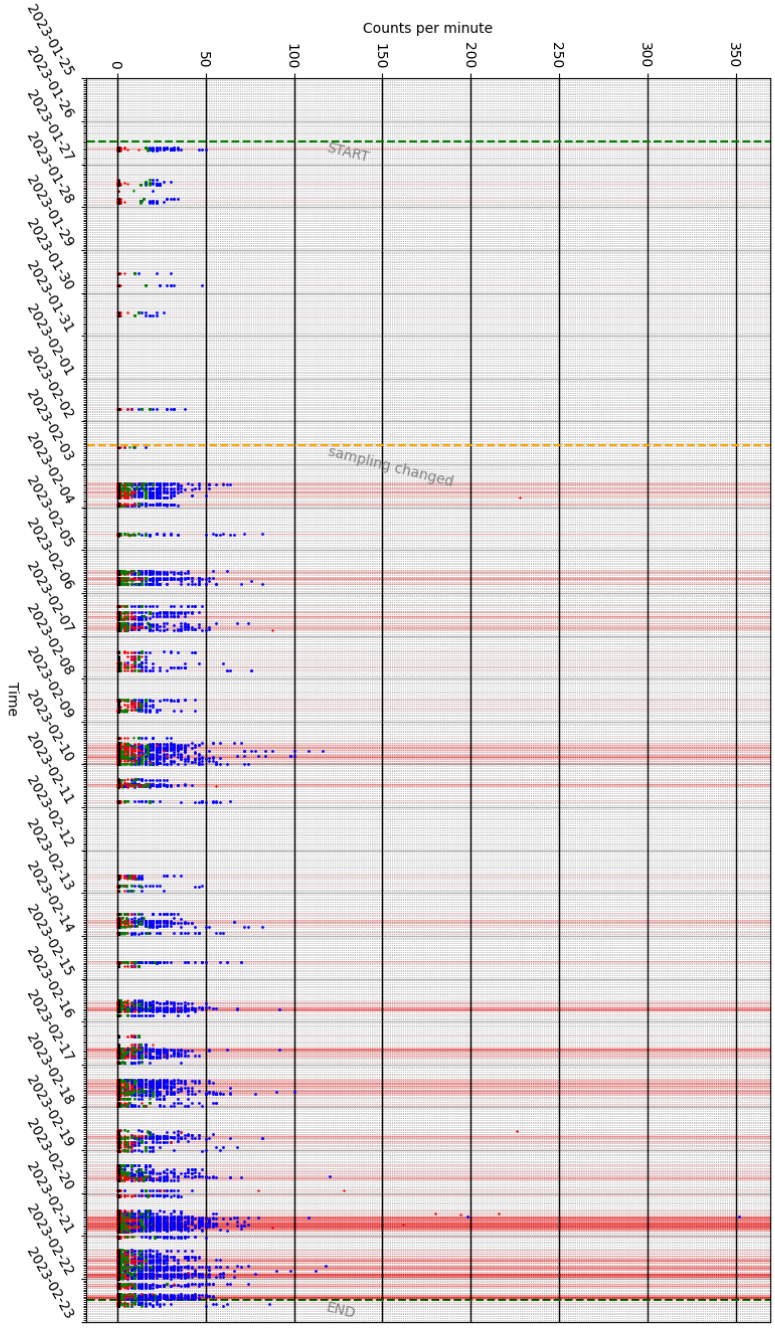


Figure 5.26: Aggregate time series of speed training of all participants

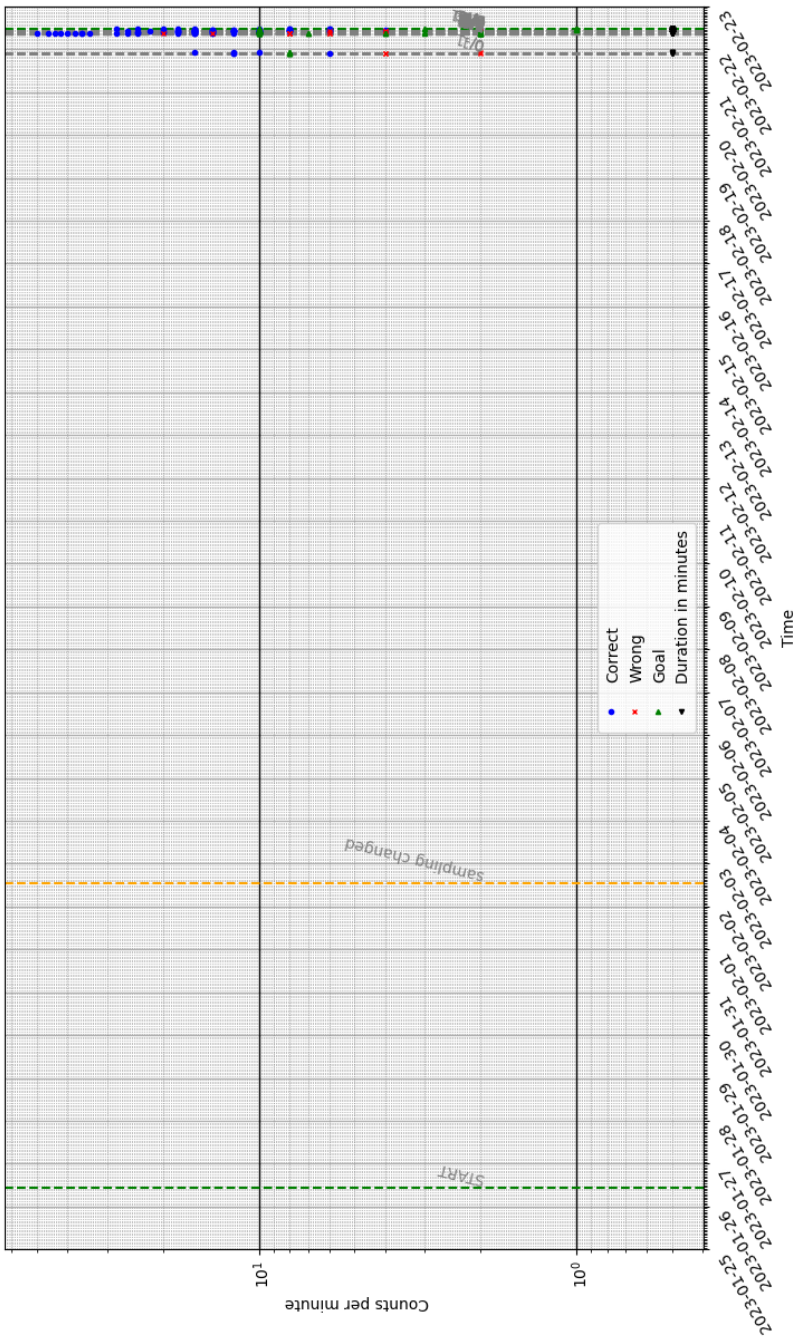


Figure 5.27: Time series for an individual who procrastinated

Chapter 6

Discussion

In this chapter, I attempt to answer the research questions, show other implications of the results and cover weaknesses of the system as it is built, limitations of the study and possible avenues of future related work.

6.1 Answering the research questions

In this section I will attempt to answer the research questions, which for clarity are repeated:

1. How effective is the application of DI and PT principles as foundation for educational gamification in teaching facts in a university computer science context?
2. How do university students react to the use of such a system?
3. How viable is such an application of these methods without tight project-integration of experts on these methods?

6.1.1 RQ 1: Efficacy

On the question of efficacy, it would be ideal to compensate for other variables than the tool itself. In the present study that would be to use the control group as such a compensation. This would factor out the learning the participants do as part of the course work itself, so that it is possible to measure if the tool improves the situation, and furthermore to what degree. This is not prudent in the present study, due to the sampling issues it has. The recruitment to the control group was too low to make a proper comparison between the groups. I will therefore not address the control group results in this section, but consider the results for the experimental group only.

The material trained in the system was more in-depth than what the participants would need for their course work, covering nearly all HTTP status codes in existence, and many participants scored above 70% on the posttest. For the retention test one can also see the score distribution fall back down a bit. This

is indicative that the system did indeed have an effect on accuracy. How big the effect was, can regrettably not be established with the current sampling.

Another measure of effectiveness could be to what degree the participants were able to answer fluently, meaning that they answer correctly and quickly. Or: how 'well' they know the material they have learned. To the best of my knowledge, none of the studies on SGs, gamification or ALSes have attempted this.

To look at this dimension, the experimental tests also measured the time that participants spent on them. I did not observe a speedup in terms of durations, which may have been presumed since the participants were doing timed trials in the training; rather, the participants seemed to be slower after training than before, at least at first glance: A statistically significant increase in test duration was observed between the pretests and the posttests. However, there was no statistically significant change in durations between the posttests and the retention tests.

The difference between the likert-scale answers for test diligence supports the claim that this difference in duration for the experimental tests may be due to the participants taking the pretest less seriously than the latter tests. Although the likert responses for the pretest are statistically inconclusive if seen in isolation, considering the contrast between them and the responses for the other diligence statements, as well as the increase in duration on the experimental tests, paint a clearer picture. I find it likely that this is the case, rather than a deficiency with the system itself.

There are however several technical deficiencies, which will be covered further in Section 6.3, that may have had an impact on the efficacy of the system in teaching the facts. Chief among them is the possible long chaining of the same prompt in PT-sessions due to the random selection of the next prompt, meaning that they got less opportunities to train the overall partition in the amount of time for each session — essentially over-training on singular items; and the tuning formula for the mastery threshold either resulting in too low a value or too high a value. Too low values in the early partitions may be some of the explanation for the wall that many participants hit when training the final part of the 4xx status codes, meaning that the system was not challenging enough in the separate partitions, such that they didn't learn the parts of the module well enough to be able to easily transition to training on all the parts at the same time. The sequencing of the curriculum itself may also be a cause of this, since the curriculum had two modules for these specific codes, covering many items. Splitting the large modules into smaller chunks would bring the learner to the summarizing training faster, which could possibly have alleviated some of this effect.

So, what is the answer to the research question? The system seems to have had an effect, although how large this effect is, is unknown due to the problems mentioned above. It is simply not feasible to calculate an effect size without a properly sized control group.

However, the percent-wise increases between the pretest and posttest observed were rather high. If we make a tentative comparison with the values for mean LGE

established in Table 2.1, we see that the mean increase in the experimental group in the present study is on par or higher than the mean among the reviewed papers. This is a promising sign that a future replication of this experiment with a proper control group might show this system well placed if ranked by efficacy against studies on SG, gamification and ALSes.

Furthermore, the median percentage point increase in scores was higher than 20, with the typical participant scoring in the neighborhood of 70% correct on the posttest. So the participants seem to have done well in absolute terms on the posttest, but also compared to the pretest where the typical participant scored just slightly better than chance on the multiple-choice test with the median around 30%.

How much of this is due to the prototype, and how much is due to the course work at large is impossible to say with the current sampling. There seems to be an effect, and it looks promising if compared to existing literature, but the research question can regrettably not be answered conclusively in terms of *how* effective the prototype was.

6.1.2 RQ 2: Student reactions

The reactions to the system were mixed. The majority of respondents to the questionnaire seem to have found the system engaging. The responses about the DI-phase of the training was, however, inconclusive, and the majority agreed to the statement that the PT-phase was repetitive. Furthermore, the results of the question about general enjoyment of the system was statistically inconclusive. In terms of usability, the results for the statement that the system was convenient to use were also inconclusive, but for the statement that they made too many typos, the majority agreed with it. The likert scale results concerning the participants' reactions toward the prototype are thus a bit unclear, but seem slightly negative when considered together.

The thematic analysis of the free-text responses offered some deeper insights, however. The fact that one had to be exact in typing the responses combined with the timed trials was a source of frustration and annoyance for the participants. The weakness of the threshold adaptation formula and sequencing mentioned in Section 6.1.1 seem to have contributed to a need, at least a perceived need, for going back in the curriculum to re-train earlier modules, giving rise to a common suggestion for improvement from the participants: some sort of chapter system where one can do exactly this. There was also several other suggestions for improvement as well. The fact that there were so many improvement suggestions indicate deficiencies in the prototype. Additionally, there was a noticeable aversion towards tests found in the analysis, which among other things manifested as the regarding of the speed training sessions as tests that one had to practice for, instead of using them as practice. But for some of the participants, the system seem to have worked as intended and was enjoyable, in the extreme case a participant described it as addictive keyboard dopamine.

The typo problem is consistent with the observations of Lovitz *et al.* [32], where misspellings was anecdotally expressed as a common annoyance among the participants. The social validity in that study was furthermore rather low for the TAFMEDS approach with the majority of the participants reporting it as their least favourite study activity. However, it should be mentioned that in that study, daily practice was mandated, even during weekends, while in the present study there was no such mandate.

Lovitz *et al.* [32] suggest allowing wild card characters in the configuration of the curriculum to make the check of the correctness of the answer less prone to typos. I would add to this that a more generic solution may be to implement some form of similarity measure between the input and correct answer, or alternately, to simply continue without stopping due to errors and focus on rate of correct responses. In the case of my system, the presence or absence of the bell sound may be feedback enough of the accuracy, while not stopping the learner as long as they achieve high enough rate of correct to beat their threshold.

As with the efficacy, there is still a lot of room for improvement. Chiefly lowering the need to be 100% accurate in the speed training and somehow combating the test aversion.

6.1.3 RQ 3: Viability without method expert integration

I do not think that I have been able to realise the full potential of DI and PT with this prototype, since the descriptions of their in-classroom use, characterised by enthusiasm [30], is not consistent with the degree of annoyance observed in the qualitative data. Many got stuck at particular modules of the curriculum, which should not happen either.

Also, I seem to have made errors in the sequencing of the curriculum, evidenced by the problems the participants had in the particular section of it concerning the 4xx codes. These errors may have been discovered if behavior analysts had been more involved in this project.

Although I presented an early prototype to behavior analysts for feedback, it was not possible to integrate them closely in the development process, and I believe the prototype suffered from it. Both DI and PT are iterative in nature [20], and how to change the approaches are not immediately evident to a layperson — at least not to me. Morningside Academy coaches their teachers [27], probably at least somewhat for this reason. Therefore, having guidance from method experts in not only the design and implementation of the software system in terms of educational procedure, but also in the shaping of the curriculum that the system teaches is necessary to evaluate such systems properly. Thus, I don't consider it viable to approach it like I have done in this study.

6.2 Other implications of the results

The movement of the floor of the test score distributions between the posttest and the retention test, seen in Figure 5.8, may suggest a background learning effect, which is plausible considering that the participants were concurrently working with the curriculum through course assignments, even though the general case scores lower on the latter tests. Though this is a cursory observation, and is by no means statistically significant, this seems to be the case for both control and experiment. As can be seen in Figure 5.9, the mean individual regression between posttest and retention test is under 10 percentage points, while there is a considerable area in the positive range up to 10 percentage points, supporting such a possible background learning effect.

These signs indicate that it may indeed be viable to repeat the experiment with randomly assigned groups, given a better recruitment strategy so as to get enough data for effect size analysis.

It seems that the criterion of no mistakes, combined with high speed, to continue through the curriculum may have made it so that making typos was more punishing than they are by themselves in a timed trial. It may also be that this typo problem is strengthening the notion that the respondents are being tested and need to practice for the tests, rather than the 'tests' being the practice itself.

Some participants seem to have misunderstood the pretest as an evaluation of them as a student, one respondent even admitting to cheating on it to get a higher score. This was separate from, and in addition to, the previously mentioned conflation of the speed training with testing. If the participants had such presumptions, unconscious or not, that they were being constantly evaluated, and furthermore repeatedly being told that they are *wrong* by a system that stops them in their tracks from simple misspelling, it may be perceived — conscious or not — as an excess punishment from the system that is supposed to help them.

The procedure of the speed training, or some factor in my communication of it, seems to make it regarded as a test that one needs to practice for, rather than an opportunity to practice in itself.

The participants also seem to have misunderstood the aims of the training, and the experiment, in more ways than one. They complained about not having time for thinking breaks of several seconds and still pass a segment, while the system is aimed at training automated responses for the prompts — answering correctly without having to think. They were also understanding the pretest as part of the 'product' or the training itself, not seeing that the experimental tests were there to evaluate the efficacy of the prototype. As previously stated they conflated the speed training with evaluation and seemed to feel the need to look up material outside the system to 'cheat' on the 'tests' instead of using the system itself to train until they were fast enough.

At the root of most of this must lie insufficient communication on my part, but at least some of it may be attributed to the participants not being used to measurement as part of learning, as this is not normal in the educational setting

they were used to. Testing however, *is* normal in that setting.

6.3 Deficiencies of the current prototype

In this instance, the learning material—the program—was not made by an expert, but myself. However, the natures of the methods themselves are data-driven, and the material and the sequencing of it can be iteratively improved by help of charting the data, which the present system does not do. Charting the speed training, and presenting it to the user after each such session should also be done so as to give feedback to them as they are learning [20].

Most of the new learning profiles that appeared after the sampling change had a TTC of 2 or 1. This is much more common than expected, compared to what should be normal if the criterion is set properly and the material is new to the subject. This may be due to that most of the students should have been familiar with the start of the learning material a week into the experiment, since they had had a couple of lectures covering it already at that point. Thus, it is likely that the learning aptitude was measured wrongly for the majority of the participants, consequently making training harder than necessary for them.

Additionally, I think that the tuning of the threshold formula has made the criterion for the short 10 second sessions too low, and should be altered in future attempts.

Too large partition sizes, paired with the dubious adaptation of the thresholds, may be part of the reason that participants struggled with the 4xx status codes. Another may be that this was effectively the first iteration of a DI-program, which should ideally be empirically validated before applied [28].

6.4 Limitations of the study

The biggest limitation of the study is that it is not possible to make a fair comparison between a control group and an experimental group. The initial sampling strategy failed to recruit large enough numbers who used the system, and the uncertainty tied with how many of the initial control group would indeed follow through with the experimental tests lead to the sampling being changed to convenience sampling, so as to ensure enough data for meaningful analysis at all. The initial assumption was that students would volunteer without external incentives. This assumption turned out to be naive. Students in the present cohort would not volunteer to any significant degree without explicit incentives. Course grade credits for participation were offered as incentives a week into the experiment, along with the sampling change, to maximise the data collected on the system. Still, these final incentives do not seem to have been good enough to achieve a decent control group.

Making this change a week into the experiment also ran the risk of invalidating the results related to the few early volunteers who trained. However, the amount

of training done during the first week was so low that I do not fear that the influence of this on the results have been large compared to the other limitations of the study.

A possible remedy for the low recruitment to the control group may have been to keep the random sampling between control and experimental groups, and offer course credits for participation only. It is however uncertain how much this would incentivise use of the system beyond just doing the experimental tests and questionnaire. The current strategy succeeded in getting a decent amount of usage on the part of the participants.

A properly controlled experiment would likely have been impossible without tangible rewards for participants of the same population. Anecdotally, the student body complained about having a high overall workload across courses this semester, which may be a contributing factor to the challenges with recruitment.

The present study also suffers from sampling bias due to the above mentioned changes. There are likely meaningful differences between the people who chose to not use the system, and those who chose to use it. For example users may have a greater degree of intrinsic motivation to work on the course than non-users, which would normally manifest as different placement within the course grade distribution between the groups. Neither grouping in this experiment can safely be considered representative of the 'typical' student in the cohort. Such differences are quite possible confounding factors if one was to compare the sparse control with the experimental condition. These factors cannot be addressed without better sampling.

Concerning the questionnaire, the confidence intervals turned out to be wide and several of the results were inconclusive due to it. The external validity of those results are therefore low. This situation would have been better with more participants.

Finally, the questionnaire was only given to the participants who had completed all of the experimental tests. A better approach would have been to give the participants the questionnaire directly after the posttest. The only part of it that would have been affected by doing so is the retention test diligence question. This was not done due to the questionnaire not being ready in time.

6.5 Future work

One aspect that should be changed for future iterations of the prototype, is to incorporate a learning aptitude test that uses guaranteed unknown material for learning, to accommodate learners with different background knowledge; or alternately, simply use the default partition size of 3. Furthermore, doing this with a multiple-choice design where the learner pairs abstract symbols with concrete symbols, like Braaten and Arntzen [39] and Arntzen *et al.* [40] have done, as opposed to the current baking-in of the measurement into the training, should further decouple the user's typing proficiency from the stratification of learners in terms of how much material to give them at a time.

Another aspect is the tuning of the necessary threshold formula. Few people have a typing speed high enough to be able to fulfill the rule of thumb of one correct response per second for longer words, or even for a phrase of two words, and for people to be able to complete the task, it needs to be adjusted to their proficiency in typing. A couple of questions arise: How much such adjustment is too much, leaving the learner with too low a level of fluency so that they do not learn the material well enough? How can this threshold be tuned so that it presents an equivalent challenge no matter the length of the typed phrase and duration of the session? Behavior analysis may have answers to the former, but the latter is a viable research question for computer science. Future work in this direction should in my opinion be an interdisciplinary effort, including researchers from both fields, so that this intersection can be tackled.

Concerning the problem of the typos, I think that as long as one is using the same learning channel of 'see-type' it may be worthwhile to remove the requirement of no errors, and focus on the rate of correct responses. Alternately, one could focus research efforts into making speech recognition accurate and fast enough to effectively use the 'see-say' channel instead. Another possible avenue, that keeps a focus on minimising errors as part of the mastery criterion, could be to implement a similarity measure threshold for if a response is correct or not, and thereby make the system less prone to misspelling.

Chapter 7

Conclusion

What is the use of learning something without learning it well? I am sure that the reader is familiar with the experience of learning interesting things only to forget them over the course of a summer. Assessing accuracy alone does not ensure this. Measuring the frequency of correct answers brings an element of fluency that is often lacking when measuring progress in learning. Being able to answer correctly and quickly necessitates practicing beyond accuracy, and it can be regarded as 'mastering' the topic in question.

In this study I attempted to add this dimension to a lightly gamified approach to practicing facts in a flashcard-like manner. Although inconclusive, the results seem promising compared to related work from the literature on serious games, gamification and adaptive learning systems. There are however several technical deficiencies with the current prototype, and also several limitations with the current study. But, it seems worthwhile to attempt another iteration with more participants and random sampling of a control and experimental group. The technical deficiencies should be addressed in such a new iteration.

However, it is not trivial to create good material and scheduling of such for utilizing the methods DI and PT without having strong guidance from experienced behavior analysts. Attempting to apply these methods in a digital system without tight integration of method experts is in my opinion not likely to bring more value than other, less evidenced, dynamic approaches like unguided gamification or conventional adaptive learning systems.

There seems to be a potential in this, but more work is needed to realise that potential or discard the notion of it. Moreover, such work should be interdisciplinary. That way, the work gains the procedural expertise from behavior analysis, as well as the professional-grade design, user interaction and algorithmic expertise that computer science can offer.

Bibliography

[1] S. Hallifax, A. Serna, J.-C. Marty and É. Lavoué, 'Adaptive gamification in education: A literature review of current trends and developments', in *European conference on technology enhanced learning*, Springer, 2019, pp. 294–307.

[2] I. Caponetto, J. Earp and M. Ott, 'Gamification and education: A literature review', in *European Conference on Games Based Learning*, Academic Conferences International Limited, vol. 1, 2014, p. 50.

[3] J. Swacha, 'State of research on gamification in education: A bibliometric survey', *Education Sciences*, vol. 11, no. 2, p. 69, 2021.

[4] S. Çiftci, 'Trends of serious games research from 2007 to 2017: A bibliometric analysis.', *Journal of Education and Training Studies*, vol. 6, no. 2, pp. 18–27, 2018.

[5] O. Irmade, N. Anisa *et al.*, 'Research trends of serious games: Bibliometric analysis', in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1842, 2021, pp. 12–36.

[6] R. Dörner, S. Göbel, W. Effelsberg and J. Wiemeyer, 'Introduction', in *Serious Games: Foundations, Concepts and Practice*, R. Dörner, S. Göbel, W. Effelsberg and J. Wiemeyer, Eds. Cham: Springer International Publishing, 2016, pp. 1–34, ISBN: 978-3-319-40612-1. DOI: 10.1007/978-3-319-40612-1_1. [Online]. Available: https://doi.org/10.1007/978-3-319-40612-1_1.

[7] S. Deterding, D. Dixon, R. Khaled and L. Nacke, 'From game design elements to gamefulness: Defining "gamification"', in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 2011, pp. 9–15.

[8] S.-K. Bjørnsen, 'Evaluating the learning gain improvements of serious games and gamification: A systematic literature review', 2022, [An unpublished paper delivered as part of course work in IMT4307, spring 2022. It is included in Appendix E].

[9] Y. Jing, L. Zhao, K. Zhu, H. Wang, C. Wang and Q. Xia, 'Research landscape of adaptive learning in education: A bibliometric study on research publications from 2000 to 2022', *Sustainability*, vol. 15, no. 4, p. 3115, 2023.

[10] D. Koutsantonis, K. Koutsantonis, N. P. Bakas, V. Plevris, A. Langousis and S. A. Chatzichristofis, 'Bibliometric literature review of adaptive learning systems', *Sustainability*, vol. 14, no. 19, 2022.

[11] F. Martin, Y. Chen, R. L. Moore and C. D. Westine, 'Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018', *Educational Technology Research and Development*, vol. 68, no. 4, pp. 1903–1929, 2020.

[12] S.-K. Bjørnsen, 'Exploring the empirical topography and quality of evidence for adaptive learning systems: A systematic review', 2022, [An unpublished paper delivered as part of course work in IMT4215, autumn 2022. It is included in Appendix F].

[13] H. Pashler, M. McDaniel, D. Rohrer and R. Bjork, 'Learning styles: Concepts and evidence', *Psychological science in the public interest*, vol. 9, no. 3, pp. 105–119, 2008.

[14] J. Cuevas, 'Is learning styles-based instruction effective? a comprehensive analysis of recent research on learning styles', *Theory and Research in Education*, vol. 13, no. 3, pp. 308–333, 2015.

[15] D. T. Willingham, E. M. Hughes and D. G. Dobolyi, 'The scientific status of learning styles theories', *Teaching of Psychology*, vol. 42, no. 3, pp. 266–271, 2015.

[16] K. Aslaksen and H. Lorås, 'The modality-specific learning style hypothesis: A mini-review', *Frontiers in psychology*, vol. 9, p. 1538, 2018.

[17] P. R. Husmann and V. D. O'Loughlin, 'Another nail in the coffin for learning styles? disparities among undergraduate anatomy students' study strategies, class performance, and reported vark learning styles', *Anatomical sciences education*, vol. 12, no. 1, pp. 6–19, 2019.

[18] A. A. Olsen, J. E. Romig, A. L. Green, C. Joswick and V. Nandakumar, 'Myth busted or zombie concept? a systematic review of articles referencing "learning styles" from 2009 to 2019', *Learning Styles, Classroom Instruction, and Student Achievement*, pp. 39–57, 2022.

[19] C. Riener and D. Willingham, 'The myth of learning styles', *Change: The magazine of higher learning*, vol. 42, no. 5, pp. 32–35, 2010.

[20] D. J. Moran and R. W. Malott, *Evidence-based educational methods*. Elsevier, 2004.

[21] E. A. Boyle, T. Hainey, T. M. Connolly, G. Gray, J. Earp, M. Ott, T. Lim, M. Ninaus, C. Ribeiro and J. Pereira, 'An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games', *Computers & Education*, vol. 94, pp. 178–192, 2016.

[22] W. D. Pierce and C. D. Cheney, 'Behavior analysis and learning', 2013.

[23] B. F. Skinner, *Science and Human Behavior*. Simon and Schuster, 1965.

[24] B. F. Skinner, 'Selection by consequences', *Behavioral and brain sciences*, vol. 7, no. 4, pp. 477–481, 1984.

[25] J. O. Cooper, T. E. Heron and W. L. Heward, *Applied behavior analysis*, 3rd ed. Pearson UK, 2020.

[26] D. J. Moran, 'The need for evidence-based educational methods', in *Evidence-based educational methods*, Elsevier, 2004, pp. 3–7.

[27] K. Johnson and E. M. Street, 'From the laboratory to the field and back again: Morningside academy's 32 years of improving students' academic performance.', *The Behavior Analyst Today*, vol. 13, no. 1, p. 20, 2012.

[28] T. A. Slocum, 'Direct instruction: The big ideas', in *Evidence-based educational methods*, Elsevier, 2004, pp. 81–94.

[29] C. Merbitz, D. ViEitez, N. H. Merbitz and H. S. Pennypacker, 'Precision teaching: Foundations and classroom applications', in *Evidence-based educational methods*, Elsevier, 2004, pp. 47–62.

[30] O. R. Lindsley, 'Precision teaching: Discoveries and effects', *Journal of Applied Behavior Analysis*, vol. 25, no. 1, p. 51, 1992.

[31] L. S. Vygotsky and M. Cole, *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.

[32] E. D. Lovitz, T. M. Cihon and J. Eshleman, 'Exploring the effects of daily, timed, and typed technical term definition practice on indicators of fluency', *Behavior analysis in practice*, 2020. DOI: 10.1007/s40617-020-00481-4.

[33] P. D. Leedy and J. E. Ormod, *Practical Research: Planning and Design*, 12th ed. Pearson Education Ltd, 2021, Global edition.

[34] G. G. Løvås, *Statistikk: for universiteter og høgskoler*, 4th ed. Universitetsforl., 2021.

[35] C. J. Clopper and E. S. Pearson, 'The use of confidence or fiducial limits illustrated in the case of the binomial', *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[36] M. Thulin, 'The cost of using exact confidence intervals for a binomial proportion', 2014.

[37] D. Mishra, S. Aydin, A. Mishra and S. Ostrovska, 'Knowledge management in requirement elicitation: Situational methods view', *Computer Standards & Interfaces*, vol. 56, pp. 49–61, 2018.

[38] V. Clarke, V. Braun and N. Hayfield, 'Thematic analysis', *Qualitative psychology: A practical guide to research methods*, vol. 3, pp. 222–248, 2015.

[39] L. F. Braaten and E. Arntzen, 'Effekten av antall noder og retningen i de trente betingede diskriminasjonene på etableringen av stimulusekvivalens', 2021.

[40] E. Arntzen, L. R. Halvorsen and C. Eilifsen, 'Respondering i henhold til stimulusekvivalens som en funksjon av antall trials vs. antall programmerte konsekvenser i etablering av baselinerelasjonene', 2021.

Appendix A

Curriculum template

The following is an export of the entire curriculum template, as it was filled out for the experiment in this thesis.

Topic title:	HTTP request methods									
Topic description:	Give the correct HTTP method that matches the description									
Topic motivation:	You will be presented with descriptions of HTTP methods. Your task is to input the correct method corresponding to the description given.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
A method that starts two-way communications with the requested resource. It can be used to open a tunnel.	connect	delete	get	head	options					
A method that deletes the requested resource on the host.	delete	patch	post	put	trace					
A method that requests a representation of a specified resource, and shouldn't include any data.	get	delete	trace	head	options					
A method that requests only the headers that would be returned from a URL if the request would have used the GET method instead.	head	patch	post	put	options					
A method that requests the permitted communication options for a given URL or server	options	delete	get	head	trace					
A method that applies partial modifications to a resource	patch	delete	post	put	trace					
A method that sends data to the server.	post	patch	get	head	options					
A method that creates a new resource, or replaces an old one.	put	patch	post	head	trace					
A method that performs a loop-back test along the path to the target resource.	trace	delete	get	put	options					

Topic title:	HTTP status code classes, part 1									
Topic description:	Give the first digit of the correct class of status code for the given description. (1xx/2xx etc.)									
Topic motivation:	HTTP requests are responded to with numerical codes, paired with a short textual descriptor, from the server, that hint to the status of that particular transaction. These are called status codes and status text. In this module you are presented with prompts for the first digit of a class of such codes. Your task is to enter the correct number.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
First digit if he response is just informational	1	4	2	3	5					
First digit if the request has been successful	2	1	4	3	5					
First digit if the request is redirected in some way	3	1	2	4	5					
First digit if an error has happened on part of the client	4	1	2	3	5					
First digit if an error has happened on part of the server	5	1	2	3	4					

Topic title:	HTTP status code classes, part 2									
Topic description:	Give the correct keywords corresponding to the prompted status code class									
Topic motivation:	This module is a reversal of the previous. You will be presented with a numerical status code class, and your task is to enter the correct keywords for it.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
1xx	information	success	redirect	client error	server error					
2xx	success	information	redirect	client error	server error					
3xx	redirect	success	information	client error	server error					
4xx	client error	success	redirect	information	server error					
5xx	server error	success	redirect	client error	information					

Topic title:	1xx HTTP status codes									
Topic description:	Give the correct HTTP status text or numerical code that corresponds to the prompt									
Topic motivation:	This module goes into detail on the first of the five classes of status codes. You will be presented with either a numerical response code or its formal status text given by the RFC document. Your task is to enter the correct code or status text in each case.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
100	continue	switching protocols	processing	early hints						
101	switching protocols	continue	processing	early hints						
102	processing	switching protocols	continue	early hints						
103	early hints	switching protocols	processing	continue						
continue	100	101	102	103						
switching protocols	101	100	102	103						
processing	102	100	101	103						
early hints	103	100	102	101						

Topic title:	2xx HTTP status codes									
Topic description:	Give the correct HTTP status text or numerical code that corresponds to the prompt									
Topic motivation:	This module goes into detail on the second of the five classes of status codes. You will be presented with either a numerical status code or its formal status text given by the RFC document. Your task is to enter the correct code or status text in each case.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
200	ok	multiple choices	continue	internal server error	bad request					
201	created	accepted	non-authoritative information	no content	reset content					
202	accepted	partial content	non-authoritative information	no content	reset content					
203	non-authoritative information	accepted	reset content	no content	reset content					
204	no content	accepted	non-authoritative information		reset content					
205	reset content	accepted	non-authoritative information	no content	no content					
206	partial content	accepted	non-authoritative information	no content	reset content					
ok	200	206	202	203	204					
created	201	206	202	203	204					
accepted	202	201	205	203	204					
non-authoritative information	203	201	202	206	204					
no content	204	201	202	203	205					
reset content	205	201	202	203	204					
partial content	206	201	202	203	204					

Topic title:	3xx HTTP status codes									
Topic description:	Give the correct HTTP status text or numerical code that corresponds to the prompt									
Topic motivation:	This module goes into detail on the third of the five classes of response codes. You will be presented with either a numerical status code or its formal status text given by the RFC document. Your task is to enter the correct code or status text in each case.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
300	multiple choices	found	see other	not modified	temporary redirect					
301	moved permanently	found	see other	not modified	temporary redirect					
302	found	permanent redirect	see other	not modified	temporary redirect					
303	see other	found	permanent redirect	not modified	temporary redirect					
304	not modified	found	see other	moved permanently	temporary redirect					
307	temporary redirect	found	see other	not modified	permanent redirect					
308	permanent redirect	found	see other	not modified	temporary redirect					
multiple choices	300	302	303	304	307					
moved permanently	301	302	303	304	307					
found	302	308	303	304	307					
see other	303	302	308	304	307					
not modified	304	302	303	301	307					
temporary redirect	307	302	303	304	301					
permanent redirect	308	302	303	304	307					

Topic title:	4xx HTTP status codes, part 1									
Topic description:	Give the correct numerical HTTP status code that corresponds to the prompt									
Topic motivation:	This module goes into detail on the fourth of the five classes of status codes. You will be presented with a formal status text given by the RFC document. Your task is to enter the correct numerical code.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
bad request	400	401	402	403	404					
unauthorized	401	400	402	403	404					
payment required	402	401	400	403	404					
forbidden	403	401	402	400	404					
not found	404	401	402	403	400					
method not allowed	405	406	407	408	409					
not acceptable	406	405	407	408	409					
proxy authentication required	407	406	405	408	409					
request timeout	408	406	407	405	409					
conflict	409	406	407	408	405					
gone	410	411	412	413	414					
length required	411	410	412	413	414					
precondition failed	412	411	410	413	414					
payload too large	413	411	412	410	414					
uri too long	414	411	412	413	410					
unsupported media type	415	416	417	418	421					
range not satisfiable	416	415	417	418	421					
expectation failed	417	416	415	418	421					
i'm a teapot	418	416	417	415	421					
misdirected request	421	416	417	418	415					
unprocessable entity	422	423	424	451	426					
locked	423	451	424	422	426					
failed dependency	424	423	451	422	426					
upgrade required	426	428	429	431	451					
precondition required	428	426	429	431	451					
too many requests	429	428	426	431	451					
request header fields too large	431	428	429	426	451					
unavailable for legal reasons	451	428	429	431	426					

Topic title:	4xx HTTP status codes, part 2									
Topic description:	Give the correct HTTP status text that corresponds to the prompt									
Topic motivation:	This module is a reversal of the previous. Your task now is to enter the correct status text corresponding to the numerical prompt.									
Training question	Training correct response (1-2 words max)	Multiple-choice incorrect 1	Multiple-choice incorrect 2	Multiple-choice incorrect 3	Multiple-choice incorrect 4					
400	bad request	unauthorized	payment required	forbidden	not found					
401	unauthorized	bad request	payment required	forbidden	not found					
402	payment required	unauthorized	bad request	forbidden	not found					
403	forbidden	unauthorized	payment required	bad request	not found					
404	not found	unauthorized	payment required	forbidden	bad request					
405	method not allowed	not acceptable	proxy authentication required	request timeout	conflict					
406	not acceptable	method not allowed	proxy authentication required	request timeout	conflict					
407	proxy authentication required	not acceptable	method not allowed	request timeout	conflict					
408	request timeout	not acceptable	proxy authentication required	method not allowed	conflict					
409	conflict	not acceptable	proxy authentication required	request timeout	method not allowed					
410	gone	length required	precondition failed	payload too large	uri too long					
411	length required	gone	precondition failed	payload too large	uri too long					
412	precondition failed	length required	gone	payload too large	uri too long					
413	payload too large	length required	precondition failed	gone	uri too long					
414	uri too long	length required	precondition failed	payload too large	gone					
415	unsupported media type	range not satisfiable	expectation failed	i'm a teapot	misdirected request					
416	range not satisfiable	unsupported media type	expectation failed	i'm a teapot	misdirected request					
417	expectation failed	range not satisfiable	unsupported media type	i'm a teapot	misdirected request					
418	i'm a teapot	range not satisfiable	expectation failed	unsupported media type	misdirected request					
421	misdirected request	range not satisfiable	expectation failed	i'm a teapot	unsupported media type					
422	unprocessable entity	locked	failed dependency	misdirected request	misdirected request					
423	locked	unavailable for legal reasons	failed dependency	unprocessable entity	misdirected request					
424	failed dependency	locked	unavailable for legal reasons	unprocessable entity	misdirected request					
426	upgrade required	precondition required	too many requests	request header fields too large	unavailable for legal reasons					
428	precondition required	upgrade required	too many requests	request header fields too large	unavailable for legal reasons					
429	too many requests	precondition required	upgrade required	request header fields too large	unavailable for legal reasons					
431	request header fields too large	precondition required	too many requests	upgrade required	unavailable for legal reasons					
451	unavailable for legal reasons	precondition required	too many requests	request header fields too large	upgrade required					

Topic title:	5xx HTTP status codes, part 1				
Topic description:	Give the correct HTTP status code that corresponds to the prompt				
Topic motivation:	This module goes into detail on the last of the five classes of response codes. You will be presented with a formal status text given by the RFC document. Your task is to enter the correct code.				
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**
internal server error	500				
not implemented	501				
bad gateway	502				
service unavailable	503				
gateway timeout	504				
http version not supported	505				
variant also negotiates	506				
insufficient storage	507				
loop detected	508				
not extended	510				
network authentication required	511				

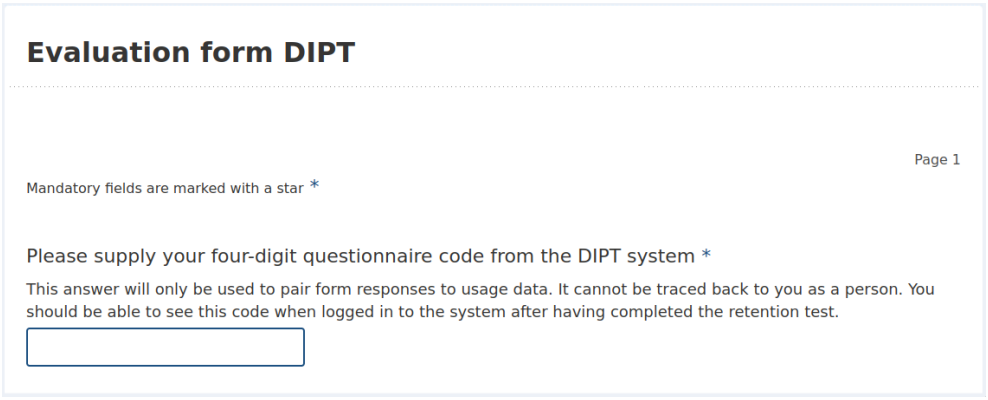
Topic title:	5xx HTTP status codes, part 2									
Topic description:	Give the correct HTTP status text that corresponds to the prompt									
Topic motivation:	This module is a reversal of the previous. Your task now is to enter the correct status text corresponding to the numerical prompt.									
Training question	**Training correct response (1-2 words max)**	**Multiple-choice incorrect 1**	**Multiple-choice incorrect 2**	**Multiple-choice incorrect 3**	**Multiple-choice incorrect 4**					
500	internal server error	not implemented	bad gateway	service unavailable	gateway timeout					
501	not implemented	internal server error	bad gateway	service unavailable	gateway timeout					
502	bad gateway	not implemented	internal server error	service unavailable	gateway timeout					
503	service unavailable	not implemented	bad gateway	internal server error	gateway timeout					
504	gateway timeout	not implemented	bad gateway	service unavailable	internal server error					
505	http version not supported	variant also negotiates	insufficient storage	loop detected	not extended					
506	variant also negotiates	http version not supported	insufficient storage	loop detected	not extended					
507	insufficient storage	variant also negotiates	http version not supported	loop detected	not extended					
508	loop detected	variant also negotiates	insufficient storage	network authentication required	not extended					
510	not extended	variant also negotiates	insufficient storage	loop detected	network authentication required					
511	network authentication required	variant also negotiates	insufficient storage	loop detected	not extended					

Appendix B

Questionnaire

The following are screenshots of the questionnaire, as the participants experienced it.

Figure B.1: Page 1 of questionnaire



Evaluation form DIPT

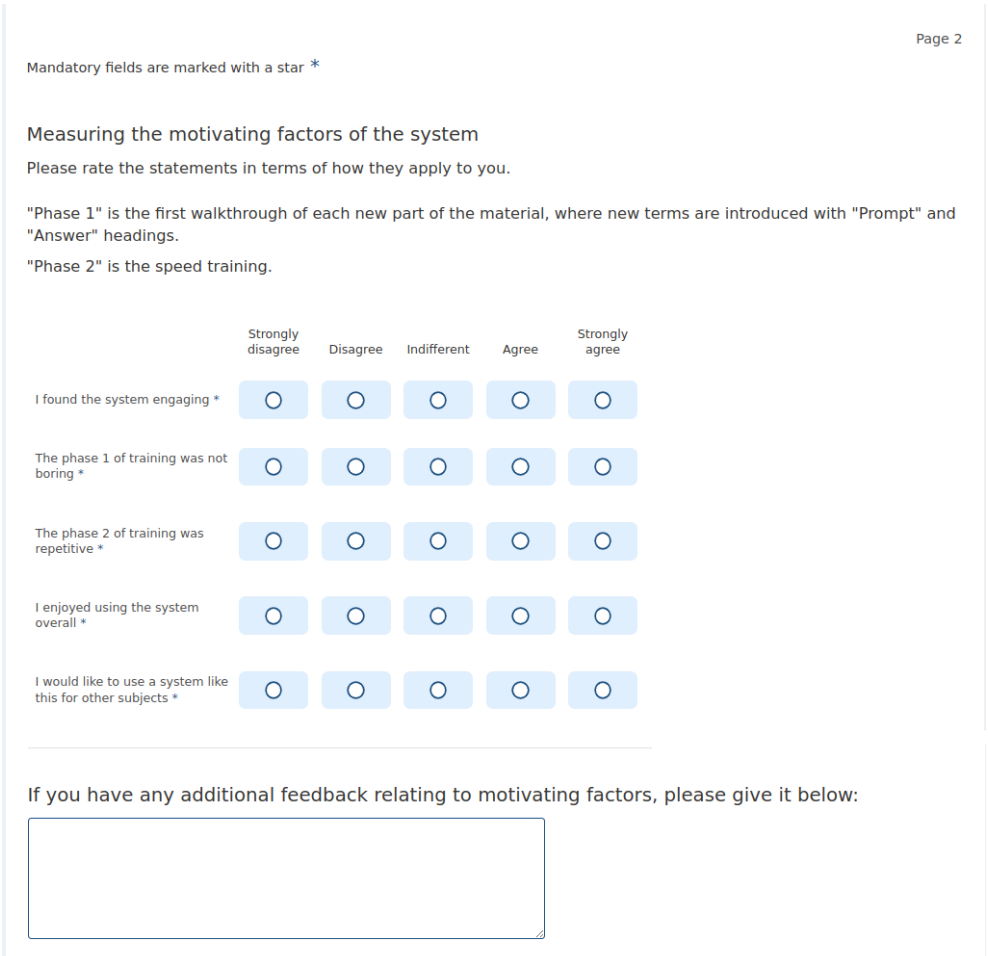
Page 1

Mandatory fields are marked with a star *

Please supply your four-digit questionnaire code from the DIPT system *

This answer will only be used to pair form responses to usage data. It cannot be traced back to you as a person. You should be able to see this code when logged in to the system after having completed the retention test.

Figure B.2: Page 2 of questionnaire



Page 2

Mandatory fields are marked with a star *

Measuring the motivating factors of the system

Please rate the statements in terms of how they apply to you.

"Phase 1" is the first walkthrough of each new part of the material, where new terms are introduced with "Prompt" and "Answer" headings.

"Phase 2" is the speed training.

	Strongly disagree	Disagree	Indifferent	Agree	Strongly agree
I found the system engaging *	○	○	○	○	○
The phase 1 of training was not boring *	○	○	○	○	○
The phase 2 of training was repetitive *	○	○	○	○	○
I enjoyed using the system overall *	○	○	○	○	○
I would like to use a system like this for other subjects *	○	○	○	○	○

If you have any additional feedback relating to motivating factors, please give it below:

Figure B.3: Page 3 of questionnaire

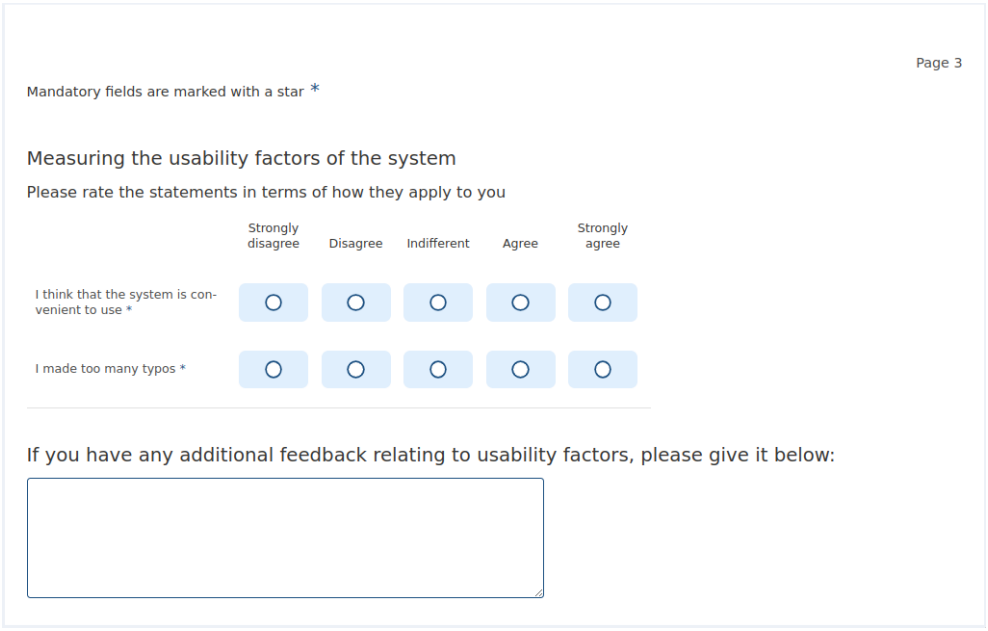


Figure B.4: Page 4 of questionnaire

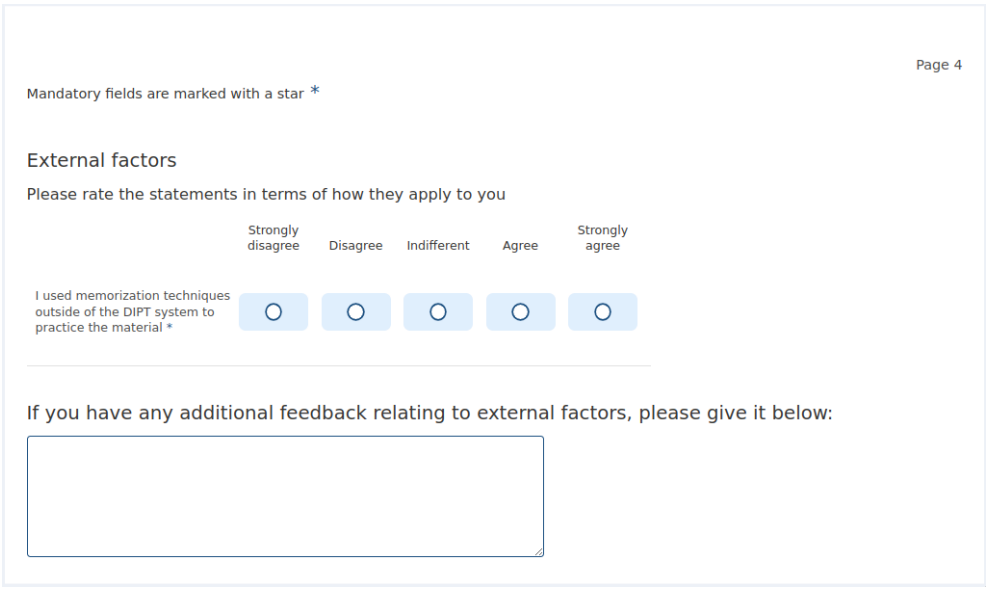


Figure B.5: Page 5 of questionnaire

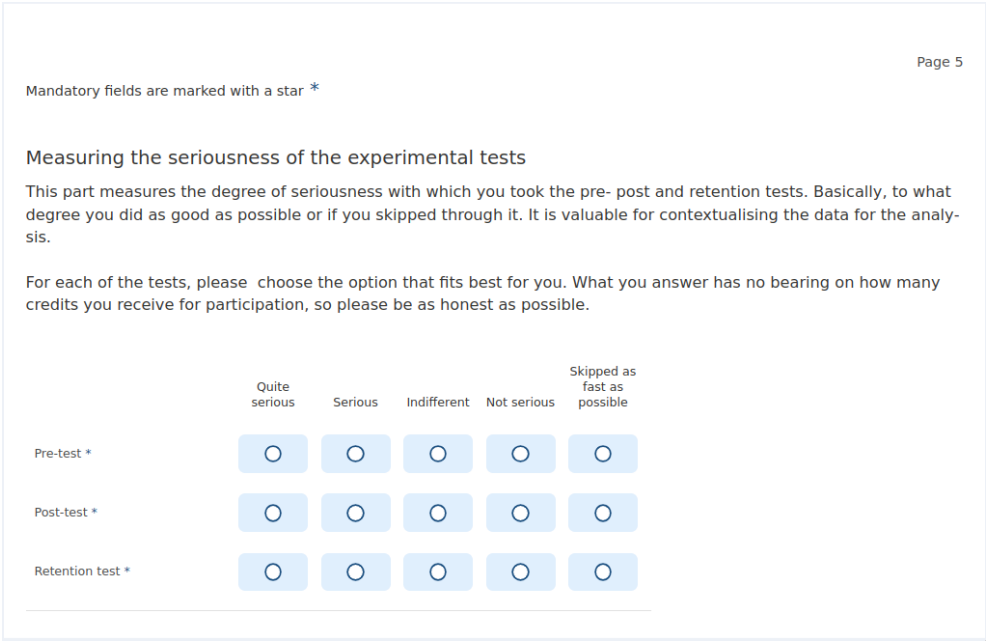
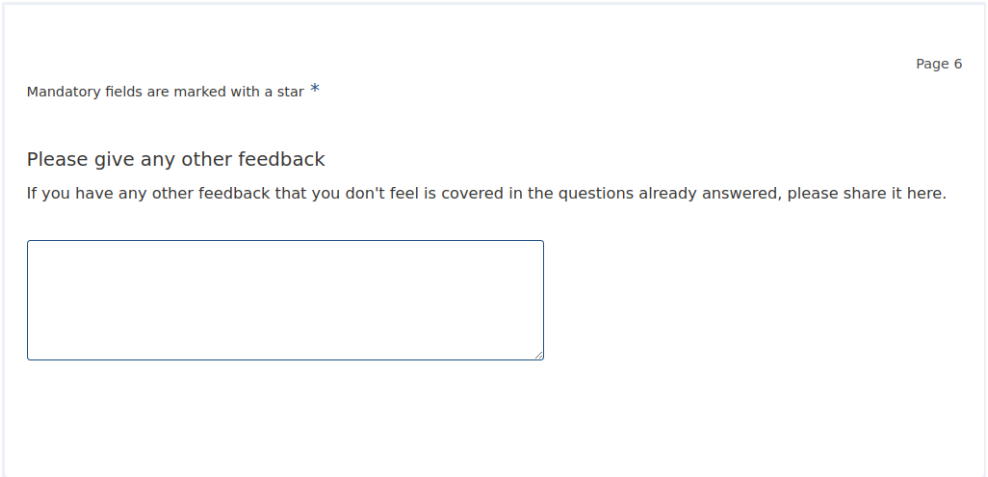


Figure B.6: Page 6 of questionnaire



Appendix C

Complete coding for the thematic analysis

The following are all identified excerpts related to the coding for the thematic analysis, organised into sections for each individual code.

C.1 Chapter system

- From response 6:
  ‘ For further improvements of the system, I think it should be possible to train for specific status codes. For instance, if I feel unconfident on the 1xx status codes, it would be great if there was an option to train for these specifically. ’
- From response 7:
  ‘ ...if there was a way for users to repeat sections at will I think the system and users would benefit a lot. ’
- From response 11:
  ‘ felt very forced. you couldn’t jump forward if you new the answers or jump back to repete hard answers. ’
- From response 13:
  ‘ ...Or even better, to revisit each module when you want. ’
- From response 21:
  ‘ I would like a way of going back to earlier codes too ensure i remember them. ’
- From response 38:

  > ...getting stuck on a specific set of codes will hinder progress both in learning other codes, and will keep you from practicing old ones. I suggest creating a “chapter” system where you can choose any old ones you have done, but new ones are not unlocked before you have completet or done a certain number of (proper) tries.

- From response 45:
  ‘ I wish it was possible to pick and choose categories to learn/test. (and maybe specific codes as well, like only 3xx codes, 1xx codes, etc.,) ’
- From response 46:
  ‘ The tool is not suited for the task, as it lacked a recap feature... ’
- From response 49:
  ‘ The site should have [...] ways of “retraing” on completed modules ’
- From response 52:
  ‘ Implement the training part to train on the specific module you’re working on, or make it possible for the use to choose module or choose which module to train on. ’

C.2 Cheating/wanting to cheat

- From response 11:

  > The tests were too long sould be more shorter test. For example a test for five and five questions, then a test for 10 (two of the earlier five questions) and 10 questions, then 20 and 20 ques-

> tions and so on. At the end you can have a big test of questions you have repeted many times so you would know the answers. Sometimes you got the same question you just answered. It was very tempting to cheet and google the answers.

- From response 47:
  ‘ I fluked on the pre-test and cheated, I looked up quite a few codes I had misundersood that this was supposed to be a memory test... ’

C.3 Enjoyed using it

- From response 2:
  ‘ I did enjoy using the tool to learn status codes. ’
- From response 3:
  ‘ Overall i really liked the system and I feel like this is a great way to learn factoids that would be pretty boring just to root learn by reading. ’
- From response 10:
  ‘ Having speedtests made it more engaging, where I felt rewarded after completing each module. The sound effects were also nice. ’
- From response 39:
  ‘ I wanted to train more but as far as I understood once you have finished training it just says “come back later” ’
- From response 42:
  ‘ It was kind of addicting. Keyboard dopamine. ’
- From response 51:
  ‘ Overall it was a good system. Having multiple forms of input for the same concepts made it easier to memorize, like having to remember both the status code and later the description. ’

C.4 Forgot early codes

- From response 2:
  ‘ ...getting stuck on 4xx codes for too long made me start to forget about the 1xx, 2xx and 3xx codes. ’
- From response 6:

  > For further improvements of the system, I think it should be possible to train for specific status codes. For instance, if I feel unconfident on the 1xx status codes, it would be great if there was an option to train for these specifically.

  Implies that the respondent, or their fellow student had this difficulty with the earlier codes
- From response 21:

‘ I spent too long on certain codes because of difficulties typing the entirety of the phrase quickly. I would like a way of going back to earlier codes too ensure i remember them. ’

- From response 32:
‘ ...i read through the 4xx status codes before the big test there to remember them all as training was looooooong, and i forgot the early ones halfway ’
- From response 48:
‘ At the end of the training set when we had to do prompts from all segments I barely remembered the 1xx status codes. ’

C.5 Looked up codes

- From response 32:
‘ ...i read through the 4xx status codes before the big test there to remember them all as training was looooooong, and i forgot the early ones halfway ’
- From response 35:
‘ I did a bit of reading on the codes outside of dipt, as the site offered no comprehensive list of the codes i had to learn. ’
- From response 36:
‘ ...I did Google to see the whole list of status codes. F.exp 5xx status codes ’
- From response 37:
‘ ...especially when it came to 4xx and 5xx status codes I read up on the codes trying to understand them and differentiate them better, as those modules were more challenging than the previous ones. ’
- From response 47:
‘ I fluked on the pre-test and cheated, I looked up quite a few codes I had misundersood that this was supposed to be a memory test... ’
This seems to only relate to the pretest.

C.6 Other feature requests

- From response 1:

  > ...Some type of scoring system might have made this more fun, for example that if you had an error, you could compensate by having more correct (e.g. one error could be compensated for by typing n correct answers in a row).
  >
  > I would also like to have an optional timer or progress bar to see the remaining time.

- From response 3:

  > If possible let students make their own sets. It would be a great addition to other courses too.

- From response 11:

> The tests were too long sould be more shorter test. For example a test for five and five questions, then a test for 10 (two of the earlier five questions) and 10 questions, then 20 and 20 questions and so on. At the end you can have a big test of questions you have repeted many times so you would know the answers.

- From response 13:
  ‘ I think there should have been a possibility to revise phase 1 as needed... ’
- From response 14:
  ‘ Would be nice to have an option to override incorrect answers. F.ex if you had a minor typo you could click a button to override the incorrect answer and make it correct. ’
- From response 17:

  > when the user is warming up it should take the data from the warm up (the typing speed) and update the expected typing speed of the user. Basically update the expected typing speed of the user, so that if the user has fast typing speed when doing the typing speed they wont be expected to answer questions fast on every speed test.

- From response 28 (in relation to pretest):
  ‘ Maybe it could be an idea to have hints in the beginning phase of such a learning tool. F.ex have html-tag abbriviations on status codes, so that when you hover the mouse over an status code you can read what is it. ’
- From response 48:

  > What I think could improve this it to add a test for after you have practised a segment. For example: I have just finished the testing for all 3xx status codes. Now i have a test that tests me on everything i have learned so far. Just like the last test, but with only what i had learned so far. And one of those after each of the big status code segments

- From response 49:
  ‘ The site should have a more detailed progress bar... ’
- From response 50:
  ‘ Should be able to lower difficulty. I needed 13 correct to continue, which is fine for 1xx status codes, but super hard for 4xx status codes. ’

C.7 Pretest was bad

- From response 28:
  ‘ The beginning of the first part was to challenging in my opinion, because when you only know the status codes 200 and 404 it feels pointless to sit and guess new status codes. ’
- From response 45:

> After the very first test I thought I would get to learn various status codes right away, but instead I was faced with lots of training on GET, PUT and POST requests.

- From response 46:
  ‘ I take full responsibility for not doing the pre-test properly, however putting 71 status codes in front of me before we had even started to learn properly what each one meant was brutal and a horrible way to do things. ’
- From response 47:
  ‘ I fluked on the pre-test and cheated, I looked up quite a few codes I had misundersood that this was supposed to be a memory test, so that is why the result is way worse at the post-test. ’

C.8 Random selection of next item backfired

- From response 2:
  ‘ ...a single code could get shown multiple times, then you barely get to learn it, and then it does not show up for another 10 tries because there are so many 4xx codes. And then you forget that one code you just learned. ’
- From response 7:
  ‘ Some things were over repeated, other not repeated enough, if there was a way for users to repeat sections at will I think the system and users would benefit a lot ’
- From response 11:
  ‘ Sometimes you got the same question you just answered. ’
- From response 46:
  ‘ Through luck you could get 3-4 of the same status code or the easy ones in a row(!!!!!) which made it largely unrepresentative in my mind. ’

C.9 Rigid scoring system in PT phase made typos more punishing

- From reponse 1:

  > I think the scoring system for Phase 2 was too rigid. One typo or error means that you have to take the test again.
  >
  > ...this may discourage trying to get as many as possible correct, because the more answers you type, the bigger the risk of a typo.
  >
  > ...after a typo, it would be less fun to type more answers as you know you failed anyway.

- From response 4:
  ‘ ...I felt like it expects you to get alot more rights answers in a shorter time and if you made 1 mistake you could not continue until you had everything right with no mistakes. ’

- From response 15:
  ‘ Questions that required a very long answer to be written perfectly time after time were very frustrating. ’
- From response 18:
  ‘ The typo fails were really annoying. You could get a lot more correct than what was needed for the test, but still fail because you accidentally hit enter twice in a row once ’
- From response 21:
  ‘ The system was to stringent on the typos. ’
- From response 22:
  ‘ A lot of typos made the experience longer than needed. ’
- From response 43:
  ‘ I was focused on beating the high score in time, which was difficult in itself, but the fact that you could not misspell once or else the speed test was invalid made it way harder. ’

C.10 Sees speed training as a test

- From response 1:
  ‘ One typo or error means that you have to take the test again. ’
- From response 3:
  ‘ I noticed that both speed and accuracy was greatly dependent on time of day / tiredness and if there were any distractions while doing the tests. ’
- From response 8:
  ‘ It was extremely demotivating during the speed tests when i was typing as fast as i could with no mistakes... ’
- From response 10:
  ‘ I think the theme of the tests, HTTP status codes, exacerbated the repetitiveness. ’
- From response 11:

  > The tests were too long sould be more shorter test. For example a test for five and five questions, then a test for 10 (two of the earlier five questions) and 10 questions, then 20 and 20 questions and so on. At the end you can have a big test of questions you have repeted many times so you would know the answers. Sometimes you got the same question you just answered.

  This one was a bit difficult to unpack. Did they mean the pretest? No, after mulling it over, the referral to the randomness issue in the last sentence must refer to the speed trials, since this would not occur in the experimental tests.
- From response 12:
  ‘ The long test phase was really difficult and repetitive. ’
  Like with response 11, the referal to repetitive indicates the speed training more than the experimental tests, especially considering how much conflat-

ing occurs.

- From response 13:
  ‘ If you stop between the parts referred to as phase 1 and phase 2 in the first question, and take it up again some days later, then you have no opportunity to refresh your memory using the system; only the test is available. ’
- From response 17:
  ‘ ...when doing the typing speed they wont be expected to answer questions fast on every speed test. ’
- From response 32:
  ‘ ...i read through the 4xx status codes before the big test there to remember them all as training was looooooong, and i forgot the early ones halfway ’
- From response 34:
  ‘ I had to make a txt file with all the promtps for many of the speed tests to study them between tries... ’
- From response 40:
  ‘ Some of the speed tests require too many words to complete. ’
- From response 43:
  ‘ Overall a good way to learn but it expected you to get everything right on a speed test before you could move on. ’
- From response 45:
  ‘ I wish it was possible to pick and choose categories to learn/test. ’
- From response 46:

  > Furthermore the 4xx status codes were a PAIN to go through, training should at least have been split into 4 parts, and the final test of that module was a pain in and of itself.
  >
  > Final test was also a nightmare, as you relied on luck there too.

- From response 48:
  ‘ What I think could improve this it to add a test for after you have practised a segment. For example: I have just finished the testing for all 3xx status codes. ’

C.11 Speed mismatch

- From response 3:

  > Sometimes I completed a module too fast and felt like I just kind of learned the factoids. Like, I knew that a given term was either this or that number, but usually either had to guess or spend a lot of time pondering.
  >
  > Sometimes the number of correct answers seemed high and really challenging, and other times very low.

- From response 4:

> The speed training was really difficult if you had fast typing speed because I felt like it expects you to get alot more rights answers in a shorter time and if you made 1 mistake you could not continue until you had everything right with no mistakes.

- From response 17:

  > when the user is warming up it should take the data from the warm up (the typing speed) and update the expected typing speed of the user. Basically update the expected typing speed of the user, so that if the user has fast typing speed when doing the typing speed they wont be expected to answer questions fast on every speed test.
  >
  > - The respondent must have encountered a speed mismatch to get this wrong impression that the system doesn't do this.

- From response 24:
  ' some of the answer were too long to type in when having a timed trial i my opinion '
- From response 46:
  ' While it will look like i tried to game the system after status code 2xx (i think) i slowed down to actually remember things, and not go at my ultimate speed. '
- From response 50:
  ' Should be able to lower difficulty. I needed 13 correct to continue, which is fine for 1xx status codes, but super hard for 4xx status codes. '

C.12 Stuck on 4xx modules

- From response 2:

  > ...getting stuck on the 4xx codes made it hard to learn and progress...
  >
  > ...there are so many 4xx codes.
  >
  > I think dividing those up into smaller bundles of codes, to get a more focused learning approach.
  >
  > ...getting stuck on 4xx codes for too long made me start to forget about the 1xx, 2xx and 3xx codes.

- From response 3:
  ' I talked to other people who said the section with 4xx codes was hard to pass. '
- From response 37:
  ' [About external factors] Yes, especially when it came to 4xx and 5xx status codes I read up on the codes trying to understand them and differentiate them better, as those modules were more challenging than the previous ones. '

- From response 46:
  ‘ Furthermore the 4xx status codes were a PAIN to go through, training should at least have been split into 4 parts, and the final test of that module was a pain in and of itself. ’
- From response 50:
  ‘ I needed 13 correct to continue, which is fine for 1xx status codes, but super hard for 4xx status codes. ’

C.13 Thinking break

- From response 8:
  ‘ It was extremely demotivating during the speed tests when i was typing as fast as i could with no mistakes, but i was thinking for less than 3s on some prompts and that caused me to not be quick enough. ’
- From response 34:
  ‘ I had to make a txt file with all the promtps for many of the speed tests to study them between tries when i simply wasn’t fast enough to think and type at the same time ’
- From response 40:
  ‘ Some of the speed tests require too many words to complete. 15 words in 30 seconds when one has to think for a second what to answer can be a little bit too much ’

Appendix D

Verbatim free-text responses from questionnaire

The following is the complete set of free-text responses from the questionnaire. Each response is numbered and annotated with which part of the questionnaire it came from. When referenced in the coding for the thematic analysis, the same numbers as in this document are used.

[[1.2.33]] ->

1. MOTIVATIONAL: I think the scoring system for Phase 2 was too rigid. One typo or error means that you have to take the test again. As you are informed of the number of correct answers needed, this may discourage trying to get as many as possible correct, because the more answers you type, the bigger the risk of a typo. Also, after a typo, it would be less fun to type more answers as you know you failed anyway. Some type of scoring system might have made this more fun, for example that if you had an error, you could compensate by having more correct (e.g. one error could be compensated for by typing n correct answers in a row). I would also like to have an optional timer or progress bar to see the remaining time.
2. MOTIVATIONAL: I did enjoy using the tool to learn status codes. However, getting stuck on the 4xx codes made it hard to learn and progress, as a single code could get shown multiple times, then you barely get to learn it, and then it does not show up for another 10 tries because there are so many 4xx codes. And then you forget that one code you just learned. I think dividing those up into smaller bundles of codes, to get a more focused learning approach. Also, getting stuck on 4xx codes for too long made me start to forget about the 1xx, 2xx and 3xx codes.
3. MOTIVATIONAL: First thought on the questionaire is that it should have been posted right after completing the training, that way you will have "fresh" feedback. There should be options to customize what you want to practice. Sometimes I completed a module too fast and felt like I just kind of learned the factoids. Like, I knew that a given term was either this or that number, but usually either had to guess or spend a lot of time pondering. I noticed that both speed and accuracy was greatly dependent on time of day / tiredness and if there were any distractions while doing the tests. Sometimes the number of correct answers seemed high and really challenging, and other times very low. Some words were also unfamiliar on a keyboard and typing them was a pain. A couple of times the system "froze" and I had to reload. It happend to a classmate too. I talked to other people who said the section with 4xx codes was hard to pass. The last section with everything could also be challenging. Overall i really liked the system and I feel like this is a great way to learn factoids that would be pretty boring just to root learn by reading. If possible let students make their own sets. It would be a great addition to other courses too.
4. MOTIVATIONAL: The speed training was really difficult if you had fast typing speed because I felt like it expects you to get alot more rights answers in a shorter time and if you made 1 mistake you could not continue until you had everything right with no mistakes. It is easier to make mistakes when typing fast yet it still expects you to answer all the questions right within such a small period of time.
5. MOTIVATIONAL: I got bored after 3-4 minutes, and i felt that it was just an endless cycle
6. MOTIVATIONAL: For further improvements of the system, I think it should be possible to train for specific status codes. For instance, if I feel unconfident on the 1xx status codes, it would be great if there was an option to train for these specifically.
7. MOTIVATIONAL: Some things were over repeated, other not repeated enough, if there was a way for users to repeat sections at will I think the system and users would benefit a lot
8. MOTIVATIONAL: It was extremely demotivating during the speed tests when i was typing as fast as i could with no mistakes, but i was thinking for less than 3s on some prompts and that caused me to not be quick enough. I was stuck on 12-13 of 14 needed for 5-10 attempts in a row, but i was simply not fast enough. It can only be described as annoying to be stuck like that
9. MOTIVATIONAL: Beacuse the curriculum did not feel useful the perception of the process

gets colored. But i think using a tool like this to memorize things could be useful

10. MOTIVATIONAL: Apart from the repetitiveness, the system was a good learning resource. I think the theme of the tests, HTTP status codes, exacerbated the repetitiveness. The system would perhaps work even better with a more carefully selected theme where it would compliment the system, rather than draw out one of it's weaker sides; repetitiveness. I realize that having breaks between sessions probably also would have helped, but unfortunately I didn't read that part of the wiki page before I was finished with all the modules. This resulted in me sitting a couple hours at a time working on the modules. So maybe display the importance of having breaks in the actual system, reminding the user to complete it over a longer period of time. Having speedtests made it more engaging, where I felt rewarded after completing each module. The sound effects were also nice.
11. MOTIVATIONAL: felt very forced. you couldn't jump forward if you new the answers or jump back to repete hard answers. The tests were too long sould be more shorter test. For example a test for five and five questions, then a test for 10 (two of the earlier five questions) and 10 questions, then 20 and 20 questions and so on. At the end you can have a big test of questions you have repeted many times so you would know the answers. Sometimes you got the same question you just answered. It was very tempting to cheet and google the answers.
12. MOTIVATIONAL: The long test phase was really difficult and repetitive.
13. USABILITY: If you stop between the parts referred to as phase 1 and phase 2 in the first question, and take it up again some days later, then you have no opportunity to refresh your memory using the system; only the test is available. So, I think there should have been a possibility to revise phase 1 as needed, and in this aspect I think the system is a bit too rigid. Or even better, to revisit each module when you want.
14. USABILITY: Would be nice to have an option to override incorrect answers. F.ex if you had a minor typo you could click a button to override the incorrect answer and make it correct.
15. USABILITY: Questions that required a very long answer to be written perfectly time after time were very frustrating.
16. USABILITY: Letting you know again that I really would like to have access to the system with my own set or factoids. I made a good deal of typos, but still it made me better at both reading carefully and I also think that it made the facts stick better as you had to go an extra round.
17. USABILITY: when the user is warming up it should take the data from the warm up (the typing speed) and update the expected typing speed of the user. Basically update the expected typing speed of the user, so that if the user has fast typing speed when doing the typing speed they wont be expected to answer questions fast on every speed test.
18. USABILITY: The typo fails were really annoying. You could get a lot more correct than what was needed for the test, but still fail because you accidentally hit enter twice in a row once
19. USABILITY: Hitting especially the longer status code texts or the ones with for example " ' " in the text was difficult, as i usually let auto-correct take care of that
20. USABILITY: (Having to use a vpn is tedious, I might have used it more if it wasn't for that") Also backspacing takes time
21. USABILITY: The system was to stringent on the typos. I spent too long on certain codes because of difficulties typing the entirety of the phrase quickly. I would like a way of going back to earlier codes too ensure i remember them.
22. USABILITY: A lot of typos made the experience longer than needed.
23. USABILITY: Good usability, intuitive to use. Important that it tested typing speed, which made the system adapt to your level.

24. USABILITY: some of the answer were too long to type in when having a timed trial i my opinion
25. EXTERNAL: Did some reading outside the system, but no active memorization techniques.
26. EXTERNAL: I'm using Anki flashcards
27. EXTERNAL: I tried not using anything to practice the codes, as I thought the tool was supposed to be used without anything else. Due to either funny code meanings (418) or actually talking about specific codes at a time, I learned some of them in a natural way.
28. EXTERNAL: The beginning of the first part was to challenging in my opinion, because when you only know the status codes 200 and 404 it feels pointless to sit and guess new status codes. Maybe it could be an idea to have hints in the beginning phase of such a learning tool. F.ex have html-tag abbriviations on status codes, so that when you hover the mouse over an status code you can read what is it.
29. EXTERNAL: I didn't use any techniques other than using the codes for work in the cloud course.
30. EXTERNAL: I used mnemonics in the end of the DIPT. But did not remember much of the mnemonics during the retention test.
31. EXTERNAL: Difficult to remember it all, and found myself searching it up again and again.
32. EXTERNAL: No, i read through the 4xx status codes before the big test there to remember them all as training was looooooong, and i forgot the early ones halfway
33. EXTERNAL: Not anything beside what I used in code in the Cloud course
34. EXTERNAL: I had to make a txt file with all the promtps for many of the speed tests to study them between tries when i simply wasn't fast enough to think and type at the same time
35. EXTERNAL: I did a bit of reading on the codes outside of dipt, as the site offered no comprehensive list of the codes i had to learn.
36. EXTERNAL: No other techniques, but I did Google to see the whole list of status codes. F.exp 5xx status codes
37. EXTERNAL: Yes, especially when it came to 4xx and 5xx status codes I read up on the codes trying to understand them and differentiate them better, as those modules were more challenging than the previous ones.
38. GENERAL: This tool could be quite useful for learning codes better. However, getting stuck on a specific set of codes will hinder progress both in learning other codes, and will keep you from practicing old ones. I suggest creating a "chapter" system where you can choose any old ones you have done, but new ones are not unlocked before you have completet or done a certain number of (proper) tries.
39. GENERAL: I wanted to train more but as far as I understood once you have finished training it just says "come back later"
40. GENERAL: Some of the speed tests require too many words to complete. 15 words in 30 seconds when one has to think for a second what to answer can be a little bit too much
41. GENERAL: I feel like this system can be really useful for other subjects, but i'm not sure how relevant it is to learn network status codes by hand.
42. GENERAL: It was kind of addicting. Keyboard dopamine.
43. GENERAL: Overall a good way to learn but it expected you to get everything right on a speed test before you could move on. I was focused on beating the high score in time, which was difficult in itself, but the fact that you could not misspell once or else the speed test was invalid made it way harder.
44. GENERAL: Very good for short term memorization, but is not enough by itself for long term memorization.
45. GENERAL: After the very first test I thought I would get to learn various status codes

right away, but instead I was faced with lots of training on GET, PUT and POST requests. I wish it was possible to pick and choose categories to learn/test. (and maybe specific codes as well, like only 3xx codes, 1xx codes, etc.,)

46. GENERAL: While i do understand how it is meant to teach, i find the product unsatisfying to say the least. I take full responsibility for not doing the pre-test properly, however putting 71 status codes in front of me before we had even started to learn properly what each one meant was brutal and a horrible way to do things. Furthermore the 4xx status codes were a PAIN to go through, training should at least have been split into 4 parts, and the final test of that module was a pain in and of itself. Through luck you could get 3-4 of the same status code or the easy ones in a row(!!!!!) which made it largely unrepresentative in my mind. While specified on git that one should not grind the status of the task in 4xx provided no other option, as doing 15 minutes each day would sabotage the final test. The system was also easily cracked but you knew that (burp suite). Final test was also a nightmare, as you relied on luck there too. While it will look like i tried to game the system after status code 2xx (i think) i slowed down to actually remember things, and not go at my ultimate speed. The tool is not suited for the task, as it lacked a recap feature, adjusted the amount of correct answers needed if you messed up one, lacked any motivation exempt from the grade. The fact that 4% of the grade was locked behind DIPT was also a source of major irritation among my peers, as it has very little to do with our actual exam. PS. DO NOT TELL US THAT IT IS NOT VERY EXAM RELEVANT, it killed motivation. DIPT is basically glorified quizlet, and not a tool suited for me atleast. TLDR: pre-test was bad, 4xx module was bad, boring to do and felt like a chore.
47. GENERAL: I fluked on the pre-test and cheated, I looked up quite a few codes I had misundersood that this was supposed to be a memory test, so that is why the result is way worse at the post-test. I do know I have gotten a little better a least
48. GENERAL: At the end of the training set when we had to do prompts from all segments I barely remembered the 1xx status codes. What I think could improve this it to add a test for after you have practised a segment. For example: I have just finished the testing for all 3xx status codes. Now i have a test that tests me on everything i have learned so far. Just like the last test, but with only what i had learned so far. And one of those after each of the big status code segments
49. GENERAL: The site should have a more detailed progress bar, and ways of “retraing” on completed modules
50. GENERAL: Should be able to lower difficulty. I needed 13 correct to continue, which is fine for 1xx status codes, but super hard for 4xx status codes.
51. GENERAL: Overall it was a good system. Having multiple forms of input for the same concepts made it easier to memorize, like having to remember both the status code and later the description. Repetitiveness should in my opinion be the biggest take away.
52. GENERAL: Implement the training part to train on the specific module you’re working on, or make it possible for the use to choose module or choose which module to train on.

Skrevet: Fri Apr 7 12:33:12 PM CEST 2023

Appendix E

Serious games / gamification review

The following is an unpublished paper delivered as part of course work during the spring of 2022. Due to the paper being referenced in the present thesis, and being a significant part of the background for it, this paper has been included as an appendix for the convenience of the reader.

Evaluating the learning gain improvements of serious games and gamification: A systematic literature review

Svein-Kåre Bjørnsen
Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
sveinkab@stud.ntnu.no

***Abstract*—This review investigated the empirical evidence on the efficacy of gamification and serious games on the improvements gained in skill- and knowledge acquisition in educational settings. Novel metrics for simple effect size comparison were evaluated as part of analyzing the results from the reviewed papers. No heed was paid to usability, self-efficacy or other "soft" outcomes of the reviewed studies, but focus was strictly laid on directly measured increases in knowledge or skill from pre-tests to post-tests. Only English journal and conference papers that had at least one experimental group and a control group with pre- and post-tests, and which reported p-values ≤ 0.05 were included, among other filters. No restriction was made on time-period. A quite small number of papers passed all filtration, 14 out of 7410, indicating a lack of good research in this specific area. Two of four metrics were found useful to give insight into the data, and showed a big variance in the efficacy of the game-based approaches reviewed. Most of the results were highly positive, but a few were highly negative in terms of the game-based group outperforming the control group. Few papers implemented known educational theory in their products. The author advises, among other things, more fine-grained data collection, standardized reporting of results, and implementation of known educational techniques for future designs to make results more consistent.**

***Index Terms*—gamification, serious game, efficacy, learning gain improvement, applied behavior analysis**

I. Introduction

The research on positive effects of games have exploded during the recent decades, and serious games for education are a big part of this, however the evidence of efficacy for skill or knowledge acquisition seems to be rather weak [1].

In this review, I investigate the effect sizes on acquiring knowledge or skills found in the empirical literature on serious games and gamification applied to educational contexts. While doing this I am only interested in quantifiable results that are not based on self-reporting, holding a behaviouristic view on learning, where affect and mental processes are not necessary for describing, predicting and controlling behavior [2].

Based on an informal preliminary exploration of the literature on serious games for education, I observed that the studies encountered were generally weak, and rarely commented on the size of the effect. Percent-based metrics were then made to compare results meaningfully. In this review, part of the aim is to present these metrics and evaluate their usefulness.

Because a very small ratio of quality papers had earlier been identified in the literature [1], [3], scripts were made to filter through a large body of papers in search of the ones that would be relevant.

From over 7400 papers, only 14 met all criteria for inclusion and avoided exclusion based on relevance and quality, confirming the suspicion of still few quality papers in this sub-set of the literature. The results were mostly quite positive, gamified applications or serious games often showing above 100% more improvement in skill and knowledge acquisition than traditional methods. However the variability was quite high and some results showed negligible improvement for both experiment and control, in addition to a few studies that showed quite lower achievement on the part of the experimental groups compared to traditional instruction.

II. Background

A. Definitions

Before going deeper into the subject of serious games and gamification, I will first define these two terms as used in the present paper.

Dörner et al [4] define a serious game as a digital game that is meant to entertain, but to also achieve at least one other goal. This definition is fitting, and coincides with other definitions that may be seen in the literature, however, it limits the scope to digital games. I think physical games are just as valid for the discussion on serious games, and for this review a serious game is a game that conforms to the above definition, except it may be a physical game as well.

However, Dörner et al's definition of *gamification* was a bit too vague: the use of game elements in non-game areas. When exploring the literature, it seemed that a highly cited definition was Deterding et al's "'Gamification' refers to the use of design elements characteristic for games in non-game contexts." [5, p. 13, formatted to sentence]. This is the one that will be used in this paper.

B. Preliminary exploration of the literature

Preliminary informal exploration preceding conception of this review showed a tendency within the serious games literature. It seemed that it might not be very rigorous in investigating efficacy of the applications in relation to the game effectiveness on improvement of learning outcomes.

The pre-test post-test quasi-experimental design seemed the most prevalent when the papers were discussing actual measurements of learning gains, and not simply usability concerns or perceived learning gains. The general state of affairs seemed quite bleak: My impression was that there were thousands of papers in the literature, but only a few percent of them were empirically rigorous in terms of measuring any concrete gain in skill or knowledge—any actual learning. Opposed to the affect or opinions or attitudes of the users of the games.

This also seemed to be a secondary concern in the literature. Games and gamification may not be valued for their direct improvements in the amount or depth of what is learned, but rather the increase in user engagement they supposedly have on the students. However, the preliminary exploration didn't yield studies that had rigorous control even for this aspect.

Another limitation of the papers found in the exploration were that they mostly didn't report on effect size at all, and in a few cases *Cohen's d* was given, but it's implications were not explained — the concrete effects on learning gain from the games were not intuitively apparent.

To give a brief explanation of *Cohen's d*: it is a statistic measure of the size of an effect on a dependent variable between two groups of a t-test. The number is a ratio of the difference between the two means over the common standard deviation. The unit of the ratio is thus the difference in number of standard deviations for the two groups, and the original measure assumes that the standard deviation is the same between the groups [6]. Alone, this number isn't very intuitive, but it can be interpreted, via a table, as a percentage of how many samples from one group have a higher score than the mean of the other. Cohen does give some "rules of thumb" for what some different ranges of the value may signify: 0.2 being "small", 0.5 being "medium" and 0.8 as "large"; however he warns that these words are relative to the domain and study [6].

Part of the preliminary exploration were the two systematic literature reviews by Connolly et al [3] and Boyle et al [1], the latter being an update of the first. These looked at the state of the research into positive effects of games across genres, categories and domains, and together they covered the decade 2004 to 2014. Both papers highlighted weaknesses with the evidence on effectiveness of serious games.

Connolly et al [3] found that the evidence for games improving learning was weak, noting a lack of randomized controlled trials (RCT) investigating knowledge acquisition. Boyle et al's update [1] found an explosion in research being conducted in this field, but the ratio of quality papers were even bleaker than before: only about 0.3% of the papers hit by their search terms were relevant and of high quality. In both cases the reviews did not only look at games related to education, but any positive effects of games, both purpose-built serious games and commercial off-the-shelf (COTS) games.

Additionally, the present review only concerns serious games in an educational setting, isolating the outcome of interest to knowledge or skill acquisition. Which presumably results in a further narrowing of the field in terms of quality.

This leads to the motivation for the current review: There seemed to be a lack of quality studies that measured learning gains directly, and I wanted to see if this was plausible or indeed correct. Also, the few studies encountered in the exploration that did measure learning gains directly, did not in any satisfactory manner tell me how much the students learned with the game compared with more traditional methods. What was generally the case, was a statistical dump of means and differences between them, as well as a p-value and a statement of statistical significance, without first establishing the threshold for the significance level was to be acceptable. The means were also often not supplied with a unit, so it could be percent-based, discrete question count or other; often it was never specified. I wanted a better way to compare the effects of the different games with each other, and came up with a few simple calculated metrics that seemed to give an intuitive insight into the effect size of the games, without having Cohen's d explicitly supplied or needing to calculate it. It would also be interesting to see how well these metrics would perform in a larger context than the small sample available at the point of the preliminary exploration.

C. Learning Gain and Learning Gain Improvement

The simple calculated metrics are what I call learning gain (LG), learning gain improvement (LGI), learning gain rate (LGR) and learning gain rate improvement (LGRI). Their specific definitions are as follows:

Learning gain (LG) of a group within a study on an educative method is the percent-wise difference between the mean performance on the pre-test and the post-test. Separate groups of participants within a study should have different learning gains. The formula for this ratio is given below:

$$LG = \frac{post\ mean - pre\ mean}{pre\ mean} \tag{1}$$

Learning gain improvement (LGI) of an educative method is the percent-wise difference between the control and experiment groups in learning gain. It is positive if the experiment had higher gains than the control, and the formula is given below for clarity:

$$LGI = \frac{LG_{experiment} - LG_{control}}{LG_{control}} \tag{2}$$

Both of these metrics are made to be insensitive to the unit of measurement in different studies, for them to be useful for between-study comparison. Using percentages, they may be more intuitive in explaining the effects of a given method. Note that they may not only apply to serious games

or gamification, but any educative method. They are simple percentage differences.

It was also relevant to see improvements in time efficiency of methods, therefore some additional metrics related to rate of learning were made: Learning gain rate (LGR), and learning gain rate improvement (LGRI).

The former is the amount of learning gain by the above definition that a group has per hour spent with an educational method, while the improvement is the percent-wise difference between the control and experiment, like LGI is for learning gain.

$$LGR = \frac{LG}{hours\ spent} \tag{3}$$

$$LGRI = \frac{LGR_{experiment} - LGR_{control}}{LGR_{control}} \tag{4}$$

D. Goal and research questions

The goal of this review is to investigate the overall effectiveness of serious games and gamification in relation to learning gains in an educational setting. New calculated metrics are proposed in an attempt to make that comparison of effect sizes more intuitive, or indeed possible, and also easier to communicate to policy makers, teachers, parents and other stakeholders in gamification and serious games endeavours. These metrics need to be tried out in practice to evaluate if they may be beneficial or viable at all to the current literature.

To meet this goal, the following research questions were formulated:

1) What are the improvements observed from gamification and serious games on learning gain and learning gain rate compared to traditional teaching?
2) What are the learning gain improvements and learning rate improvements produced by the application of the different learning theories used?
3) How useful are the proposed learning gain metrics for gaining insight into the literature?

III. Method

A. Search terms

The search terms were designed to catch as much as possible of the literature relevant to both serious games and gamification in an educational context, focusing on empirical findings, and attempting to disregard literature reviews. There are two equivalent versions due to differences in the syntax of the search engines used, both are shown in table I.

B. Literature databases used

The following databases were selected because of their ease of bulk exporting of citation information.

- Web of Science
- ScienceDirect
- Emerald Insight

TABLE I
Search terms

Database	Search terms
IEEE Xplore	(”serious game” OR ”game for learning” OR ”gamification”) AND (”analysis” OR ”study” OR ”empirical”) AND (”training” OR ”learning” OR ”knowledge”) AND NOT ”review”
All others	(”serious game” OR ”game for learning” OR ”gamification”) AND (”analysis” OR ”study” OR ”empirical”) AND (”training” OR ”learning” OR ”knowledge”) -review

- ACM
- ERIC
- IEEE Xplore
- PubMed
- PubMed Central (PMC)

C. Procedure

A protocol was made to keep track of the many filtration steps needed for both inclusion and exclusion. Following the protocol, all databases were queried with the search terms. The citation information for all results were downloaded in RIS or BibTeX format, whichever was available, but RIS was preferred. The entirety of the search and downloading of citation information was done on 6th of April, 2022. This information was then used to filter the results for inclusion and subsequent download before exclusion filters were applied. All the inclusion filtration was performed programmatically by a set of tailor-made python scripts. The number of papers after filtering for inclusion was initially still in the thousands, and because of time and resource constraints, an extra inclusion filter was added. This filter looked for either of the terms ”efficiency”, ”efficacy”, ”pre-test” and ”post-test” in the abstracts. At least one of the terms had to occur for the paper to be included for download.

The complete set of inclusion filters were as follows:

1) Remove duplicates by title in citation information.
2) Include all results that are English journal articles or conference papers containing ”serious game”, ”game(s) for learning”, or ”gamification” in title, or abstract.
3) Include only papers which citation information has content in all the following fields: title, authors and abstract.
4) Narrow down the selection further by only including papers where at least one of the terms ”efficiency”, ”efficacy”, ”pre-test” and ”post-test” can be found in the abstract.

After filtering for inclusion, all the remaining articles were attempted downloaded manually and a preliminary scripted exclusion filter, searching for terms in the whole content of the downloaded files was applied before filtering the rest manually. During the download step two papers were found written in a non-English language and excluded; and a further 26 papers were not possible to download because of a lack of subscription on the side of my university.

The exclusion filters were as follows:

1) Exclude papers that don't contain any of the terms "pre-test", "pretest", "post-test", "posttest", "pre-" and "post-" in the entire text.
2) Manually exclude duplicate papers that are still in the collection.
3) Manually exclude papers that aren't empirical studies. This also includes meta-analyses and literature reviews that have gotten through previous filters as well as qualitative studies, since these wouldn't have the data needed for extracting the metrics in question. Protocols for studies aren't "finished" studies, and are thus also excluded
4) Manually exclude papers that don't investigate the efficacy of gamification or serious games in an educational setting. Learning gain in the form of an increase in either knowledge or skill must be measured, and in relation to traditional methods. Attitudes, affect, usability and other factors are not of interest for this review. If uncertain, the article will be excluded. Measuring motivation scores, or perceived learning, is not sufficient, knowledge or level of skill must be measured directly.
5) Manually exclude papers that don't have both a pre-test and a post-test.
6) Manually exclude studies without a means of control.
7) Manually exclude papers where none of the reported p-values are equal or below 0.05.
8) Manually exclude papers where there has been employed an independent samples t-test between pre- and post-tests of groups without thoroughly demonstrating that the group members are homogeneous enough for an independent samples test to be valid [7]. A paired t-test, or a different statistical test must otherwise have been used instead. Papers that don't mention which statistical tests have been done are also excluded at this step.
9) Manually exclude papers where it is still impossible to calculate any of the proposed metrics.

The remaining papers after final exclusion filtration were examined for extraction of data and calculation of the metrics in question, as far as possible. The papers were also examined for any applied learning theories or principles, reference of effect size and any values for *Cohen's d* were extracted.

The papers were also classified into two categories, "serious game" and "gamification", based on what sort of application they were investigating. I disregarded what the texts themselves classified the instance as, but looked at the implementation and followed the definitions given in section II-A. The final classification is given in table VIII.

IV. Results

The search hit 7410 papers in total, the largest contributors of which were *PubMed*, *ACM* and *ScienceDirect*, as can be seen in table II. The final number of papers after filtration was 14, as can be seen in table III, along with specification of every step of the filtration.

Group sizes are given in table IV, the extracted data for the pre- and post-tests are given in tables V and VI respectivley.

TABLE II
SEARCH RESULTS

Database	No. of results
ACM	1465
Emerald Insight	144
ERIC	688
IEEE Xplore	809
PMC	230
PubMed	2366
ScienceDirect	1207
Web of Science	501
Total	**7410**

TABLE III
FILTRATION OF SEARCH RESULTS

Step	Removed	New total
Deduplication	131	7279
Main inclusion filter	4189	3090
Abstract-based inclusion filter	2809	281
Download	26+2	253
Scripted, non-pre-post content	86	167
Manual deduplication	8	159
Manual, non-empirical studies	17	142
Manual, not educational efficacy	83	59
Manual, not both pre- and post-test	13	46
Manual, no control	21	25
Manual, no p-value ≤ 0.05	1	24
Manual, inproper use of t-test	6	18
Manual, unable to calculate metrics	4	14
	Final total:	**14**

TABLE IV
GROUP SIZES FOR THE EXPERIMENTS IN THE REVIEWED PAPERS

Paper	Grouping	Size control	Size experiment
[15]	—	22	22
[16]	a	26	25
	b	26	17
	c	26	27
[17]	knowledge	50	50
	skill	50	50
[18]	linguistic	207	339
	play	207	339
[19]	b	26	15
	c	9	12
[20]	—	46	57
[21]	—	125	144
[22]	—	88	79
[23]	—	39	38
[24]	—	47	44
[25]	badge	23	27
	goal	23	22
	both	23	25
[26]	a	37	62
	b	37	35
[27]	—	155	73
[28]	—	58	62
Mean		61.36	71.09
SD		58.58	91.51

TABLE V
PRE-TEST DATA IN THE REVIEWED PAPERS

Paper	Grouping	Mean Control (SD)	Mean Experiment (SD)
[15]	—	37.27 (?)	36.31 (?)
[16]	a	2.88 (0.993)	2.56 (0.821)
	b	2.88 (0.993)	2.24 (0.97)
	c	2.88 (0.993)	2.78 (0.892)
[17]	knowledge	78.64 (16.74)	74.23 (10.56)
	skill	82.57 (11.11)	87.91 (12.67)
[18]	linguistic	2.63 (0.92)	2.45 (1.07)
	play	0.44 (0.84)	0.35 (0.78)
[19]	b	0.784 (0.141)	0.769 (0.146)
	c	0.855 (0.084)	0.821 (0.122)
[20]	—	11.39 (2.26)	12.02 (2.23)
[21]	—	3.8 (?)	3.4 (?)
[22]	—	11.78 (8.88)	11.78 (9.94)
[23]	—	14.15 (2.23)	12.63 (3.37)
[24]	—	43.4 (16.67)	55.15 (16.03)
[25]	badge	7.17 (2.44)	7.81 (2.68)
	goal	7.17 (2.44)	6.95 (2.17)
	both	7.17 (2.44)	6.8 (2.06)
[26]	a	12.83 (3.08)	14.2 (3.27)
	b	13.49 (3.86)	12.37 (3.36)
[27]	—	6.25 (1.3)	6.46 (1.3)
[28]	—	4.49 (1.5)	4.46 (1.32)
Mean		16.13 (4.00)	16.57 (3.79)
SD		23.53 (5.14)	24.51 (4.60)

TABLE VI
POST-TEST DATA IN THE REVIEWED PAPERS

Paper	Grouping	Mean Control (SD)	Mean Experiment (SD)
[15]	—	54.32 (8.06)	74.54 (7.05)
[16]	a	4.08 (1.831)	5.52 (2.163)
	b	4.08 (1.831)	4.71 (1.863)
	c	4.08 (1.831)	5.22 (2.207)
[17]	knowledge	79.53 (8.26)	82.56 (9.35)
	skill	84.35 (15.88)	94.23 (6.89)
[18]	linguistic	3.04 (?)	3.05 (?)
	play	2.9 (?)	3.0 (?)
[19]	b	0.837 (0.141)	0.907 (0.089)
	c	0.889 (0.089)	0.904 (0.089)
[20]	—	11.76 (2.26)	16.46 (1.86)
[21]	—	4.7 (?)	5.0 (?)
[22]	—	21.47 (10.2)	20.94 (9.93)
[23]	—	19.916 (?)	22.744 (?)
[24]	—	71.9 (15.55)	76.51 (16.01)
[25]	badge	11.74 (3.25)	13.63 (3.77)
	goal	11.74 (3.25)	11.73 (3.52)
	both	11.74 (3.25)	11.48 (4.75)
[26]	a	23.49 (5.06)	22.76 (4.19)
	b	15.89 (5.59)	15.54 (5.35)
[27]	—	6.42 (1.5)	6.94 (1.4)
[28]	—	9.02 (0.95)	6.4 (1.44)
Mean		20.81 (4.93)	22.94 (4.55)
SD		26.24 (4.85)	29.42 (4.08)

The question marks in these tables signify that the value wasn't supplied by the paper.

In almost half the papers reviewed, there were more than a single experimental group. In these cases, data were extracted for all experimental groups using the game or gamified method, and if there was only a single control group, it was used for calculating the learning gain metrics for all experimental groups. Experimental groups using alternative methods for control against the novelty effect or other reasons were ignored for data extraction, since there was essentially

TABLE VII
LEARNING GAIN METRICS FOR THE REVIEWED PAPERS

Paper	Grouping	LGC	LGE	LGI
[15]	—	45.75%	105.29%	130.15%
[16]	a	41.67%	115.63%	177.50%
	b	41.67%	110.27%	164.64%
	c	41.67%	87.77%	110.65%
[17]	knowledge	1.13%	11.22%	891.56%
	skill	2.16%	7.19%	233.49%
[18]	linguistic	15.59%	24.49%	57.09%
	play	559.09%	757.14%	35.42%
[19]	b	6.76%	17.95%	165.46%
	c	3.98%	10.11%	154.23%
[20]	—	3.25%	36.94%	1037.10%
[21]	—	23.68%	47.06%	98.69%
[22]	—	82.26%	77.76%	-5.47%
[23]	—	40.75%	80.08%	96.52%
[24]	—	65.67%	38.73%	-41.02%
[25]	badge	63.74%	74.52%	16.92%
	goal	63.74%	68.78%	7.91%
	both	63.74%	68.82%	7.98%
[26]	a	83.09%	60.28%	-27.45%
	b	17.79%	25.63%	44.04%
[27]	—	2.72%	7.43%	173.17%
[28]	—	100.89%	43.50%	-56.89%
Mean		62.31%	85.30%	157.80%
SD		114.99	153.92	274.10

TABLE VIII
CLASSIFICATION OF PAPERS

Paper	Classification
[15]	gamification
[16]	serious game
[17]	gamification
[18]	serious game
[19]	gamification
[20]	serious game
[21]	serious game
[22]	gamification
[23]	gamification
[24]	gamification
[25]	gamification
[26]	serious game
[27]	serious game
[28]	serious game

TABLE IX
NUMBER OF PAPERS PER CATEGORY

Category	No. of papers
Gamification	7
Serious Games	7
Total	**14**

TABLE X
NUMBER OF EXPERIMENTAL GROUPS PER CATEGORY

Category	No. of experimental groups
Gamification	11
Serious Games	11
Total	**22**

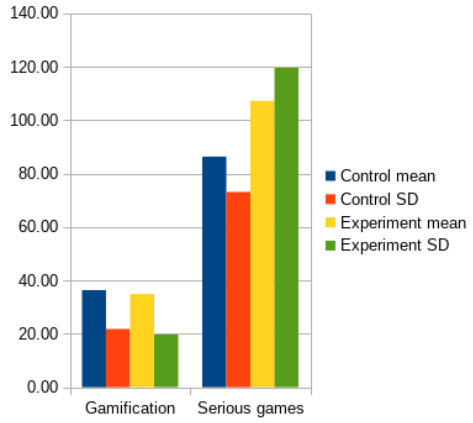


Fig. 1. Group sizes means of gamification vs. serious games.

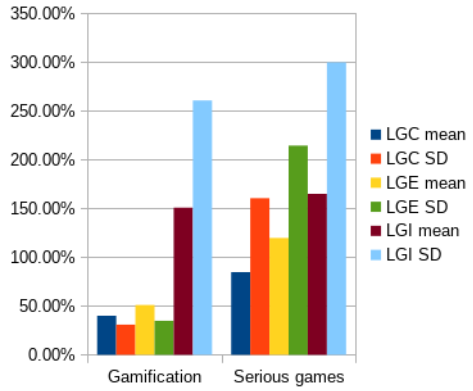


Fig. 2. Learning gain metrics across categories.

just two such cases. The calculated learning gain metrics are given for each experimental group within each paper in table VII. Note that the standard deviations in final row of this table given in percentage points.

The time spent on app or game versus traditional instruction was not precisely measured for the great majority of papers, so the attempt of calculating the rate-based metrics was quickly discarded.

An interesting coincidence is that the number of results classified as "gamification" and "serious games" were *equal* both when looking at the number of papers and the number of experimental groups. As can be seen in tables IX and X. The detailed classification is shown in table VIII.

There are a few differences between the gamification and serious games, as reviewed in this paper. The larger group sizes are found within serious games, as shown in figure 1. Serious games also show larger learning gains, for both control and experiment, and also LGI, than gamification, as shown in figure 2

A. *Specific remarks on the reviewed papers*

Barros et al [16] investigated the effects of a serious game on learning polynomial operations. They had three experimental groups and one control group, which were convenience sampled as four different 8th grade classes at a Portuguese middle school. The difference between the experiments were the length of the time period that the participants had access to the game. Group *a* had access to the game for 22 days, group *b* had access to the game for 12 days and group *c* had access for 6 days. The control group didn't play the game. All the experimental groups were compared with the same control in this review with respect to the calculation of the metrics.

Garcia and Revano [17] investigated improvements in knowledge, skills, self-efficacy and attitude from participating in a gamified course teaching python programming. Only knowledge and skill were of interest in this review, so the data for the latter two were not extracted. Note that the grouping for this study in tables V, VI and VII signifies the type of learning measured, since the study split it into the different categories. This study also shows an absolutely incredible value for LGI for knowledge. However, the reason for this becomes apparent when taking a look at the learning gains for both groups. The mean score changed quite little, but the experimental group did in fact improve significantly more in comparison. Examining the standard deviations of the pre- and post means for this study, one can see that although the overall improvement was negligible, the variance in the group decreased. This indicates that the tail end approached the mean for the study, and noting that all of the participants had previous programming knowledge, it might also indicate that the low improvement in the groups can be due to the ceiling effect—the test content not being advanced enough to differentiate the improvement of the upper portion of the performers [8]. One may assume that most of the students knew much of the material from the beginning.

Manero et al [18] investigated the effects of a serious game on the interest and knowledge of high-schoolers in relation to theatre. They measured linguistic knowledge, theatre play knowledge, and student interest in theatre. Data for the knowledge tests were extracted. This study had two control groups, one of which used traditional teaching methods, whereas one used theatre actors to drive the session, thus attempting to also control for the novelty effect. For this review, however, only comparisons with the traditional control group were made for calculating the metrics.

Jodoi et al [19] compared learning of critical thinking through a non-gamified app, a gamified app, and no app at all but only normal university teaching. There were several groups from two different universities, and also a third collection where the participants had experience in debating. Data was extracted primarily only from the groupings that had a control subgroup without any app use at all, since in this instance the comparison in question is against traditional educational settings.

The control group in [20] did not receive teaching with traditional methods while the experimental group received the gamified experience. The gamification was as such "on top" of the usual curriculum and methodology. The control did however get access to the gamified experience after the experiment was completed—they were wait-listed.

Morris et al [25] compared the use of badges and learning goals, as well as the combination of them on learning about Turkish culture. They also had multiple experimental groups against a single control. Data for all three experimental groups were extracted and compared against the same control for calculating the metrics.

Su et al [26] investigated the effectiveness of a serious card game, teaching the regulation of human immunology. They ran their evaluation in two semesters, yielding two sets of experimental and control groups, and reported them separately. Both sets of results were also extracted separately for metric calculation.

Of all the papers reviewed, only three ([18], [26] and [28]) gave numbers for statistical effect size of their experiments one of which only had numbers for pairwise comparison of the experimental and control groups, but not between control and experiment [26], while the others reported numbers of effect between experiment and control. These numbers were given as *Cohen's d*. The numbers aren't reproduced here because being so few give little means of comparison between the articles reviewed, but they ranged from around -2 to 2, illustrating a big variance, and "huge" effects [9].

Only two papers cite specific theory that had been applied in the making of the solution to be evaluated, [16] and [17], these were *Gee's learning principles* and the *jigsaw teaching strategy* respectively. One paper [21] reported on a variety of elements being implemented, but it was unclear how any of them came into play, since the game was not described in detail. The reported elements were: Problem-based learning, spaced recall and active restitution. Two other papers, [22] and [26], cited traditional learning theories like Vygotsky's zone of proximal development [10] and Bandura's self-efficacy [11], however these were more of a commenting character than something that had been explicitly attempted applied. However, it should be mentioned that many of the reviewed papers measured self-efficacy through questionnaires as an additional affective measure in their experiments.

V. Discussion

Although the rate-based metrics were essentially impossible to evaluate with the designs in the present state of the research, LG and LGI showed some real promise. Even when the paper in question doesn't comment on effect size, they give a simple way for the reader to quickly and intuitively grasp this, and if given two examples of games to pick from, a reasonable comparison can be made.

However, the example of Garcia and Revano [17], it is clear that the LGI metric cannot be used in isolation to gain fruitful insight into the efficacy of a study. An incredibly large value may not mean that the experimental group learned a whole lot. In that study, students using the gamified solution improved 11%, while students in the control improved only 1%. The tests measuring gains may have been flawed, meaning that they were too simple to capture the upper bound of the actual learning happening, or students may simply have gained very little from the course overall, knowing most of the material from before. I am confident that I wouldn't have been able to make this inference from looking at the raw data alone.

One of the chief observations that one can derive from the results is the great variance in the results, standard deviations for the learning gain metrics categorically being higher than the means themselves, and indeed almost double the mean in all cases. This is a problem if one were to recommend the use of a serious game or gamified app in a classroom. An intuitive explanation is that they are quite simply different games and applications, that employ a range of different underlying mechanics, leading to differing results. This may stem from the design of these applications not being data-driven, overall, but specifically toward learner achievement, and generally lacking theoretical foundations. Neither theory or data has explicitly been used to guide design decisions in the majority of the reviewed papers.

Digital games have the potential of making education data-driven in a much larger scale than has been done traditionally [12], and the leap to include gamified apps in that sentiment is not big. Traditional learning management systems, like *Blackboard*[1] or *It's Learning*[2], simply do not have the possibility of measuring students' performance on a micro-level. They're dependent on a deliverable being delivered, and in the traditional case also for someone to rate that deliverable. With games, there is the possibility of knowing from second to second how each and every one student is performing in a given skill or piece of knowledge, and they can even dynamically adjust to findings therein [12]

Morningside Academy[3] in Washington, USA, has a 40-year long success-story of guaranteeing that their students will improve at least two grade levels in their worst academic skill in the space of a single year. Their approach is highly data-driven, and progress of each student is measured daily through the methods they use, although these methods are analog and physical in nature. This laboratory school credits much of their success to their use of the techniques known as Direct Instruction (DI), a near errorless instruction method, and Precision Teaching (PT), a fluency building technique used after DI, in the first years of attendance [13].

Direct Instruction has half a century of high-quality research showing *consistently* statistically significant and medium to quite high effect sizes as measured by *Cohen's d* [14]. The technique itself is based on programmed instruction, meaning that teachers follow a script when teaching the student group. The script is sequenced in such a way that it facilitates many responses from all of the students, at a high pace, and so that they make very few mistakes. All students also answer simultaneously, so that the teacher can retrace and make sure that the lesson doesn't continue until everyone understands the current step without mimicry. Fine-grained grouping is used to make sure that the students in any given group are as homogeneous as possible skill-wise so that progress through

[1] https://www.blackboard.com/
[2] https://www.itslearning.com
[3] https://morningsideacademy.org/about-morningside-academy/

the material can be as smooth as possible, and benefit every student of a given group [13]. Another important aspect of DI is that if some curriculum of this nature fails to be efficient, it is revised until it is; if a student fails to learn, the program—or the teacher's execution—is at fault, not the student [13], [14]. No signs of this could be seen in the papers reviewed, but this sort of "programmed" material may in the future find a place as an underlying mechanic within gamification or serious games for education. Thereby using well-established techniques for learning underneath the allure of gamification—both DI and PT rely on positive reinforcement, and games are flush with possibly reinforcing elements.

A. Answers

In terms of answering the research questions, the results have yielded some grounds to talk about them.

RQ 1 can not be answered clearly from this review. The data is quite variable, and the odds of getting an average of either a 5% improvement, 200% improvement, or a *decrease* of 50%, is not a clear cut one. Although, the results are *generally* quite positive, and many times over 100% better than traditional teaching with a teacher, which strongly indicate that serious games and gamification can have sizable positive effects on learning. Concerning the rate-based metrics, the underlying data didn't manifest through the reviewed papers, and it is impossible to say from this if there are any general learning rate benefits to using serious games or gamification.

RQ 2 was formulated to look for any connections between effects and the learning theories applied in the experiments reviewed. However, since very few of the reviewed papers contain actual application of theory, this questions is not possible to answer at this point. It may be a question for further empirical investigation.

RQ 3, how useful the proposed metrics are for gaining insight into the literature, is a bit easier to answer. At the present time, the rate-based metrics are useless, since almost none of the papers reported fine-grained time spent on both methods. The learning gain metrics themselves, however, LG and LGI, turned out to work well, yielding interesting discussion points on several fronts, and in my opinion they do give a more intuitive language for talking about the concrete effects of the games on learning. An added bonus is that one can compare effects of different papers, even if the authors themselves haven't given any statistical effect size, as long as the study uses pre- and post-tests. One possible draw-back is that they are only defined for situations where all the means in the study are non-negative values. Negative mean values give a quite incorrect picture of the effects, and render useless values for all the metrics. Although encountered only a single time during this review, it is something to keep in mind for future studies that wish to employ these metrics or facilitate their calculation.

Regrettably, the prevalence of papers that indeed have pre- and post-tests is miniscule, as measured by this review, lowering the usefulness in regards to the present literature. However, I can only presume that the trend between Connolly et al [3] and Boyle et al [1] of a dilution of quality studies in the explosion of literature may be continued and also true for this subset of the literature—gamification and serious games for learning knowledge and skills. A future, more thorough, review may however uncover more instances, and higher quality papers.

B. Limitations

The elephant in the room is the Hawthorne effect. How much of these incredible improvements are simply due to the novelty of the game or application? Only one of the reviewed studies had a control against this, measuring the effect of not only the game or application against traditional instructional approaches, but also the effect of some other arbitrary novelty, that didn't have obvious overlapping effect with gamification, as well. Some of the experiments, however, lasted for longer periods of time, making the novelty effect less likely to play it's part on the results.

Another limitation is the relatively small number of studies finally looked at. This could have perhaps been larger if not limited by having to download every paper one by one. Due to this necessity, the number of papers that could be screened by script was severely limited. This forced an aggressive inclusion filter based on abstracts alone. Some relevant articles may very well not have been included because of the terms filtered for didn't appear in the abstract. Granted, this is ultimately the fault of the present closed nature of scientific literature, and one can only hope that with time this will no longer be an issue.

The number of reviewed papers may also have been larger if more databases could have been incorporated into the search. Initially, several other databases were attempted searched, but the ones used were the only ones that I could comfortably download citation information for the entirety of the search results. This was still a laborious process, requiring downloading citations only for a single page of results at a time. *PubMed* and *ERIC* were good examples, where citations for large portions of, or even the entire search, could be downloaded in bulk.

C. Future work

In a future review, if having more resources and time, the authors may refrain from my abstract-based inclusion filter and search the entire content of a greater amount of papers. This may yield a larger number of articles for review, but the problem of manual downloads will remain for the foreseeable future. No matter how far the Open Access movement has come, programmatically aiding research is evidently still a challenge. However, a team should be able to download more papers if spreading the load on several people and over a longer period of time.

Another line of questions for further work comes up when deliberating the low number of papers uncovered and the variability of the results: Is there really such a difference as observed between the effects of gamification and serious games? And if it is genuine, is it large enough that the extra

budget of graphical and sound design, which a full-fledged game often needs, is worth it? And if the difference observed is indeed a general phenomenon, then why is it so? Are gamification attempts simply of lesser entertainment value, or is there some fundamental mechanical differences in how these two categories work upon the user to induce learning, one being more effective than the other?

With such few papers uncovered after filtration, one can only conclude that there is dire need for more rigor in this field. Researchers aren't focusing enough on the size of the learning effects that their games or applications are having on their users. The applicational value of the products made are not measured, and future work should address this.

Future work should also put effort toward consistency, both in methodology and results. We need to know what aspects of games or gamified applications work to the improvement or indeed detriment of actual learning in our users. I believe standardising how we report our results, and in this consistently report on the size of the effects we're observing, is a good first step. Another will be to measure the learning gains of users more fine-grained, to be able to improve the results of the applications we create by letting the data guide future designs.

Another way to approach consistency may be to look elsewhere for well-established procedures to use as mechanics underneath the gamified experiences of gamification and serious games. One such example is DI. It is doubtful that this has been done previously to any extent. Originally, this review was intended to look at not only the effects of serious games and gamification, but also the effects of teaching methods from applied behavior analysis (ABA), like DI and PT, used by Morningside and others, and any crossover from that literature into the literature on gamification and serious games; performing a three-pronged search. However, after discovering a systematic error too late for the time allotted for this review, this was discontinued in favor of looking at gamification and serious games separately. Future work should in my opinion look for this crossover, or indeed attempt to introduce it.

VI. Conclusion

The results seem to indicate that serious games may be more effective for learning than gamification efforts. However, the variability of results are immense, and the results of these 14 papers are not enough to speak for the general applicability of these approaches in the whole world. The truth is that, although the effects seem to often be positive, and often highly so, we still don't know enough of how these games and applications work on the users to make them learn. At least not isolated to the serious games and gamification literature.

We do not yet know enough, and our measurements are evidently not detailed enough to know *why* the approaches work better or worse. Yes, one could take closer looks at results of affect, but I stand with Freire et al [12] in that education should become more data-driven through data extracted from the games or applications themselves, and with Skinner [2] in that one doesn't need to refer to mental processes or emotional states to explain and control human behavior and learning. My chief proposal is thus that future work in gamification and serious games should consciously implement behavior-analytical techniques, like Direct Instruction and Precision Teaching to not only attempt to get the variability down, but also to have more fine-grained data to drive development. I believe the future of education contains digital gamified tools for the teacher, that are evidence-based beyond questionnaires of affect and usability, or simple pre- and post-tests with low p-values and no mention of effect size. That future, from the results reviewed here, seems far away still. But the proposed metrics, Learning Gain and Learning Gain Improvement, are at least a promising first step to start comparing apples to apples.

References

[1] E. A. Boyle, T. Hainey, T. M. Connolly, G. Gray, J. Earp, M. Ott, T. Lim, M. Ninaus, C. Ribeiro, and J. Pereira, "An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games," *Computers & Education*, vol. 94, pp. 178–192, 2016.

[2] B. F. Skinner, *Science and human behavior*. Simon and Schuster, 1965, no. 92904.

[3] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & education*, vol. 59, no. 2, pp. 661–686, 2012.

[4] R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer, *Serious games*. Springer, 2016.

[5] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining" gamification"," in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 2011, pp. 9–15.

[6] J. Cohen, *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, 1988.

[7] M. Helbæk, *Statistikk. Kort og godt*, 3rd ed. Oslo: Universitetsforlaget, 2011.

[8] W. P. Vogt, "Ceiling Effect," in *Dictionary of Statistics & Methodology*, 3rd ed. SAGE Publications, 2005.

[9] S. S. Sawilowsky, "New effect size rules of thumb," *Journal of modern applied statistical methods*, vol. 8, no. 2, pp. 597–599, 2009.

[10] L. S. Vygotsky and M. Cole, *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

[11] A. Bandura, "Self-efficacy: toward a unifying theory of behavioral change." *Psychological review*, vol. 84, no. 2, p. 191, 1977.

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

[13] K. Johnson and E. M. Street, "From the laboratory to the field and back again: Morningside academy's 32 years of improving students' academic performance." *The Behavior Analyst Today*, vol. 13, no. 1, p. 20, 2012.

[14] J. Stockard, T. W. Wood, C. Coughlin, and C. Rasplica Khoury, "The effectiveness of direct instruction curricula: A meta-analysis of a half century of research," *Review of Educational Research*, vol. 88, no. 4, pp. 479–507, 2018.

Reviewed papers

[15] F. Grivokostopoulou, I. Perikos, and I. Hatzilygeroudis, "An innovative educational environment based on virtual reality and gamification for

learning search algorithms," in *2016 IEEE Eighth International Conference on Technology for Education (T4E)*. IEEE, 2016, pp. 110–115.

[16] C. Barros, A. A. Carvalho, and A. Salgueiro, "The effect of the serious game tempoly on learning arithmetic polynomial operations," *Education and Information Technologies*, vol. 25, no. 3, pp. 1497–1509, 2020.

[17] M. B. Garcia and T. F. Revano, "Assessing the role of python programming gamified course on students' knowledge, skills performance, attitude, and self-efficacy," in *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. IEEE, 2021, pp. 1–5.

[18] B. Manero, J. Torrente, Á. Serrano, I. Martínez-Ortiz, and B. Fernández-Manjón, "Can educational video games increase high school students' interest in theatre?" *Computers & education*, vol. 87, pp. 182–191, 2015.

[19] K. Jodoi, N. Takenaka, S. Uchida, S. Nakagawa, and N. Inoue, "Developing an active-learning app to improve critical thinking: item selection and gamification effects," *Heliyon*, vol. 7, no. 11, p. e08256, 2021.

[20] A. J. Q. Tan, C. C. S. Lee, P. Y. Lin, S. Cooper, L. S. T. Lau, W. L. Chua, and S. Y. Liaw, "Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: A randomized controlled trial," *Nurse education today*, vol. 55, pp. 38–44, 2017.

[21] L.-B. Jaunay, P. Zerr, L. Peguin, L. Renouard, A.-S. Ivanoff, H. Picard, J. Griffith, O. Chassany, M. Duracinsky *et al.*, "Development and evaluation of a new serious game for continuing medical education of general practitioners (hygie): double-blinded randomized controlled trial," *Journal of medical Internet research*, vol. 21, no. 11, p. e12669, 2019.

[22] J. R. Rachels and A. J. Rockinson-Szapkiw, "The effects of a mobile gamification app on elementary students' spanish achievement and self-efficacy," *Computer Assisted Language Learning*, vol. 31, no. 1-2, pp. 72–89, 2018.

[23] S. Wichadee and F. Pattanapichet, "Enhancement of performance and motivation through application of digital games in an english language class," *Teaching English with Technology*, vol. 18, no. 1, pp. 77–92, 2018.

[24] B. Hazan, W. Zhang, E. Olcum, R. Bergdoll, E. Grandoit, F. Mandelbaum, G. Wilson-Doenges, and L. A. Rabin, "Gamification of an undergraduate psychology statistics lab: Benefits to perceived competence," *Statistics Education Research Journal*, vol. 17, no. 2, pp. 255–265, 2018.

[25] B. J. Morris, C. Dragovich, R. Todaro, S. Balci, and E. Dalton, "Comparing badges and learning goals in low-and high-stakes learning contexts," *Journal of Computing in Higher Education*, vol. 31, no. 3, pp. 573–603, 2019.

[26] T. Su, M.-T. Cheng, and S.-H. Lin, "Investigating the effectiveness of an educational card game for learning how human immunology is regulated," *CBE—Life Sciences Education*, vol. 13, no. 3, pp. 504–515, 2014.

[27] R. Baños, A. Cebolla, E. Oliver, M. Alcañiz, and C. Botella, "Efficacy and acceptability of an internet platform to improve the learning of nutritional knowledge in children: the etiobe mates," *Health education research*, vol. 28, no. 2, pp. 234–248, 2013.

[28] N. Charlier and B. De Fraine, "Game-based learning as a vehicle to teach first aid content: a randomized experiment," *Journal of school health*, vol. 83, no. 7, pp. 493–499, 2013.

Appendix F

Adaptive learning systems review

The following is an unpublished paper delivered as part of course work during the autumn of 2022. Due to the paper being referenced in the present thesis, and being a significant part of the background for it, this paper has been included as an appendix for the convenience of the reader.

NTNU
Kunnskap for en bedre verden

Department of Computer Science (IDI)

IMT4215

Exploring the empirical topography and quality of evidence for adaptive learning systems: A systematic review

Student Name:
Svein-Kåre Bjørnsen

15th December 2022

Abstract

Adaptive learning systems have been an emerging research trend for at least the past 20 years. These systems foster active learning on part of the learner, and attempt to adapt to the individual user's needs, characteristics and proficiency. This work searches for the empirical evidence on such systems within nearly 1500 papers published between 2015 and 2022. A semi-automatic trawl-and-sift approach was used to filter the results using a computer program by analysing the citation information of the papers, as well as the entire text once downloaded, to a number manageable for manual review.

A few commercial systems were identified to have good empirical evidence for positive impact. The existing systems seem to be heavily influenced by cognitive theories, and there is a dearth of other evidence-based educational approaches. There also seems to be a research opportunity in performing a large-scale meta-analysis on the entire literature corpus for investigating the empirical evidence at large.

1 Introduction

Adaptive learning is an educational principle that is based on keeping the learner in their zone of proximal development continuously [20].

The zone of proximal development is a theoretical construct that reflects the level of learning challenge that is 'just enough' so that the learner is able to do more with help than without, and not get frustrated or bored by the level difficulty [29].

An adaptive learning system (ALS) can be defined as: '[...] a specific platform that provides structured learning activities or sequenced learning paths; or for the purpose of targeting a specific learning population' [23, p. 1920].

For my future master thesis I am creating an adaptive learning system that attempts to use evidence-based educational methods from behavioral science as a foundation for educational gamification, so in this instance I am only interested in concrete digital systems for adaptive learning. The technology itself more than the philosophy. I have already looked into the quality of the evidence on educational gamification and serious games [16], however there seemed to be limited overlap between these fields, so an investigation into adaptive learning systems were in order.

I was unable to find any previous literature reviews that covered the empirical evidence on digital adaptive learning systems to a satisfactory degree, so I embarked on this review, employing a semi-automatic trawl-and-sift approach to cover as much of the literature as possible while still being only a single person.

I searched the literature from the period 2015–2022, and found a few commercial systems that have good empirical support. It was challenging to take a stance on ALSs in general due to the diverse nature of the research uncovered. I also found that cognitive theory is heavily present as foundation for these systems, and future work could benefit from attempting evidence-based methods from behavioral science for contrast. I also conclude that there should be enough of a corpus in the literature at this point to do a meaningful meta-analysis on the empirical evidence of these systems. Something that I have not found so far.

1.1 Related work

Martin et al. [23] investigated the research designs, contexts, strategies and technologies in the period 2009–2018, reviewing 61 papers in total. This review was more of a high-level mapping than a literature review, and the authors didn't pay much heed to the empirical strength of adaptive learning systems. Among other things, they remarked that the greatest concentration of research in terms of discipline lies in computer science. Learning style was the most frequent learner characteristic.

Of the mentioned previous works in the paper, only one asked the effectiveness question. Back in 2008. The rest focused on characteristics and traits of the learner as a target for adaptation. Like personal traits and learning styles.

Kumar et al. [21] looked at papers published between 2001 and 2016, reviewing 78 papers in total, 12 of which were empirical. The study focused on adaptive tutoring systems based on learning styles, as such it was far more specific than what I was looking for, and is also starting to get dated.

Verdu et al. [28] look at the effect sizes of empirical papers in the literature, and reported medium to high effect sizes using Cohen's d for the 15 papers they reviewed. However the review was non-systematic and didn't report how paper selection was done. Additionally it is quite dated at this point.

Nakic et al. [25] looked at papers in the period 2001–2014, reviewing 98 papers in total. The word 'review' is a bit misleading here, since what they actually did was map the literature. Investigating the empirical evidence was not a focus.

Vandewaetere et al. [27] investigated the contribution of learner characteristics in the development of digital ALSs. They wanted to know how and to which degree learner characteristics had been been incorporated into the implementation of ALSs. They noted high variability and a lack of empirical validation, as well as standard frameworks to design experiments. Effect sizes are not commented on, and it is quite dated at this point.

Normadhi et al. [26] reviewed 78 papers from the period 2010–2017, providing an overview of the personal traits of learners and the techniques that have been used to identify them. Effectiveness and effect sizes was not a concern.

Akbulut and Cardak [14] looked at publications

in the period 2000–2011, specifically adaptive hypermedia that accommodated learning styles. They reviewed in total 70 studies. This review was non-systematic, and includes less than ideal details on the procedure that the authors followed. It is also quite specific in looking only at hypermedia systems and learning styles. In addition is is also quite dated.

1.2 Problem description

There is a lack of recent literature reviews that look at the empirical effectiveness of adaptive learning systems in general. There is also no crossover into the fields of educational gamification and serious games that I am aware of, even though ALSs seem to be quite similar to gamified systems.

Thus I formulate the following research questions:

1. What are the current empirically supported digital adaptive learning systems in the literature?

2. To what degree are current digital adaptive learning systems evidence-based, and what is the quality of the evidence?

1.3 Learning Gain and Learning Gain Improvement

Adaptive learning systems have much in common with educational gamification, and depending on the author, educational gamification and serious games may be considered as adaptive learning systems themselves [23]. In an earlier work [16], I reviewed the empirical literature on educational gamification and serious games. I had found that the controlled pretest-posttest quasi-experimental design was common, but that effect size reporting was scarce. To have a simple measure of effect size that I could quickly calculate to compare approaches, even with limited data. I formulated two intuitive percent-based metrics, that I called Learning Gain (LG) and Learning Gain Improvement (LGI). They are not without flaws, but in most cases give a picture of the experimental results that is very easy to understand. For this review, I found it of interest to calculate LGI for any eligible studies that I found to contrast them to my previous findings. Their formulas are given below.

$$LG = \frac{post\ mean - pre\ mean}{pre\ mean} \tag{1}$$

$$LGI = \frac{LG_{experiment} - LG_{control}}{LG_{control}} \tag{2}$$

2 Methodology

I did several iterations of test searches trying out different search strings, and judging the first few pages of results on different search engines. I looked for a high number of results, and for the first couple of pages to show results that seemed relevant. The final search terms were as follows:

('adaptive learning system' AND ('efficacy' OR 'effectiveness' OR 'empirical'))

The following databases were chosen for this search. From previous experience I knew that all of them would give me citation information for the results in bulk, at least all results on each page in one download. They also aligned thematically with the current domain.

- Web of Science
- ScienceDirect
- IEEE Xplore
- Scopus
- ERIC
- Emerald

The inclusion criteria for this review were as follows:

- Research papers, including conference papers, but not book chapters
- Published in the period 2015–2022
- English language

Exclusion criteria:

- Papers not related to adaptive learning systems at all
- Literature reviews
- Papers concerning machine learning unrelated to adaptive learning systems
- Incomplete works, for example parts of dissertations or protocols
- Non-empirical research

- Papers with a quality score of 4 or less

For the quality assessment of the papers, I opted for and adapted version of the *QualityScore* of Bertolino et al. [15]. Using an aggregated quality criteria score to sort my results by quality before picking the best papers. This seemed like a good way to somewhat check my personal bias, even though it would be better to do this with more people.

The original score [15] was the sum of the assessment of a paper with five different criteria formulated as yes/no questions, each valued between 0.0–1.0. A criterion score of 1 meant 'yes', a criterion score of 0 meant 'no' and a criterion score of 0.5 meant 'partial'. The authors did not specify if they allowed criterion scores between 0–0.5 or 0.5–1.0.

However, I found the original criteria a bit loose, so I adapted them to something more concrete, and I also decided that I would allow giving in-between scores, like 0.2 or 0.6. The following are the quality criteria I used, giving a minimum quality score of 0, and a maximum of 7:

1. Is the focus on adaptive learning systems clearly defined?
2. Is the problem of the study clearly defined?
3. Is the contribution of the study clearly defined?
4. Is the study design sound and valid?
5. Are the results clearly communicated?
6. Is the data analysis sound in terms of methodology?
7. Are limitations and future directions clearly stated?

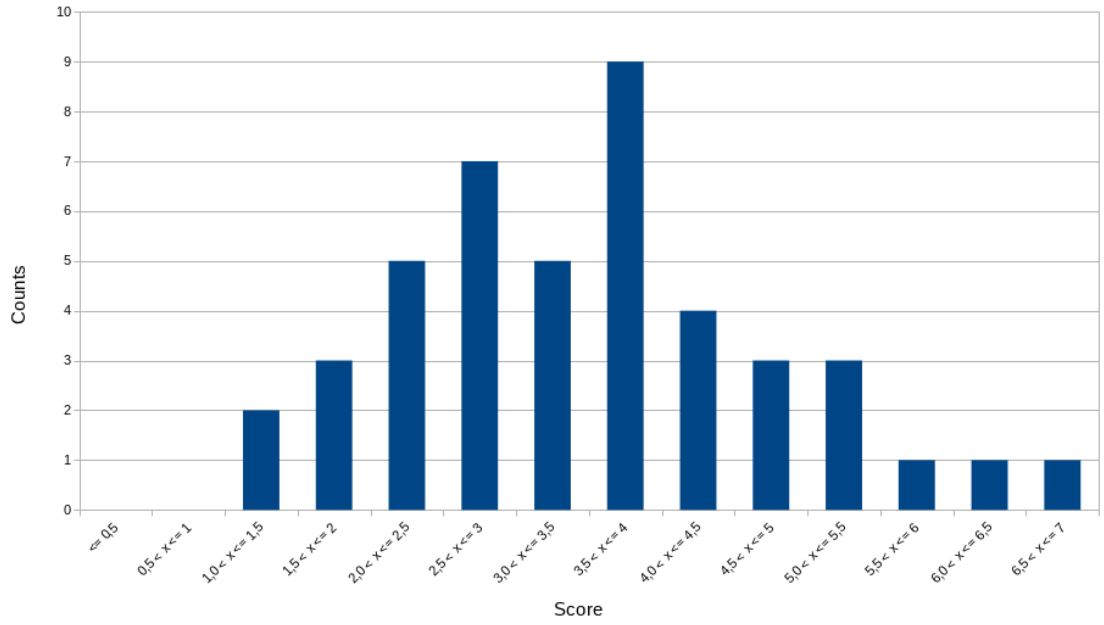
As a guiding note on the scoring of criterion 4 in particular, I set examples of the scores 0 and 1 to help me in the assessment:

0.0 = Empirical study without control or other grave deficit

1.0 = Randomised controlled trial (RCT) with sizable population

To save time, I only partially scored many of the papers that had such a low score that they would not make the cut for my review even if the last criterion got a full score. Figure 1 shows the final plotted distribution, with an assumption that the last criterion of the still partially scored papers—which is still 14 papers in the lower half of the distribution—would be 0.5.

Figure 1: Quality score paper distribution



2.1 Procedure

The inclusion and exclusion criteria were codified into several filters, some of which were automatic, meaning they were done by a tailored python script. These are detailed further below.

Checkboxes on the search engines were used where possible to roughly narrow down the scope of inclusion. Then the citation information for all results across the different databases was downloaded in bulk and converted to a single CSV file for further processing. The entire search and download of citation information was done on 30th September 2022. The total number of search results before removing duplicates was over 1600.

After the citation information had been compiled to a single file, the following filters were applied to it with a python script:

- Remove entries where one of the following fields were blank: *title, abstract, authors, year*
- Remove duplicate entries by comparing the title fields converted to lowercase for equality
- Remove entries where the number in the year field is lower than 2015
- Remove entries that do not contain the phrase 'adaptive learning system' in either keywords or title
- Remove entries that do not have a *type* field that correspond to a conference paper or journal article.

The resulting filtered CSV file was loaded into a spreadsheet and further processed manually. First doing a deduplication pass on titles to catch any cases where the program was unable to remove the duplicate. If a conference paper had same title as a journal article, only journal article was included. After this, the list still contained papers that were highly irrelevant, for example relating to neurological research or marketing research, or papers considering adaptive learning algorithms within machine learning. Therefore I did a manual filtering for relevance by judging the titles, marking any that I was in doubt on; then I judged the doubtful papers by their abstract. The papers that were obviously irrelevant from title or abstract were discarded, and any remaining papers that were still doubtful after reading their abstract were also excluded. After this manual filtering, the remaining papers were downloaded as PDFs for full-text filtering, quality evaluation and reading.

The downloaded papers were first manually screened for papers that were literature reviews, then two programmed filters were run on the collection of PDFs to remove irrelevant papers. First one that excluded any paper that in its full text did not contain the word 'education', to hopefully eliminate any papers that didn't relate to any form of educational use of an adaptive learning system. The second did the same, but excluded any paper that didn't contain either of the terms 'pre-test', 'pretest', 'post-test', 'posttest', 'pre-' or 'post-'. This was to remove non-empirical papers. Granted, this also excludes any papers that employ a design that doesn't use either a pretest or a posttest. My reasoning was that such papers would most likely be excluded by my subsequent quality assessment anyway, so there should be no harm in them disappearing at this stage.

After this, the papers were manually scanned in two passes to first eliminate incomplete works, and then any remaining non-empirical papers.

The collection of remaining papers was then assessed for quality.

While performing the quality evaluation, some of the papers turned out to still be completely missing the mark in terms of relevance, resulting in them being excluded at this stage.

After eliminating the lower-quality papers for the final selection, the papers were read for comprehension. I looked for common themes and patterns related to my research questions, as well as extracting learning gain metrics where applicable.

3 Results

The number of results before filtering was 1 604. After completing all filtration steps, including quality assessment and cutoff, the resulting number of papers for final review was 13. The details of how the different steps affected the number of papers can be seen in table 1. 13 papers were excluded while downloading due to lack of access.

Table 1: Filtration of search results

Step	Removed	New total
Blank fields	12	1592
Deduplication	120	1472
Year filter	407	1065
Main exclusion filter	887	178
Manual deduplication	4	174
Manual, relevance	37	137
Download	13	124
Manual, literature review	13	111
Scripted, full-text 'education'	14	97
Scripted, full-text 'pre-' & 'post-'	34	63
Manual, incomplete works	1	62
Manual, not empirical	12	50
Manual, quality screening, irrelevant	6	44
Manual, quality score cutoff	31	13
	Final total:	**13**

After filtration and quality assessment, the papers where read for comprehension, and I attempted to establish types of adaptive learning systems, applied learning theory. How the systems were adapted and what they adapted, as well as other possible differentiators there were. These findings are summarised in table 2.

While reading the papers, I iteratively expanded my lists of categories that the papers fell into in relation to different aspects mentioned above. These were numerically coded to condense the table I used while working, and are preserved in table 2. The codes are given below.

Type of paper:

1. Evaluations of ALSs vs other ALS
2. Evaluations of ALSs vs traditional or 'business as usual'
3. Evaluations of methodology for ALS
4. Evaluations of algorithm or sub-function for ALS
5. Other

Theoretical foundation:

1. Mastery learning (based on Bloom's taxonomy [17])
2. Learning styles
3. Zone of proximal development
4. Not clear
5. Spaced repetition
6. Other

Adaptive focus, what was adapted:

1. Learning style
2. Difficulty
3. Topic path
4. Content by proficiency
5. Unknown

Furthermore there were also two main types of approaches to adaptivity. The users could on one side personally adapt the system manually, or the system could do it automatically. This also somewhat relates to the type of dynamic of the system, if the system is manually adapted, it has been mentioned as an 'adaptable' system, whereas if it only adapts automatically, it has been called an 'adaptive' system [3]. I also found that some cases combined these types of dynamics.

Table 2: Details of the reviewed papers

Ref.	Type	Theory	Design	Focus	Year	Approach	Dynamic
[3]	3	4	other	3	2015	negotiation	combination
[4]	3	4	quasi-expr.	4	2020	system	adaptive
[9]	2	5,6	within-subj.	4	2019	system	adaptive
[10]	2	2,6	quasi-expr.	1,4	2019	user and system	combination
[11]	3	4	within-subj.	3	2020	system	adaptive
[5]	3	2	RCT	1	2022	system	adaptable
[1]	2	4	quasi-expr.	3	2020	system	adaptive
[13]	2	4	RCT	5	2020	unclear	unclear
[2]	1,2	multiple	quasi-expr.	multiple	2016	multiple	multiple
[7]	3,4	6	RCT	3	2020	system	adaptive
[6]	5	6	other	3	2017	system	adaptive
[8]	3	6	RCT	4	2018	system	adaptive
[12]	3,4	6	other	4	2021	system	adaptive

The relative prevalence of different study designs is shown in figure 2, The quasi-experimental design was tied with the pretest-posttest controlled experimental design—also known as randomised controlled trial (RCT)—covering over half the papers reviewed. It is however important to note here that this prevalence is most likely due to my aggressive filtering and picking of only the top section of the papers I got in terms of quality, and so should definitely not be taken as a sign of the field at large. However, it does show that these sorts of designs can easily be found in the literature. The 'other' category is for any design that could not fairly be put into any category found in a regular scientific methodology text book [22]. These studies were also so heterogeneous that finding a common label was infeasible.

Table 3 shows the extracted learning gain metrics for those papers that these could be calculated. The data for Siddique et al. [10] has been averaged, since they ran 12 different group pairs, each with a different combination of prior knowledge, working memory capacity and learning style. The detailed data on this specific study is listed in table 4. Standard deviation is for the learning gain metrics listed as percentage points, noted as *pp* for clarity. In regards to the calculated means, there are a few caveats. Nye et al. [8] did not report raw mean values, but only the raw learning gain, so the learning gain metrics for experiment and control for this paper are excluded from the mean and standard deviation calculation. The learning gain improvement calculation is insensitive to the unit of the learning gains, so it could still be calculated, and *is* included in the calculations. Wang et al. [13] found a *decrease* in the learning gain of the control group in one of their two experiments, the formula for LGI does not handle negative numbers for the control group results, so this calculation is not applicable, and is excluded from mean and SD calculation. Another thing to note is that Chou et al. [3] did two experiments, first a pilot and then a full study on different cohorts of students, they used the same control cohort as control. I included the pilot in table 3, since it mostly used the same method as the full experiment and had a fairly large number of participants.

Table 3: Learning gain improvement extracted from eligible papers

Paper	Grouping	N exp	N con	LG exp	LG con	LGI
[10]	**mean**	92	92	70.32%	38.96%	120.95%
[1]	pilot	48	112	17.61%	14.45%	21.90%
	full	183	112	23.74%	14.45%	64.31%
[13]	vs large class	87	68	15.64%	2.48%	531.66%
	vs small group	48	36	8.57%	-0.57%	**n/a**
[7]	–	25	25	110.13%	83.44%	31.98%
[8]	–	28	48	**31.70**	**28.40**	11.62%
	Mean	73.00	70.43	47.49%	30.76%	130.40%
	SD	55.13	35.78	40.44pp	31.60pp	200.51pp

Table 4: Learning gain improvement specification for [10]

Grouping	N exp	N con	LG exp	LG con	LGI
1*	8	8	91.11%	53.12%	71.50%
2*	8	8	116.31%	71.17%	63.43%
3*	8	8	101.72%	62.14%	63.70%
4	8	8	117.76%	75.97%	55.01%
5	8	8	117.76%	76.00%	54.95%
6	8	8	103.95%	44.74%	132.35%
7	8	8	40.20%	13.66%	194.26%
8	8	8	35.38%	14.05%	151.92%
9	8	8	41.20%	11.07%	272.30%
10	8	8	30.99%	23.01%	34.63%
11	8	8	24.63%	17.10%	44.02%
12*	4	4	22.79%	5.51%	313.33%
		Mean	70.32%	38.96%	120.95%
		SD	40.49pp	27.70pp	94.14pp

*: $p < 0.05$

In a similar review on the literature on educational gamification and serious games that I did previously, these metrics were also extracted [16]. There were no overlapping papers between that review and this one. Table 5 shows a comparison of the mean value for LGI between the three types of applications, using the data from the previous review. As can be seen, the mean for adaptive learning systems, as measured by the present review, is lower than both gamification and serious games, serious games coming out on top. But this is based on relatively low numbers, and the metric isn't statistically sound, especially for comparison, so this should be regarded as an interesting observation that would need proper study to validate.

Table 5: ALS LGI compared with data on gamification and serious games

Measure	Gamification	Serious games	ALS
LGI mean	150.70%	164.91%	130.40%
LGI SD	260.60pp	299.58pp	200.51pp

Brasiel et al. [2] report on a large-scale study evaluating 11 different ALSs for mathematics instruction within the K-12 segment of education in the US. The purpose was to measure the impact of the supplemental use of these systems on student proficiency on standard state-issued tests; as well as discover common themes reported by teachers on their implementation in classrooms.

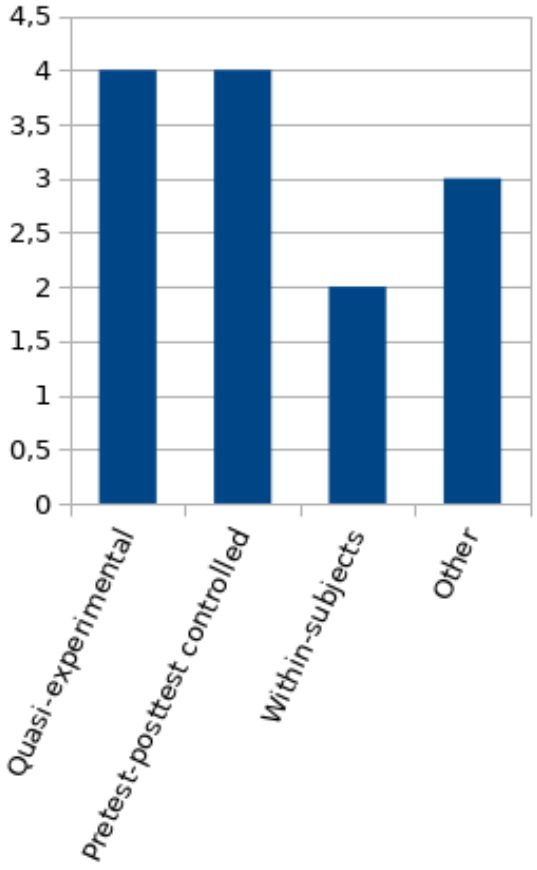
Of the 11 systems that were part of the study, only 6 were part of the final impact evaluation. These were: *ALEKS*, *Catchup Math*, *i-Ready*, *MathXL*, *ST Math* and *Think Through Math*. All of these showed a considerable impact on at least some level, all of them beating the baseline effect size in at least one of their samples. Several of the systems where tested on both a full sample and a fidelity sample, and in these cases, only the fidelity sample beat the baseline effect size. However, only *ALEKS* and *i-Ready* also had a significance level of < 0.05.

Siddique et al. [10] measured several dependent variables against the independent variables (1) prior knowledge, (2) working memory capacity and (3) learning style. They looked for effects on (1) learning outcome, measured through pre- and posttest; (2) learning efficiency, measured by activity completion time; (3) retention, measured as difference between first and second posttest; and (4) satisfaction, measured with a likert-scaled survey. Experimental and control groups were sub-grouped according to combination of prior knowledge level, working memory capacity and learning style, and results were analysed separately. Of the 12 groupings, only 4 showed statistically significant results ($p < 0.05$). These are marked with * in table 4. However, most of the rest had a p-value < 0.1.

Standen et al. [11] investigated the effect of including affective state measured by different behavioral measurements through interactions with the system, including posture, facial expressions etc. in an adaptive learning system

for teaching students with intellectual disabilities. It was a within-subjects experiment, where the students alternated using the system with and without the affective part. Achievement was a factor in both experiment and control phases. They found statistically significant positive correlations between time spent engaged or frustrated with an increase in achievement. This increase was stronger with engagement than with frustration. Boredom had a similar statistically significant negative correlation as engagement, time spent bored reduced achievement. Their ALS, *MaTHiSiS*, had statistically significant positive effects on engagement and achievement, with effect sizes ranging from 0.4–0.8 (Cohen's d).

Figure 2: Prevalence of different study designs



Alwadei et al. [1] investigated the hypothesis that student achievement when using an adaptive learning system should be at least as good as that of face-to-face instruction. The authors performed two convenience sampled experiments, one formative pilot and one summative experiment afterward. In both cases the intervention group scored significantly higher on the final exam compared to the same control. The ALS is not described in enough detail to reproduce the study, but it was put together by an instructor through some sort of adaptive learning platform, however which product this was was not disclosed. Effect sizes were not reported.

Wang et al. [13] This study investigated the efficacy of *SquirrelAI*, and adaptive learning system that has been deployed on a large scale in China for some time. They ran two separate experiments, and 'substantial' improvement effects were found for the intervention groups in both, while not for the control groups. The effect size compared to control was not reported for either experiment.

Niknam and Thulasiraman [7] addressed curriculum sequencing by creating a learning path recommendation system that used clustering to group learners by prior knowledge for selection of a proper learning path per group. The authors used an ant colony optimisation algorithm to search for suitable learning paths continuously as the learners improve. The effectiveness of the system was evaluated through a database course offered to students. The experiment took the form of a randomly controlled experiment with a pretest and using the final exam scores as the posttest.

The theoretical foundation for this learning system was *meaningful learning theory* which is part of early cognitivism, but has apparently been developed further more recently.

They found a statistically significant learning gain difference between experiment and control, with an effect size of $d = 0.58$ (Cohen's d).

Liu et al. [6] used the commercial adaptive learning system *LeaP* to analyse student usage patterns while interacting with it. The aim of the study was to investigate the patterns, see if and how the patterns differ between low- and high-performing students, as well as what kinds of relationships existed between the discovered patterns. The participants were first-year students in a pharmacological study program. The experiment utilized pre- and posttests but was exploratory in nature, looking for correlations, and there was no control group. They found that students that had higher prior knowledge made more attempts at testing themselves within the system before the posttest, while also attaining higher end scores than others. The authors note that while the latter is expected, the former is surprising, and may indicate an inherent mastery goal orientation with those students, since they seemed more motivated. They did correl-

ation and regression analyses as well, and they suggested that the more students took a test, the higher their final score was. The authors attribute this to the students actively retrieving the learned information and became more familiar with the makeup of the tests.

Nye et al. [8] investigated the learning outcomes and user perceptions from combining *AutoTutor*, a conversational tutoring system, with *ALEKS*, an adaptive learning system for mathematics, into a hybrid intelligent tutoring system. The system talked students through step-by-step examples that should help them solve algebra problems outside the system. The theoretical foundation for the system was *worked examples*, *self-explanation* and *impasse-driven learning*.

The system was evaluated using a randomly controlled experiment with pre- and posttests, as well as user perception surveys both before and after.

Results were mixed, both for the learning outcomes and the user perceptions, and the strongest influence on learning across users was the time spent studying.

The study of van der Velde et al. [12] attempted to mitigate the 'cold start' problem of ALSs by employing different strategies. The cold start problem refers to the initial unadapted state of an adaptive learning system before it has enough data to dial in its adaptation to the new user.

Earlier work had shown promise for three different strategies inspired by recommender systems, however these had used post hoc simulations instead of experimentation to validate their approaches. The mitigation strategies in question were: making assumptions on level of prior knowledge from the first interactions with the system; clustering learners into similar levels of skill based on previously collected data; and using background information like grade-level, gender etc. to group them.

They conducted two experiments, using a fact-learning system, the first one to compare mitigation strategies, and the second to introduce more variability in the difficulty of the learned facts.

They found that addressing the cold start problem with mitigation strategies improved learning as long as there was enough variability in the difficulty of the material. They reported specific confidence intervals, but no p-values.

4 Discussion

The studies on ALSs seems to often be quite specific, and yield results that may not be very generalisable to ALSs in general. Wang et al. [13] may lead us to believe that ALSs are less effective for small groups than larger classes, but this was *one type* of ALS, and results shouldn't be extrapolated in a wider sense. Future studies should investigate the how effect sizes differ across different types of ALS with the same subject matter and population of participants.

I also notice an emphasis on cognitive theory as the basis for the design of ALSs. I have yet to see a single instance of an ALS that incorporates evidence-based educational methodology from behavior science [24], neither during the present review nor in the related works. This is an echo of the state of the art within educational gamification and serious games as welli [16]. Even though some of those methods are suitable for digitalisation. This is a clear gap in the literature on adaptive learning systems.

In relation to research question 1, which current digital ALSs in the literature are empirically supported, I end up with the following list: *ALEKS* and *i-Ready*, due to their good results in the large-scale US-based study [2], *SquirrelAI* and *MaTHiSiS* for their statistically significant effect sizes. Other systems that have strong evidence also surfaced in this review, but these were not named, and thus are probably not something directly available for other researchers or educators.

Research question 2 asked to what degree current digital ALSs are evidence-based, and what the quality of the evidence is. From the results of the present review, I must conclude that the evidence for ALSs is generally not very good. But this is not because well evidenced ALSs do not exist, but rather that the researchers are often creating their own instead of using an existing one, in addition to many studies investigating specific methods or approaches for use in future ALSs. Instead of an ALS being treated as a unit, it is treated as a divisible container that can house many different methodologies, philosophies and learning theories to help the learner learn. It is impossible to take a stance on ALSs as a whole from the state of the literature now, as it is reflected by this review.

However, this review has obvious limitations. Granted, this is true for any literature review that relies on a human to make value-judgements on the search, inclusion and exclu-

sion of publications; but they must nevertheless be covered, and in this case there are a couple of special ones.

4.1 Limitations and future work

A cautionary note on table 5 is in order. Properly statistically analyzing the results from the research on gamification versus serious games versus adaptive learning systems, in the form of a large scale meta-study, may not be possible yet, due to a quite heterogeneous research corpus [18, 19], though I haven't seen any recent large-scale literature reviews that have tried to aggregate data on the former two domains since 2016 [18]. Suffice it to say that the differences observed in table 5 may not be real. There is a gap for large-scale comparison here, but the scope is larger than possible for the present review.

This review also has more general methodological issues. This review was conducted by myself alone, and I used a subjective quality scoring system to exclude low-quality papers. Doing this injects bias into the selection process, and traditional systematic literature reviews often offset this by using multiple people to assess the papers for quality and note the rate of agreement. This was not an option here. However, much of the relevance filtration was codified into computer scripts looking for the presence of specific terms in the papers themselves, and I regard this as somewhat offsetting my bias in this instance, since it at least is non-volatile and deterministic. A value-judgement on whether or not a paper fulfills a more abstract criterion is not deterministic and is quite possibly volatile even keeping the reviewer the same across papers and time. In addition, the quality distribution seen in figure 1 is smooth, and has no substantial gaps where one could wonder where all the papers scoring X have gone. This distribution not having obvious holes can be seen as a sign that my sample of papers, if not entirely representative, at least doesn't have obvious deficits.

Which brings me to my last point. The sample of papers is rather small. If having been more lenient in my filtration, I might have been able to uncover more high-quality papers, but that would also include more work in assessing the quality of the papers, and the constraints on this review in context of course work did not allow for a deeper dive without the inclusion of more people.

Future reviews can take into account a larger portion of the literature, given more manpower and time. I think there may be a possibility of a meaningful large-scale meta-analysis on the empirical effect sizes of ALSs at this point in time, if one takes the entire corpus of research under that lens. We have papers from at least two decades now, but I am not aware of any such undertaking having been attempted.

Future ALS implementations should try to incorporate evidence-based methods from behavioral science, as these may fit well in a digital form, and have not been tried at all yet in this domain as far as I have been able to tell.

5 Conclusion

In this review I have scraped the literature on adaptive learning systems in the period 2015–2022, in the search for the empirical support for these systems. A few concrete commercial systems have in this period been found to have considerable effects, however the strength of the evidence can still improve. The theoretical foundation for these systems is heavily rooted in cognitivism, and there could exist benefits to in new research attempt to use evidence-based educational methods from behavioral science as a basis instead. There seems to be a gap in the literature, and data support for, a large-scale meta-analysis of the empirical evidence on adaptive learning systems. Both of the latter points are real possibilities for future work.

Reviewed papers

[1] A. H. Alwadei, A. S. Tekian, B. P. Brown, F. H. Alwadei, Y. S. Park, S. H. Alwadei, and I. B. Harris. Effectiveness of an adaptive elearning intervention on dental students' learning in comparison to traditional instruction. *Journal of Dental Education*, 84(11):1294–1302, 2020.

[2] S. Brasiel, S. Jeong, C. Ames, K. Lawanto, M. Yuan, and T. Martin. Effects of educational technology on mathematics achievement for k-12 students in utah. *Journal of Online Learning Research*, 2(3):205–226, 2016.

[3] C.-Y. Chou, K. R. Lai, P.-Y. Chao, C. H. Lan, and T.-H. Chen. Negotiation based adaptive learning sequences: Combining

adaptivity and adaptability. *Computers & Education*, 88:215–226, 2015.

[4] G.-J. Hwang, H.-Y. Sung, S.-C. Chang, and X.-C. Huang. A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, 1:100003, 2020.

[5] Y. Lin, S. Wang, and Y. Lan. The study of virtual reality adaptive learning method based on learning style model. *Computer Applications in Engineering Education*, 30 (2):396–414, 2022.

[6] M. Liu, J. Kang, W. Zou, H. Lee, Z. Pan, and S. Corliss. Using data to understand how to better design adaptive learning. *Technology, Knowledge and Learning*, 22 (3):271–298, 2017.

[7] M. Niknam and P. Thulasiraman. Lpr: A bio-inspired intelligent learning path recommendation system based on meaningful learning theory. *Education and Information Technologies*, 25(5):3797–3819, 2020.

[8] B. D. Nye, P. I. Pavlik, A. Windsor, A. M. Olney, M. Hajeer, and X. Hu. Skope-it (shareable knowledge objects as portable intelligent tutors): overlaying natural language tutoring on an adaptive learning system for mathematics. *International journal of STEM education*, 5(1):1–20, 2018.

[9] S. Ruan, L. Jiang, J. Xu, B. J.-K. Tham, Z. Qiu, Y. Zhu, E. L. Murnane, E. Brunskill, and J. A. Landay. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[10] A. Siddique, Q. S. Durrani, and H. A. Naqvi. Developing adaptive e-learning environment using cognitive and noncognitive parameters. *Journal of Educational Computing Research*, 57(4):811–845, 2019.

[11] P. J. Standen, D. J. Brown, M. Taheri, M. J. Galvez Trigo, H. Boulton, A. Burton, M. J. Hallewell, J. G. Lathe, N. Shopland, M. A. Blanco Gonzalez, et al. An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *British Journal of Educational Technology*, 51 (5):1748–1765, 2020.

[12] M. van der Velde, F. Sense, J. Borst, and H. van Rijn. Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates. *Computational Brain & Behavior*, 4(2):231–249, 2021.

[13] S. Wang, C. Christensen, W. Cui, R. Tong, L. Yarnall, L. Shear, and M. Feng. When adaptive learning is effective learning: comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, pages 1–11, 2020.

References

[14] Y. Akbulut and C. S. Cardak. Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. *Computers & Education*, 58(2):835–842, 2012.

[15] A. Bertolino, G. D. Angelis, M. Gallego, B. García, F. Gortázar, F. Lonetti, and E. Marchetti. A systematic review on cloud testing. *ACM Computing Surveys (CSUR)*, 52(5):1–42, 2019.

[16] S.-K. Bjørnsen. Evaluating the learning gain improvements of serious games and gamification: A systematic literature review. 2022. [An unpublished paper delivered as part of course work in IMT4307, spring 2022].

[17] B. S. Bloom, M. D. Engelhart, E. Furst, W. H. Hill, and D. R. Krathwohl. Handbook i: cognitive domain. *New York: David McKay*, 1956.

[18] E. A. Boyle, T. Hainey, T. M. Connolly, G. Gray, J. Earp, M. Ott, T. Lim, M. Ninaus, C. Ribeiro, and J. Pereira. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*, 94:178–192, 2016.

[19] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2):661–686, 2012.

[20] P. Karkazis, H. C. Leligou, P. Trakadas, N. Vretos, S. Asteriadis, P. Daras, and P. Standen. Technologies facilitating smart pedagogy. In *Didactics of smart pedagogy*, pages 433–451. Springer, 2019.

[21] A. Kumar, N. Singh, and N. J. Ahuja. Learning styles based adaptive intelligent tutoring systems: Document analysis of articles published between 2001. and 2016. *International Journal of Cognitive Research in Science, Engineering and Education*, 5 (2):83, 2017.

[22] P. D. Leedy and J. E. Ormod. *Practical Research: Planning and Design*. Pearson Education Ltd, 12 edition, 2021. Global edition.

[23] F. Martin, Y. Chen, R. L. Moore, and C. D. Westine. Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, 68(4):1903–1929, 2020.

[24] D. J. Moran and R. W. Malott. *Evidence-based educational methods*. Elsevier, 2004.

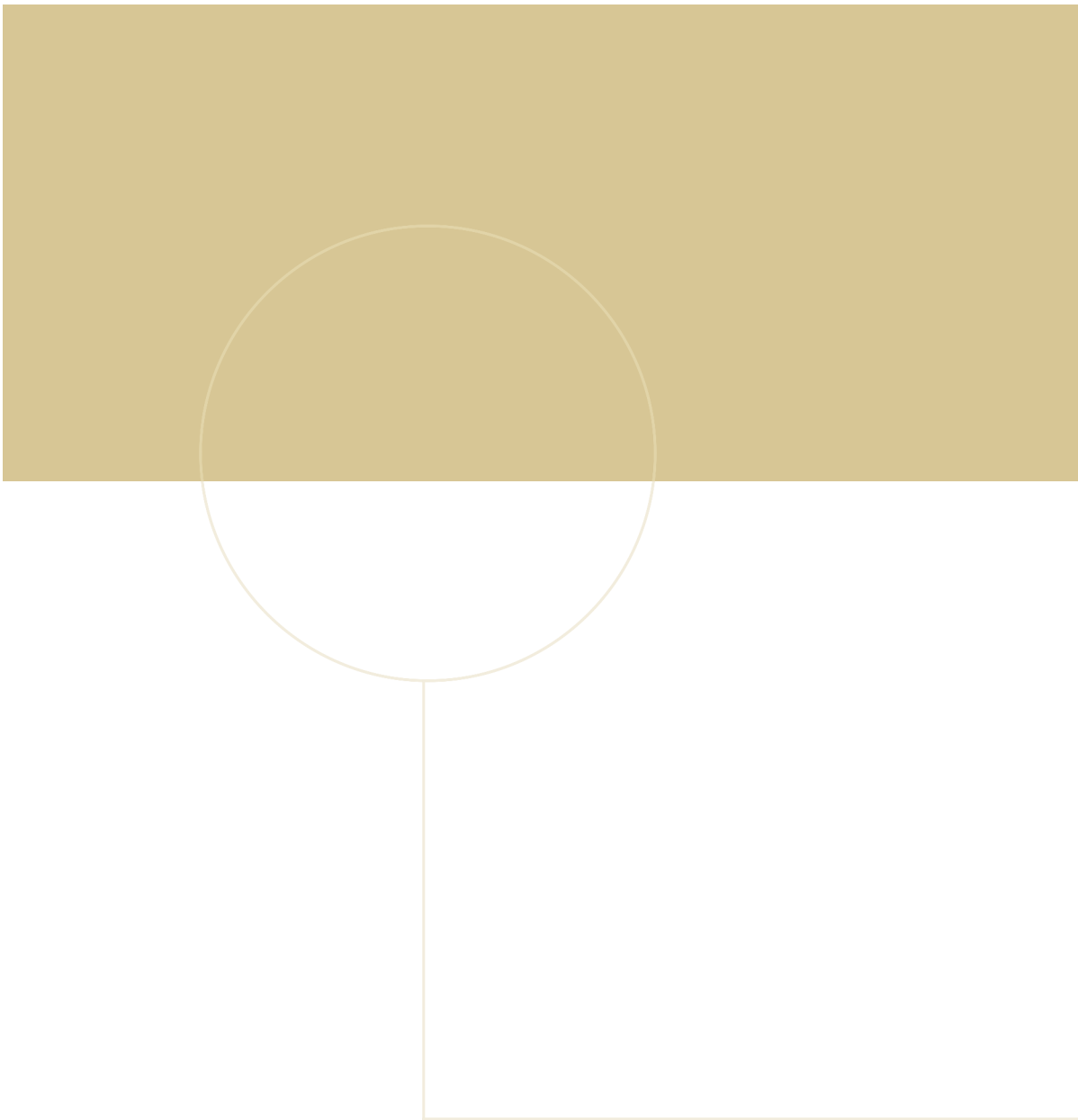
[25] J. Nakic, A. Granic, and V. Glavinic. Anatomy of student models in adaptive learning systems: A systematic literature review of individual differences from 2001 to 2013. *Journal of Educational Computing Research*, 51(4):459–489, 2015.

[26] N. B. A. Normadhi, L. Shuib, H. N. M. Nasir, A. Bimba, N. Idris, and V. Balakrishnan. Identification of personal traits in adaptive learning environment: Systematic literature review. *Computers & Education*, 130:168–190, 2019.

[27] M. Vandewaetere, P. Desmet, and G. Clarebout. The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior*, 27(1):118–130, 2011.

[28] E. Verdu, L. M. Regueras, M. J. Verdu, J. P. De Castro, and M. A. Pérez. Is adaptive learning effective? a review of the research. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, volume 7. World Scientific and Engineering Academy and Society, 2008.

[29] L. S. Vygotsky and M. Cole. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.



NTNU

Norwegian University of
Science and Technology