ulm university universität
uulm

Universität Ulm | 89069 Ulm | Germany

Faculty of Engineering, Computer Science and Psychology

How to incentivize efficient learning choices in digital learning environments? A reinforcement learning approach applied to an educational game

Master thesis at Ulm University in the program Cognitive Systems

Submitted by:

Reena Charlotte Pauly

reena.pauly@uni-ulm.de

Thesis advisors:

Dr. Falk Lieder

Rationality Enhancement Group, Max Planck Institute for Intelligent Systems, Tübingen

Prof. Dr. Dr. Daniel Braun

Ulm University, Ulm

Supervisors:

Victoria Amo, M.sc. and Lovis Heindrich, M.sc.

Rationality Enhancement Group, Max Planck Institute for Intelligent Systems, Tübingen

December 2021

Name: Reena Charlotte Pauly Matriculation number: 1713839

Declaration

I hereby declare that I wrote the master thesis independently and used no other aids than those cited. In each individual case, I have clearly identified the source of the passages that are taken word for word or paraphrased from other works.

Tübingen, December 3^{rd} 2021 .....

Reena Charlotte Pauly

Abstract

Learners often struggle to overcome the tendency to value short-term pleasure higher than long-term goal fulfilment. The recent trend to add game-like elements such as points and badges to educational contexts in order to enhance learner's motivation has not been able to consistently address this issue. That is likely due to the fact that such interventions often make the existing motivational dynamics more salient but do not change them. Research has shown that focusing incentives on effort rather than performance can have a positive impact on learner's academic achievements. The concept of optimal brain points developed by Xu, Wirzberger & Lieder (2019) demonstrates that methods from the field of reinforcement learning allow to align short-term rewards for learning choices with their expected long-term benefit. In this thesis project, a principled and scalable approach to incentivizing efficient learning choices is developed based on these insights and applied to a real-world educational game. Specifically, the approach entails a formal model of the educational game and can compute the choice expected to maximize the learner's progress. Its evaluation in a controlled online experiment with a simplified learning task produced promising results, showing that the derived incentives can positively impact both learners' choice behaviour and their learning outcomes. Therefore, an evaluation of the approach in the educational game is planned as a result of this thesis project.

Contents

List of Figures

List of Tables

1 Introduction

Do you have a long list of bookmarks to interesting articles you want to read one day? Or do you maybe keep getting reminders to continue your attempt to freshen up your Spanish from before your last vacation? Or are you still subscribed to that YouTube channel offering free guitar lessons without having touched your guitar in months? You are not alone.

Nowadays, more people can access digital educational resources than ever before. However, many people struggle to fully exploit the offered resources in the pursuit of their personal learning goals [1]. One possible reasons for this is an aversion towards failure, leading to a preference for easy tasks over tasks that might allow them to really progress [1]. This can be viewed as a form of procrastination, as it stems from the tendency to value short-term pleasure (succeeding at an easy task) over long-term benefit (developing one's skills or knowledge) [1, 2].

In the strive to help learners persist through such motivational difficulties, gamification has often been the tool of choice [3]. Gamification means the incorporation of game elements in a non-game context and is often applied in the form of points, badges or a leaderboard [4]. Another approach is to convey lessons in form of educational games [5].

The popularity of gamification is partly based on it being linked to positive effects on motivation and retention [3]. On the other side, it has been shown that gamification applied to educational contexts can also elicit negative effects. These can include indifference, loss of performance and an increase in undesired behaviour [3] and are often attributed to poor design [4]. Specifically, emphasizing momentary performance through game elements can intensify rather than alleviate the problem of striving to avoid failure [6]. Additionally, incentivizing schemes can often be gamed, meaning that a learner can engage in behaviour that will serve to accumulate points or badges without increasing the efficiency or success of their actual learning efforts [1, 6]. Overall, applications of educational games or gamification in educational contexts have produced mixed results and in part lack a solid theoretical foundation [1].

A suitable foundation could be the concept of implicit theories of intelligence [7, 8]. Research based on this concept has shown that it is more beneficial to praise the effort exerted by learners than their momentary performance [7]. Xu, Wirzberger & Lieder (2019) built on that research and developed a computational approach that aligns short-term rewards and expected long-term learning progress by calculating the value of practice [1]. Thereby, it addresses both issues identified so far: what to incentivize and how to do so.

The goal of this thesis project is to develop and test a principled, general approach to incentivizing self-directed learning in digital learning environments in such a way that students learn as much as possible as efficiently as possible. This goal is to be reached by extending the method developed by Xu, Wirzberger & Lieder (2019) in a way that it can be generally applied to different educational environments and scale up to the complexities found in such real-world scenarios.

To that end, we discuss relevant findings concerning the application of gamification in educational contexts and present the reinforcement learning methods forming the basis of the approach in Chapter 2. In Chapter 3, the general approach to incentivizing self-directed learning in digital learning environments is presented and its application to the the studied educational game explained. Chapter 4 details the design and the results of a controlled online experiment evaluating the method developed in Chapter 3. A field evaluation of the method with the actual educational game was unfortunately out of the scope of this thesis. Chapter 5 features other future directions of research, before the thesis reaches its conclusion in Chapter 6.

2 Background

This chapter gives an overview of the most relevant findings and concepts building the base for this thesis. Firstly, we discuss the concept of gamification, its application to educational contexts and the implications of different empirical findings for its design. Secondly, an introduction to the field of reinforcement learning and the concepts and methods relevant for the approach developed in Chapter 3 is given.

2.1 Gamification Applied to Educational Contexts

Gamification is the introduction of game elements to non-game contexts [9]. Namely, those game elements are most often points, badges, levels and leaderboards [3, 10]. Points allow to provide quantified feedback and are often tied to levels. Badges are typically rewarded for completing tasks outside the main scope. Leaderboards communicate how individuals are doing in the gamified system and thereby can instill a sense of competition [10].

Gamification interventions are thought to offer motivational affordances, leading to a change in behaviour [9]. However, precise theories concerning the way gamification impacts behaviour are still underdeveloped [1, 9, 10]. Nevertheless, the interest in gamification has spiked in the last decade [9], as can be seen in Figure 2.1. A large number of corporations have applied gamification methods hoping to enhance their employees' motivation [4]. Out of the same reasons, designers of educational environments have been drawn to add gamification to their toolboxes [3]. Aside from being put to practical use, gamification has also increasingly been the subject of academic interest over the past years [9].

But before we delve into the research on gamification in education, the next section discusses an important conceptual differentiation.

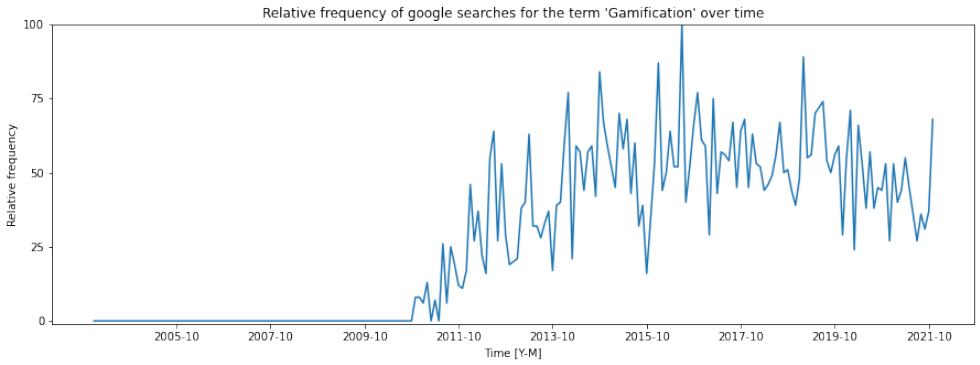


Figure 2.1: Increase in interest in gamification: The y-axis shows the number of google searches for the term "Gamification" relative to the maximal observed value (100%) over time denoted along the x-axis. Data source: Google Trends (https://www.google.com/trends).

2.1.1 Gamification and Game-Based Learning

Closely related to gamification, and sometimes mixed-up with it, is game-based learning [11]. Game-based learning means the presentation of learning material within a game [11]. To illustrate the difference, let us consider two examples: A math teacher gives out stickers to students for handing in extracurricular exercise sheets and keeps a poster in their classroom showing how many stickers each student has earned. That is gamification. If the math teacher gives the homework to spend 30 minutes playing a video game teaching math, that is game-based learning. The motivational affordances offered by incentive structures such as points, badges and leaderboards are often part of game-based learning, but do not necessarily need to be. This thesis deals with the mechanisms and potential design possibilities of gamification. This is done with the example of an educational game that employs both game-based learning and prominently features gamification mechanisms. In principle, the method that is to be developed should also be applicable to pure gamification approaches which do not involve an actual game. This claim will be revisited in Chapter 5.

Due to these reasons, we will primarily focus on empirical research concerning the application of gamification in education as opposed to game-based learning, which of course also includes research tackling both topics at once.

2.1.2 Empirical Research on Gamification in Education

On the one hand, applications of gamification in education have been reported to improve students' motivation for and performance in learning new material [3, 12]. On the other hand, gamification has also been linked to inconsistent and even negative effects on learners [3, 13]. Several studies report that the studied interventions elicited a positive effect only on a subset of the participants [3]. More concerning, gamification often does not only fail to produce the desired effects but does instead lead to undesired behaviour [3].

One possible reason for that is a misalignment between the behaviour that is to be incentivised and the motivational affordances offered by the gamification scheme [4]. A study highlighting the importance of carefully choosing which behaviour to incentivize was conducted by Hakulinen, Auvinen, & Korhonen (2013) [14]. Computer science students were awarded badges for handing in assignments early to encourage better time management. There were also badges encouraging carefulness, which was operationalized as handing in assignments with as few errors as possible, to reduce trial and error approaches to passing assignments. It was found that students aiming to earn time-management badges did so by sacrificing carefulness, which was of course not the intended effect [14]. A more general example is wanting to enhance learning but rewarding performance, which can lead learners to intensify their efforts to avoid experiencing failure by sticking to tasks they have already mastered [1].

A second factor contributing to undesired behaviour is users gaming the system deliberately. Gaming the system, in this context, means learners trying to succeed strictly in the terms of a gamified environments reward structure without engaging in the learning task itself [6, 15]. Across studies, it becomes apparent that if the gamification method can be gamed, it most probably will be gamed. In a study evaluating gamifying an e-learning platform for university students, participants reported figuring out that they could earn badges by uploading empty screenshots [16]. Another university released a gamification initiative that - among other things - rewarded points for clicking a link to reading material as a measure of the students' reading efforts. Students engaged in a competition concerning who could write the fasted script for automatically clicking the link and thereby achieving astronomically high levels within the first day of the gamification program [17].

It can be concluded that gamification interventions need to be well-planned and based on a solid theoretical foundation in order to ensure positive impacts on the learners [3]. The goal behaviour should be carefully chosen with the help of relevant research

on learning and motivation. Additionally, it should be made sure that the designed intervention is not gameable. One important framework in motivational psychology is the concept of implicit theories of intelligence [7]. This framework and its implications for designing gamification interventions are discussed in the next sections.

2.1.3 Implicit Theories of Intelligence and their Impact on Motivation and Performance

A large body of research shows that a person's implicit theory of intelligence, also called mindset, influences their learning motivation, persistence in face of challenges and academic achievements [7, 8]. Explicitly, Carol S. Dweck distinguished the fixed mindset and the growth mindset.
A person with a fixed mindset believes intelligence and abilities are innate and not subject to change. Because of that belief, they are more likely to seek out challenges which allow them to validate their abilities than those which provide them with opportunity to improve them [7]. Consequently, they are likely to attribute experiences of failure to a lack of necessary intelligence [7]. Along the same lines, the necessity to excert effort is more likely to be interpreted as an indicator of insufficient abilities [7].
A person with a growth mindset, on the other hand, believes that intelligence and abilities can develop with practice and instructions from others. Therefore, they are likely to perceive excerting effort as a mean to growing their abilities. Accordingly, they seek challenges allowing them to build up their abilities and to seek support in facing them. When faced with an experience of failure, they are more likely to attribute it to a lack of effort or use of suboptimal strategies than to a lack of ability [7].
The effect the described mindsets may have on the people's academic achievement has been studied on numerous occasions [7]. For example, US-American middle school students holding a growth mindset achieved higher overall grades and were more likely to choose to enrol in a more challenging math course than those holding a fixed mindset [18]. In an analysis of data from a complete cohort of Chilean 10th graders, it was shown that the mindset held by a student can be used to predict their academic achievement in a standardised test almost as well as their socio-economical background [8]. Interestingly, students from lower-income families were more likely to lean towards a fixed mindset while at the same time benefiting more from a growth mindset than students from higher income families [8]. This is likely due to the fact that a growth mindset can help a person to overcome challenges, of which there are many for disadvantaged pupils [8].

Seeing how the mindset held by a person can influence their achievement and development leads to the next important question: Can mindsets be taught? In studies comparing the effect of the type of praise offered to students, it was found that the mindsets held by the children can be influenced. Specifically, praising intelligence was shown to induce a fixed mindset in comparison to praising effort and use of strategies [7]. The next section discusses the results of an online intervention employing gamification in the context of an educational game designed to induce a growth mindset.

2.1.4 Brain Points can Foster a Growth Mindset in Educational Games

O'Rourke, Haimovitz, Ballweber, Dweck and Popovic (2014) designed an intervention to promote a growth mindset using a point-based reward system within an educational game [5]. The game aimed to teach the concept of fractions to pupils in elementary school [5]. The experimental intervention consisted of an animated narrative addressing that intelligence can be trained and brain points awarded for usage of strategies. The name "brain points" was chosen to support the narrative that applying effort and using strategy trains the brain [5]. The control condition featured a neutral narrative and points rewarded based on performance. It was found that children interacting with the experimental version of the game played for longer and used more of the incentivized strategies than those children playing the control game [5]. In a follow-up study [19], the differential effects of the intervention's components were disentangled. The results suggest that the animations used to transport the growth mindset narrative did not increase how long players stuck with the game. This is attributed to the overall effect that many players quit the game during those introductory animations, possibly out of impatience [19]. The positive effect of brain points on player retention is corroborated [19]. Furthermore, brain points rewarding use of strategies were contrasted with brain points presented at random points in time during game play. Players receiving meaningful brain points persisted longer than those awarded random points. This finding allows the conclusion that brain points impact player retention positively by specific incentives for strategies associated with a growth mindset, rather than through general encouragement [19]. It is important to note that the rewarded strategies are specific to the studied game. A generalization of these findings would therefore strongly depend on the possibility to translate the incentive structure to work for different contexts [19]. Xu, Wirzberger & Lieder (2018) developed a more principled and general approach for encouraging persistent learning efforts, which will be presented in the next section.

2.1.5 Optimal Brain Points Reflect the Long-Term Value of Mastering New Skills

As discussed in Section 2.1.2, hand-designed incentive schemes risk being gamed. For the brain points studies discussed above, that means that learners might use strategies in order to trigger the brain point rewards without being invested in learning about fractions or strategies. The concept of optimal brain points [1] addresses that issue by leveraging a principled computational approach for designing incentives. The core idea is to incentivize learners' study choices and effort allocation in a way that short-term rewards align with long-term benefits [1]. This is achieved by calculating the expected value of investing effort into acquiring a new skill opposed to exploiting a less effective skill which has already been mastered [1]. That value, in combination with the expected progress toward skill acquisition, is used to provide learners with optimal feedback regarding their study choices in a simple learning paradigm. In that paradigm, participants could choose between steering a space ship to a goal position using arrow keys or trying to find out which of the letter keys teleports the spaceship directly to the goal. It was found that participants perceiving optimal brain points were more persistent in attempting to master the more difficult but also more efficient skill than those who only got information on the action cost and goal reward [1]. The optimal brain points a player in the experimental condition receives depend on their choices, not on the obtained outcome. In other words, optimal brain points reward strategy rather than performance, thereby adhering to the principles of fostering a growth mindset in the same fashion as the brain points intervention [1, 5, 19].

The goal of this thesis project is to extend the concept of optimal brain points to develop a scaleable method for incentivizing study choices in more complex and realistic environments than steering a spaceship icon across a grid. To that end, the next chapter presents the methodological foundations for the computational approach behind the optimal brain points [1], which also form the second pillar for this thesis project.

2.2 Reinforcement Learning

The field of reinforcement learning (RL) concerns itself with the question how agents can learn to maximize their future rewards trough trial and error learning [20]. One key characteristic of RL problems is poor prior knowledge, meaning that the environment

is only accessible to the agent through interaction [20]. A further characteristic is that the agent's decisions are sequential and can have long-term consequences. That can lead to the temporal credit assignment problem, or the challenge of learning good decision sequences from delayed rewards [20].

Machine learning models using RL taking on these kinds of problems have a strong normative grounding in psychological research on learning behaviour and have been brought to use successfully [21]. For example, an artificial RL agent gained worldwide public attention in 2015, when Google DeepMind's AlphaGo defeated a human expert player in the game of Go [22]. Nevertheless, the application of RL agents is by far not restricted to game play. They form an important part of control strategies in robotics [23]. Additionally, they have been employed to optimize both traffic signaling [24] and chemical reactions [25], to name just a few examples.

RL tasks are commonly modeled as Markov Decision Processes, which will be presented in the next section.

2.2.1 Markov Decision Processes

A Markov Decision Process (MDP) is defined by a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition function $P(s'|s,a)$, a reward function $R(s_t, a_t, s_{t+1})$ and the discount factor γ [20].

At each time step t of the discrete agent-environment interaction, the state $s_t \in \mathcal{S}$ conveys the information about the environment that is available to the agent [20]. It forms the basis for the agent to choose an action a_t from the set of actions $\mathcal{A}$ available in the state. The states satisfy the Markov property, meaning that the probability of moving to a state depends only on the previous state and the action taken in it (Eq. 2.1) [20].

$$P(s_t|s_{t-1}, s_{t-2}, ..., s_1, a) = P(s_t|s_{t-1}, a) \tag{2.1}$$

This allows to define the transition function $P(s'|s,a)$. The reward function returns the immediate reward signal the agent receives for transitioning from s to s' [20]. The described interaction is summarized in Figure 2.2.

In order to solve the MDP and the RL task , the optimal policy π^* that maximizes total future reward (Eq. 2.2) has to be found [26]. The long-term performance function $J(\pi)$ is defined as in Eq. 2.3. The discount factor γ regulates to what extent future

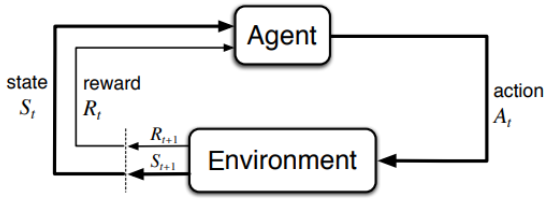


Figure 2.2: Agent-Environment Interaction, taken from [20]

rewards are discounted compared to immediate rewards.

$$\pi^* = \arg\max_{\pi} \ J(\pi) \tag{2.2}$$

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{2.3}$$

In the case that $P(s'|s,a)$ and $R(s_t, a_t, s_{t+1})$ are known, one possible algorithm to find the optimal policy π^* is the value iteration [26]. Its reasoning is described in the following section.

2.2.2 Value Iteration

For each state $s \in S$, the utility of being in state s is defined as the cumulative future reward that can be obtained by an agent starting in that state and following policy π (Eq. 2.4) [26].

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | \pi, s_0 = s\right] \tag{2.4}$$

This transforms into the linear Bellmann Equation (Eq. 2.5), which recursively defines the utility of a state as the sum of the immediate reward of the current state and the expected utility of the next state, assuming that the agent follows policy π [26].

$$V^{\pi}(s_t) = \sum_{a_t}(a_t|\pi)\left\{r_{s_t,a_t} + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) * V^{\pi}(s_{t+1})\right\} \tag{2.5}$$

The optimal policy needed to solve the MDP will produce the action maximizing the future reward. Therefore, the optimal value function (Eq. 2.6) is obtained from Eq. 2.5 by assuming the agent follows the optimal policy π^* [26].

$$V^*(s_t) = \max_{a_t} \left\{ r_{s_t,a_t} + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) * V^*(s_{t+1}) \right\} \tag{2.6}$$

As the name of the algorithm hints at, the optimal value function is found by iteratively updating an initial guess $V_0(s)$ (Eq. 2.7) [26].

$$V_{i+1}(s_t) \leftarrow \max_{a_t} \left\{ r_{s_t,a_t} + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) * V^*(s_{t+1}) \right\} \tag{2.7}$$

In order to perform value iteration, one can define a convergence threshold δ. Once the updates to the utilities become smaller than δ, the value function is considered to have converged and therefore the optimal value function. Once the optimal value function has been found, the optimal policy is extracted with one more pass over all states $s \in S$. For each non-terminal state, the policy function returns the action which maximizes future reward $\pi(s) = \arg\max_a \{r_{s_t,a_t} + \gamma V_i(s_{t+1})\}$ for that state.

Value iteration is not possible in cases in which the model of the environment, $P(s'|s, a)$ and $R(s_t, a_t, s_{t+1})$, is not known to the agent. Then, the optimal policy has to be found trough interaction rather than deliberation [20]. One common approach is Q-learning and will be discussed next.

2.2.3 Q-Learning

In Q-learning, the quality of an action taken in a state is considered and captured in the state-action value function $Q(s, a)$. Instead of calculating a sum over all possible actions weighed by their probabilities and all possible resulting next states (Eq. 2.4), the state-action value function (Eq. 2.8) directly estimates the usefulness of taking one specific action in that state.

$$Q(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) * max_{a'} Q(s', a') \tag{2.8}$$

These Q-Values are updated during the agents interaction with the environment, as

Algorithm 1 Value Iteration

Initialize $\mathcal{S}, \mathcal{A}, R(s, a, s'), \gamma$
$\delta = 10^{-7}$
procedure VALUE ITERATION($\mathcal{S}, \mathcal{A}, R(s, a, s'), \gamma, \delta$)
    while maxdiff $<= \delta$ **do**
        maxdiff = 0
        for $s \in S$ **do**
            if s is a terminal state **then**
                V(s) = 0
            else
                $r_t = R(s_t, a_t, s_{t+1})$
                $V_{i+1}(s_t) \leftarrow \max_{a_t} \left\{ r_t + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) * V^*(s_{t+1}) \right\}$

                maxdiff $= \max($maxdiff$, V_{i+1}(s) - V_i(s))$
    for $s \in S$ **do**
        if s is a terminal state **then**
            $\pi(s) = None$
        else
            $\pi(s) = \arg\max_a \{r_{s_t,a_t} + V_i(s_{t+1})\}$

specified in (Eq. 2.9).

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right) \tag{2.9}$$

During interaction, the agent chooses an action according to some sampling strategy. Often, the ϵ-greedy sampling strategy is used (Eq. 2.10). It handles the exploration-exploitation dilemma that presents itself in reinforcement learning: In order to learn to act in a way leading to high rewards, an agent has to both acquire knowledge about its environment by exploring it and obtain rewards by exploiting the knowledge it already gathered. In many scenarios, these strategies are mutually exclusive [27]. In the ϵ-greedy sampling strategy, exploration and exploitation are balanced by the parameter $\epsilon \in [0, 1]$. Each time an action has to be selected, it will be the one the agent currently assumes to yield the highest future rewards with a probability $p = 1-\epsilon$ and a randomly chosen action with probability $p = \epsilon$. Over the course of the training, ϵ is usually decayed, allowing the agent to explore more in the beginning and to hone

in on its found strategy toward the end [26].

$$p(a|s) = \begin{cases} 1-\epsilon, & \text{if } a = \arg\max_a Q(s,a) \\ \epsilon * \frac{1}{|A|}, & \text{otherwise} \end{cases} \tag{2.10}$$

Another parameter shaping the learning process is the learning rate $\alpha \in [0,1]$. It determines the magnitude of the updates applied to the Q-Values and with that how long past interactions with the environment can influence the Q-Values. If α is small, the Q-Values are only updated slightly and past interactions have a lot of weight. If α is closer to 1, new information is weighed more strongly.

Algorithm 2 Q-Learning

```
Initialize S, A, R(s, a, s'), α
procedure Q-LEARNING(S, A, R(s, a, s'), ε, α, γ)
    for t in T do
        ε = 1 − t/T
        s_0 = startstate
        while s is not a terminal state do
            a = ε − greedy(s)
            r = R(s, a, s')
            Q(s, a) ← Q(s, a) + α (r + γ max_a' Q(s', a') − Q(s, a))
```

The biggest constraint of Q-learning and other value-iteration based methods is that they can only effectively deal with very low-dimensional state spaces.

2.2.4 Deep Q-Networks

In order to find the optimal policy in high-dimensional state spaces, one has to parameterize the policy function and directly estimate the optimal one, which is termed policy iteration [21]. A very successful approach to derive the optimal policy for high-dimensional state spaces is deep Q-learning, which uses neural networks as a function approximator [21].

Neural networks are well suited to extract relevant features from a high-dimensional input. Combined with parameterized Q-Functions $Q(s,a,\theta)$, a Deep Q-Network (DQN) is obtained [21]. The input is passed through several layers in order to obtain the output: the estimated Q-values for each possible action in that state. In order to calculate the loss between these Q-Values, emitted from the so-called policy network, and the

target Q-Values, the next state s' is passed trough a second network named the target network. The loss is used to perform gradient descent and backpropagation on the weights of the policy network as described by Equation 2.11 [28].

$$\theta_{t+1} = \theta_t + \alpha \left(Y_t^Q - Q(s, a, \theta_t) \right) \Delta_{\theta_t} Q(s, a, \theta_t) \tag{2.11}$$

With the target Y_t^Q being estimated with the parameters of the target network θ^-:

$$Y_t^{DQN} = \left(r_{t+1} + \gamma \max_a Q(s', a, \theta_t^-) \right) \tag{2.12}$$

The weights of the target network remain fixed to ensure stability in the learning process. They are only updated once every predefined number of iterations by copying the weights of the policy network [21]. In addition to the fixed target Q-Values, another improvement to the algorithm was introduced: experience replay [29]. Instead of learning directly from the interactions with the environment as they occur, experiences consisting of the initial state s, the action taken a, chosen for example following an $\epsilon - greedy$ policy, the resulting state s' and the observed reward r are stored in a replay memory. From this memory, experiences are then uniformly sampled and fed to the policy network described above. This alleviates a problem inherent to reinforcement learning with policy iteration: The experiences made during interaction with the environment are highly temporally correlated, thereby violating the assumption of identically and independently distributed data which are necessary to optimally update the weights using gradient descent [30].

2.2.5 Reward Shaping

Pure trial-and-error learning can become unfeasible in large state or action spaces, as it becomes more and more difficult for an agent to infer a strategy maximizing the reward for a sequence of actions from sparse rewards granted for single actions [31]. This is the temporal-credit-assignment problem [31]. One option to overcome this problem is to provide the agent with additional information concerning the reward landscape of their environment. This is called reward shaping [31, 32].

This guidance towards optimal behaviour is achieved by letting the agent interact with an altered MDP $M' = (S, A, P, R')$ instead if the original MDP $M = (S, A, P, R)$ it has to solve [32]. The altered reward function R' consists of the original reward function $R(s, a, s')$ as well as of the shaping function $F(s, a, s')$ (Eq. 2.13). The rewards

obtained from the shaping function are often called pseudo-rewards [33, 2]. In order to be truly helpful, the shaping function has to be crafted very carefully. Otherwise, the agent might end up maximizing its reward in a manner neither foreseen nor desired without solving the actual task at hand [31, 32].

$$R'(s, a, s') = R(s, a, s') + F(s, a, s') \tag{2.13}$$

Avoiding those pitfalls can be achieved by choosing a shaping function such that any optimal policy π^* for M' is also an optimal policy π^* for M [32]. This condition, which is called invariance property, can only be guaranteed to be fulfilled by using a potential-based shaping function (Eq. 2.14), as has been proven in the shaping theorem [32, 33]. The shaping function is then defined as the difference between the discounted potential of the new state $\gamma\phi(s')$ and the potential of the old state $\phi(s)$ [32].

$$F(s, a, s') = \gamma\phi(s') - \phi(s) \tag{2.14}$$

According to Ng, Harada and Russel (1999) [32], the optimal potential function equals the optimal value function (Eq. 2.15). The optimal potential-based shaping function (Eq. 2.16) consequently grants the agent a positive pseudo-reward if its action moved it into a state with a higher utility than that of the previous state.

$$\phi^*(s) = V^*(s) \tag{2.15}$$

$$F^*(s, a, s') = \gamma V^*(s') - V^*(s) \tag{2.16}$$

Optimal pseudo-rewards therefore dissolve the conflict between long-term and short-term reward maximization. The agent obtains the highest possible long-term reward by only considering the highest reward for the next step [33, 2].

In many real-world applications, $V(s)$ is neither known nor easy to compute. For those cases, distance-based or subgoal-based have been proposed as potential functions [32].

Reward shaping is a crucial concept in the pursuit of a principled approach to incentivizing learning choices and the methodology that allowed to compute the optimal brain points discussed in Section 2.1.5. In another study, Lieder, Chen, Krueger and Griffiths (2019) [2] used the principles of reward shaping to develop an optimal

gamification method for decision support. This was tested in several behavioural experiments. It was shown that presenting people optimal pseudo-rewards which align the immediate reward gained by an action with its long-term value can support them in making more far-sighted decisions. Furthermore, the results suggest that optimal gamification can assist people in overcoming procrastination and in prioritizing when dealing with a multitude of tasks [2].

2.3 Summary

Gamification can be a useful tool to help learners overcome motivational obstacles, but is has to be brought to use with care [3]. Specifically, we have seen that the objective has to be well defined and the gamification prompts need to align with it [14]. Useful input to how to choose what to incentivize can be drawn from the research on mindsets [7]. It offers insight into why some gamification incentives focusing on performance have differential effects on different users and allows to conclude that it is a more promising approach to instead reward effort and use of strategies [5].

Additionally to adequately choosing which behaviour to incentivize, it is important to carefully design how the incentives should be distributed in order to ensure they cannot be gamed and thereby lead to undesired behaviour [17]. A useful resource to do so has shown to be reward shaping [2], as it allows to align short-term rewards for behaviour with its expected long-term benefits.

The research question tackled in this thesis is whether these principles allow to develop a scalable method for deriving incentives for real-world educational environments. To that end, the next Chapter details the development of a general approach to calculate incentives based off the methodology used to derive optimal brain points [1]. Furthermore, the educational game serving as a use case is introduced and the application of the approach to it specified.

3 Development of a Principled Approach to Incentivizing Self-Directed Learning in Digital Learning Environments

3.1 Modeling the Choice of Educational Activities as a Markov Decision Process

For the purpose of developing a scaleable principled method for incentivizing learners to choose their learning activities in an efficient way, that choice is modeled as a Markov Decision Process (MDP), which has been introduced in Section 2.2.1. In the following section, the necessary information from the educational environment the method is to be applied to is specified.

First of all, the skills that are to be trained in the educational environment have to be defined. Those can be a set of distinct skills (e.g. math, English and chemistry knowledge) or aspects of a single skill (e.g. English vocabulary, grammar and writing skills), depending on the context of application. Both concepts will be summarized with the term skill from here on. The number of skills to be trained is denoted as N_S. Additionally, a way to measure the learner's competence c for each skill has to be provided. Thereby, for each skill a linearly ordered set of competence values $\mathcal{C}$ can be defined. Higher competence values are considered superior to lower competence values. In the context of gamification, increasing competence values is the objective of the intervention. The number of possible competence values c the i^{th} skill can assume is denoted as N_{c_i}. The set of competence values for the i^{th} skill is defined in Eq. 3.1.

$$\mathcal{C}_i = \{c_1, c_2, ..., c_{N_{c_i}}\} \tag{3.1}$$

Another crucial parameter which has to be provided is the learning goal $\boldsymbol{g}$. It defines

a goal competence value $c_g \in \mathcal{C}$ for each skill which has to be reached by the learner in order to consider the learning goal completed (Eq. 3.2).

$$\boldsymbol{g} = (c_{g_1}, c_{g_2}, ... , c_{g_{N_S}}) \tag{3.2}$$

As already introduced in Section 2.2.1, a state s of the MDP needs to consist of all information relevant to satisfy the Markov assumption. In this model, we can distinguish learner-specific information θ and system-specific information β (Eq. 3.3). θ comprises all parameters needed to compute the current competence values $c \in \mathcal{C}$ of each skill (Eq. 3.4). Additionally, if the environment's model of skill improvement relies on additional factors β, these are also included in the state. Those are all factors that allow to differentiate between states in which the competence levels are equal but the transition probabilities are not.

$$s = (\theta, \beta) \tag{3.3}$$

$$\boldsymbol{c} = (c_1, c_2, ... , c_{N_S}) = f(\theta) \tag{3.4}$$

Consequently, we can define whether the learning goal is satisfied by a state by calculating the current competence values $\boldsymbol{c}$ and comparing them against the set learning goal as defined in Eq. 3.5.

$$d(\boldsymbol{c}, \boldsymbol{g}) = c_i \geq c_{g_i} \qquad \forall i \in \{1..N_s\} \tag{3.5}$$

The learning environment further has to comprise more than one educational activity the learner can choose between. The choice of activity is modelled as the actions making up the action space of the MDP (Eq. 3.6). The number of activities a learner can choose between and with that the number of actions is denoted as N_a.

$$\mathcal{A} = \{a_1, a_2, ..., a_{N_a}\} \tag{3.6}$$

The transition function of the MDP $P(s'|s, a)$ depends on the way the educational environment models skill improvement. Specifically, a way to define how the different educational activities are expected to impact the skills depending on the current state is needed.

Two estimations are needed in order to construct the MDP's reward function $R(s', a, s)$: First, an estimation of the effort required to carry out each learning activity r_a. Furthermore, an estimation of the usefulness of reaching the learning goal $\boldsymbol{g}$ r_g is needed. Consequently, the reward function is defined in Eq. 3.7. Note that $\boldsymbol{c}$ can be calculated from s as specified in Eq. 3.4.

$$R(s', a, s) = \begin{cases} r_g - r_a, & \text{if } d(\boldsymbol{c}', \boldsymbol{g}) \wedge \neg d(\boldsymbol{c}, \boldsymbol{g}) \\ -r_a, & \text{otherwise} \end{cases} \tag{3.7}$$

In order to solve the MDP, the optimal policy π^* which maximizes future reward (Eq. 3.8) has to be found. In the context of the application to a learning environment, that equates to reaching the learning goal without excerting more effort than necessary to do so.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{3.8}$$

As has been discussed in Chapter 2, learners often choose suboptimally when it comes to investing their time and effort in a way that allows them to efficiently reach their long-term goals. Hence, the choice of activities is to be incentivized by presenting brain points $B(s, a)$ to the learner (Eq. 3.9). For the purpose of evaluating the effectiveness of the developed approach with simulated agents, we therefore modeled learners as myopic agents. Myopic means short-sighted and describes decision strategies that weigh the short-term outcomes much higher that the long-term consequences. In order to model that behaviour, the simulated agent looks one step into the future and chooses the action with the highest expected immediate reward.

In Section 2.2.5, we discussed that reward shaping allows to dissolve the conflict between long-term and short-term reward maximization [32]. The key to deriving brain points therefore is to define a shaping function $F(s', a, s)$ which offers additional information concerning the MDP's reward landscape. As explained in Section 2.2.5, the simulated agents then interact with the altered MDP with the reward function $R'(s, a, s) = R(s, a, s') + F(s, a, s')$.

Because the brain points are meant to incentivize the choice and therefore are presented to the learner before carrying out an action, they can only depend on s and a and not on s', as the latter is not known at the time the brain points are calculated. Therefore, they are calculated as the expected value across all possible s' in Eq. 3.9.

Depending on the complexity of the specific application of this model, a solution or approximated solution of the MDP is used to define the shaping function. A learner trying to myopically maximize the brain points they receive will then simultaneously approximate the optimal policy. That means that they improve their skills efficiently, without wasting effort on educational activities which are unlikely to to lead to an increase in skill in the given state.

$$B(s,a) = \sum_{s'} P(s'|s,a)(R(s',a,s) + F(s',a,s)) \tag{3.9}$$

The specific educational environment that will be modeled and used to evaluate the effect of presenting optimal brain points to learners is presented in the next section.

3.2 The Use Case: An Educational Game for Learning English

This section first introduces the organisation that published the educational game and the reasoning behind their most important design decisions. This is followed by an introduction to the game itself, after which we take a look at how users interact with it.

3.2.1 Solve Education

As laid out in Chapter 2, a large range of educational games have been and are being developed. However, these resources are often designed to meet the needs of students who are able to pay for the respective products [34]. Founded in 2015, Solve Education is a non-profit organisation aiming to provide access to quality education technology to marginalised young people around the globe [34].
To that end, they have identified two main challenges often overlooked in the discussion concerning educational games. The first relates to the learner's motivation. Most providers of educational software assume their users are generally motivated to learn, whether intrinsically or by their surroundings. For learners excluded from the education system, this can be a very shaky assumption. The second challenge is posed by the digital infrastructure available to those students. Solve Education addresses these challenges by creating a highly engaging learning game that can be run on a

low-end smartphone with intermittent internet connectivity [34, 35].
The game is called "Dawn of Civilisation" [36] and provides lessons for acquiring or improving English literacy. The goal is to support the users to develop into independent learners who have gained confidence in their own abilities [34]. Furthermore, competence in the English language may open up access to other free educational resources. It has been shown that improving English language skills benefits both the individual and their society at large [37].

3.2.2 Dawn of Civilisation

In the game Dawn of Civilisation (see Figure 3.1a), the user assumes the role of a city's mayor (see Figure 3.1b). The goal is to build and develop the city. In order to gain the resources for adding buildings and decorations to their city, the user completes different minigames (see Figures 3.1c and 3.1d). Within those minigames, a wide range of English lessons are provided. Specifically, the sixteen different minigames provide lessons to train the user's vocabulary, their grammar, their listening and speaking skills as well as their reading and writing skills.

For completing a minigame, the user gets rewards proportional to their performance in the minigame. In detail, a user is awarded reward cards (see Figure 3.1e). One reward card is given for the attempt, and one additional one per 20% of correct answers. Accordingly, the maximal number of reward cards a user can be awarded for a completing a minigame is 5, which would be the case if they solved at least 80% of that minigame. Each reward card contains different in-game currencies such as cash, stars or hearts with the reward's magnitude varying randomly. Those can then be invested in buying buildings or decorations.

3.2.3 How Do Users Interact with Dawn of Civilisation?

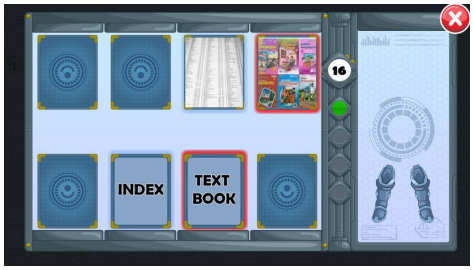
In this section, some key aspects of the way users interact with Dawn of Civilisation will be presented. We performed the following analysis based on user data gathered by Solve Education between March 2020 and January 2021. In that time period, over 80,000 minigames were played by over 5,000 users. For purposes of analysis, only regular users were considered. To that end, users who played on less than 5 occasions were excluded. That leaves roughly 16,500 games played by 370 users.

The first aspect to consider is whether the users exploit the whole range of exercises

(a) Dawn of Civilisations

(b) City Building

(c) Choosing Minigames



(d) Example Minigame

(e) Reward Cards

Figure 3.1: Impressions of Dawn of Civilisations: Figure 3.1a shows the logo of Dawn of Civilisations. In Figure 3.1b, the home screen in which users can design their city is depicted. Is has to be noted that the author only played the game to familiarize herself with it and the shown city is therefore gruesomely underdeveloped. The selection of minigames is depicted in Figure 3.1c and an example of a minigame is given in Figure 3.1d. Finally, the reward cards given after the completion of a minigame are shown in Figure 3.1e.

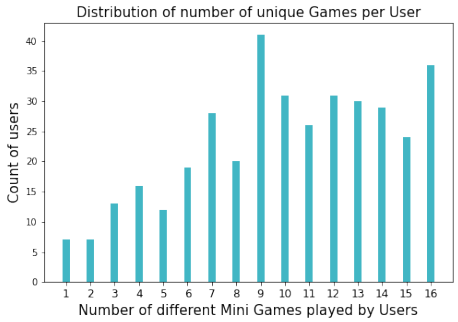
and content offered to them by Dawn of Civilisation. As introduced in the previous section, Dawn of Civilisations contains 16 distinct minigames, each training some facets of the English language. It is visible in Figure 3.2a that many users only interact with a subset - for some larger, for some smaller - of the games they could engage in. Figure 3.2c shows the mini games ranked in their popularity.

The follow-up question to this finding is whether the six skills are trained equally even though many users only engage in a small subset of the offered minigames. As depicted in Figure 3.2b, users tend to interact more with games that train their listening and vocabulary skills and less with those addressing their writing skills.

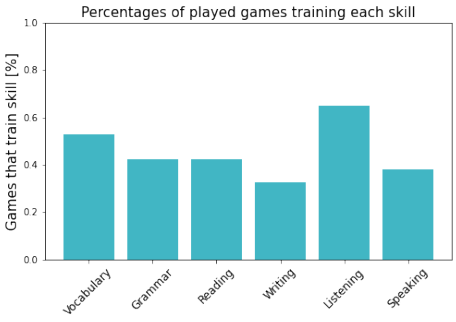
Figure 3.2d shows the mean number of reward cards earned in each mini game. The mean number of reward cards earned in each mini game is moderately correlated with the number of times the game is played (Spearman's r = 0.37). This could be interpreted as a result of the user's preference for easier games as well as the result of improved performance in frequently played games, or as a combination of both effects.

Figures 3.2d and 3.2e show that the minigames also differ in the mean time needed to complete them once and in their mean completion rates.

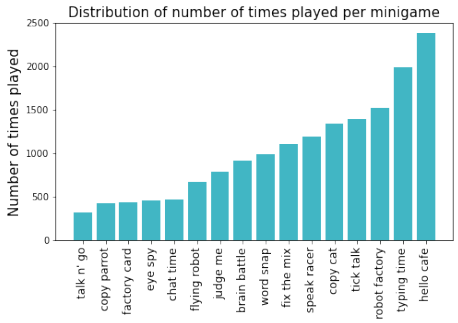
While these characterisations of the way users interact with Dawn of Civilisation are exploratory by nature, some patterns are interesting to discuss within the context of the research findings presented in Chapter 2. Namely, some minigames are vastly more popular than others and that popularity is correlated with the average number of reward cards obtained from the game. This might be an effect of the way rewards are distributed. The number of reward cards a player receives is based on their performance and the actual magnitude of reward is determined randomly. O'Rourke et al. (2016) found that brain points increase persistence compared to random points [19]. Xu, Wirzberger & Lieder (2018) showed that incentivising the value of practice rather than momentary performance can encourage learners to make more efficient study choices. Additionally, it was reported that especially disadvantaged learners, who are the target population of Dawn of Civilisations, can benefit from interventions promoting a growth mindset and the associated behaviour [8]. Therefore, it can reasonably be hypothesized that the users of Dawn of Civilisation might benefit from the planned intervention detailed in the next chapter.



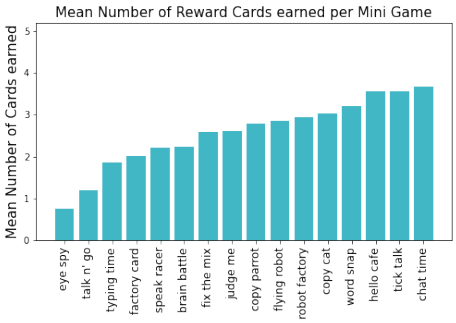
(a) Which games are played?



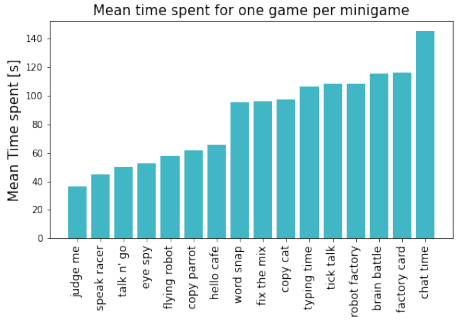
(b) Which skills are trained?



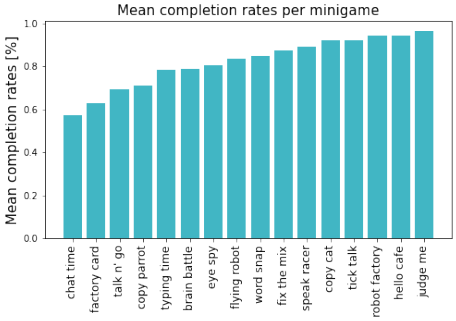
(c) How many different games are played?



(d) How much reward is achieved per game?



(e) How much time is spent per game?



(f) How often are games completed?

Figure 3.2: How users interact with Dawn of Civilization

3.3 Applying the Educational Choice Model to Dawn of Civilisations

The next step towards the development of an approach to incentivize learning choices in Dawn of Civilisations (DoC) is to specify how the model of choosing educational activities defined in Section 3.1 can be applied to the game. To that end, both a detailed description of the game mechanics and the user data provided by Solve Education were utilized. The first question to be answered was whether all users could be reasonably summarized with one set of parameters, or if it would be necessary to construct several models for distinct prototypes of users. Based on results from a k-means clustering approach [38] in which no meaningful clusters emerged (see Appendices B and C), the decision was made to model a single user. In the following, the details of Dawn of Civilisation and how they can be translated into the MDP will be elaborated.

3.3.1 Skills and Competence Values

Dawn of Civilisations is an educational environment for learning English. Specifically, six aspects of learning English are trained: Vocabulary, Grammar, Writing, Speaking, Listening and Reading. Therefore, $N_S = 6$.

A learner's competence in each of these skills is measured in 16 levels, $N_{C_i} = 16$. These levels are constructed as sub-levels of the 6 language levels ranging from A.1 to C.2 defined by the Common European Framework of Reference for Languages (CEFR) [39].

These levels are assessed continuously throughout the learner's interaction with the game. That is possible because the learning material presented to the learner by the minigames is also used to evaluate the learners current competence. Concretely, each skill and level are associated with a set of questions. These questions are presented to the learner in the minigames they play. Once 80% of those questions have been completed by a learner, their corresponding skill is considered to have reached the corresponding level.

When is a question considered complete?
The completion of a question is governed by a spaced repetition review schedule. Spaced repetition is a common tactic, especially in second language acquisition [40,

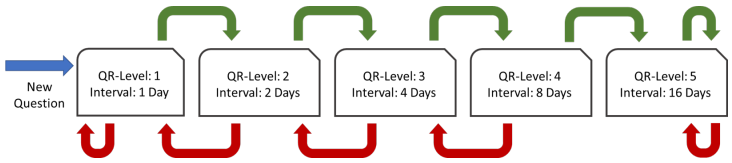


Figure 3.3: QR-Process: The boxes represent the QR-levels questions can assume in Dawn of Civilisations and their corresponding delay periods. The green arrows represent the change in levels following a correct response while the red arrows show the changes following an incorrect response.

41]. It is based on the forgetting curve first postulated by Ebbinghaus [42], which describes the probability of successful recall of learned information as a function of time passed since the last recall. With each additional recall, the memory trace is assumed to become stronger. Hence, the forgetting curve becomes flatter, meaning that the probability of successful recall decays less steeply with passing time. Consequently, the optimal interval between recalls id assumed to increase exponentially.

Spaced repetition is applied in Dawn of Civilisations in form of an adaption of the Leitner-system, which was originally designed for learning with flashcards [40, 43]. While flashcards are put in distinct boxes, the question in DoC are assigned a Question-Recycler level (QR-level). A question enters the system the first time it is encountered by the learner. If it is answered correctly, it gets assigned a QR-level of 2. If it is not answered correctly, it gets assigned a QR-level of 1. Once in the system, the QR-level of a question will increase by 1 every time it is answered correctly and decrease by 1 every time the learner gets it wrong. The delay interval that needs to pass before the question is presented to the learner again increases exponentially depending on the question's QR-level. It ranges from one day for questions at level 1 to 16 days for questions at level 5. Returning to the question opening this paragraph, a question is considered complete - and its content mastered by the learner - when it reaches a QR-level of 5. Once a question reaches the completion level, its QR-level will not decrease again, even if it is answered incorrectly. As an exception to this rule, questions belonging to the first two sub-levels (called "Pre-A") are considered completed once they reach a QR-level of 3.

The described process is summarized in Figure 3.3.

It is important to note that the review schedule as it has been described here is an idealised version. During actual game play, the review intervals additionally depend on how often a user plays minigames, how many they play in a row on days they

decide to play and which games they choose to play when they play. Intervals may become longer if a user only plays occasionally. They might become shorter when they play only a small subset of minigames and those rather excessively. In such a case, a minigame can out of questions with expired delays and has to present questions which are not scheduled for review yet.

In summary, Dawn of Civilisations is an education environment training 6 different skill aspects of learning English which are measured at 16 levels of competence. The levels of competence are determined by the number of relevant questions which have been mastered by the learning according to the rules of the QR-System.

3.3.2 Learning Goal

Dawn of Civilisations itself does not define specific learning goals for its users. Based on the CEFR levels for assessing someone's language proficiency, balanced learning goals were defined for the brain points approach [39]. A balanced learning goal means that the same goal competence value is chosen for all six skills. As competence values can assume 16 distinct values the way they are measured by DoC, 16 learning goals are defined. For a learner starting out without any prior skills would therefore have $g = (1, 1, 1, 1, 1, 1)$ as their learning goal. A learner whose current skill competence values are $\boldsymbol{c} = (2, 3, 1, 1, 2, 2)$ would have $\boldsymbol{g} = (2, 2, 2, 2, 2, 2)$ as their learning goal. Throughout the application of this model, a sequence of goals was considered. That means that once a goal is reached, the learning continued with the next-highest goal until the highest possible goal $\boldsymbol{g} = (16, 16, 16, 16, 16, 16)$ was reached.

In principle, any combination of skill competence levels could be set as a learning goal. For example, some learners might have reasons to want to advance their speaking skills more than their writing skills. However for this first development and evaluation of the feedback method, balanced goals were deemed to be the most generally beneficial learning goals across all learners.

3.3.3 Action Space

Each action in the action space represents the choice to engage in one of the minigames Dawn of Civilisation offers to its users. The action space $\mathcal{A}$ of Dawn of Civilisation therefore comprises 16 actions. Different skill combinations are taught in the different

Module	Skills / Mini Game	Vocabulary	Grammar	Reading	Writing	Listening	Speaking
1	Fix the Mix	■			■		
2	Speak Racer	■					■
3	Copy Cat	■				■	
	Robot Factory	■				■	
	Flying Robot	■				■	
	Copy Parrot	■				■	
	Typing Time	■				■	
	Factory Card	■				■	
4	Brain Battle		■	■			
5	Word Snap		■	■			
6	Hello Cafe		■	■	■	■	■
	Chat Time		■	■	■	■	■
	Tic Talk		■	■	■	■	■
7	Eye Spy						■
8	Judge Me!		■	■			
9	Talk n' Go						■

Figure 3.4: Skills trained by different mini games: Add Description

minigames, which are visualized in Figure 3.4. Each minigame pertains to one of 9 modules. Minigames which belong to the same module present the same questions.

3.3.4 State Space

In Section 3.1 it was stated that a state s_t can be defined based on the parameters θ needed to compute the competence values of each skill at time t. In the case of Dawn of Civilisations, the competence values are a function of the QR-levels of the relevant questions, as has been explained in Section 3.3.1. With $\mathcal{Q}_{c_{i,j}}$ being the set of questions relevant to the ith skill at competence level j and $\mathcal{H}_{c_{i,j}}$ being the set of completed questions thereof, c_i is defined in Eq 3.10.

$$c_i = \max j \in \{1, 2, ..., 16\} \quad \text{subject to} \quad |\mathcal{H}_{c_{i,j}}| >= 0.8 * |\mathcal{Q}_{c_{i,j}}| \tag{3.10}$$

Furthermore, the amount of time left until a question's scheduled review and the number of games played on the current day n_d are relevant to differentiate states as they influence the transition function detailed in Section 3.3.5.

Consequently, with regard to the definition of the states in Equation 3.3, θ could be all the questions' QR-levels while β could be all the questions remaining delays and

the number of games played on the current day. However, in order to reduce the complexity of the state space, an abstraction is applied to this representation [44]. Instead of including each question's QR-level and remaining delay in the state representation, the number of questions with a specific QR-level and delay are counted at each competence level and for each module. For the first two levels, for which the maximal QR-level equals 3, and thereby the maximal interval period equals 4 days, there are 11 such categories per module per level, resulting in $2 * 9 * 11$ parameters. For the remaining 14 levels, with a maximal QR-level of 5 and intervals ranging from 0 to 16 days, there are 37 possible states a question can be in, resulting in additional $14 * 9 * 37$ parameters. Including n_d, the resulting states have a length of 4861.

The resulting number of parameters defining a state in the MDP are calculated in Equation 3.11.

$$|s| = 2 * 9 * 11 + 14 * 9 * 37 + 1 = 4861 \tag{3.11}$$

3.3.5 Model of Skill Improvement

As laid out in Section 3.1, the way in which an educational environment defines skill improvement governs the transition probability $P(s'|s, a)$ of the MDP. That is the probability of ending up in state s' if action a is taken in the current state s. Dawn of Civilisation's model of skill improvement builds on the QR-system described in Section 3.3.1.

When a user chooses a minigame to play, the questions to be presented by that minigame are selected from the pool of potential questions by their priorities. Highest priority is given to questions that have not been asked yet. Second highest priority is given to questions which are not yet completed and scheduled for review, in ascending order of their QR-level. Third highest priority is given to questions which are not yet completed within the delay period, ordered by their closeness to the end of it. Lastly, completed questions are asked, ordered as well by their delay status. The number of questions presented by each game n_q varies between minigames. Also, the same minigame will not necessarily present the same number of questions each time it is played. Therefore, discrete probability distributions $P(n_q|a)$ were derived from the user data to estimate how probable a minigame is to present n_q questions (see Figure 3.6). As the probabilities were estimated across different levels, which was done because data was not available for each level, $P(n_q)$ only depends on a.

In addition to determining how many and which questions are impacted by action a in state s, an estimate of how likely those questions are to be answered correctly is needed in order to calculate $P(s'|s,a)$. To that end, the probability $P(\text{correct}|a)$ has been calculated from the user data (see Figure 3.5). Ideally, the probability of a correct answer would have been conditioned on several factors apart from the chosen minigame, such as level of competence and QR-level of the question. That would allow us to get a more precise estimate of $P(\text{correct}|a)$ and consequently a more precise estimate of $P(s'|s,a)$. However, the data did not provide enough examples to do so, and conditioning only on the minigame was the feasible option. Answering each question correctly is assumed to be an independent event. Consequently, the probability of observing an answer pattern z to all n_q questions is defined in Eq. 3.12, with x being the number of correct questions in z. $K(x,n_q)$ expresses the number of combinations of answers to n_q questions leading to the same number of correct answers x.

$$P(z|x,a,n_q) = P(\text{correct}|a)^x * (1 - P(\text{correct}|a))^{n_q - x} * K(x,n_q) \tag{3.12}$$

$$K(x,n_q) = \begin{cases} \frac{n_q!}{x! * (n_q - x)!}, & \text{if } x < n_q \\ 1, & \text{if } x = n_q \end{cases} \tag{3.13}$$

The final aspect of the transition between states is independent from the action a, as it captures how the delay until the scheduled review of the questions is decreased by the passage of time. Two estimations were made from the user data in order to address this issue. Firstly, how likely a user is to play a certain number of in a single day, $P(N_d)$. Secondly, how likely it is to observe a delay b between days on which a user plays any minigame, $P(b)$. Based on that information, the probability that the minigame played is the last of the day, $P_{n_d}(N_d = n_d)$, can be estimated given the number of games already played at the current day n_d, which is included in state s. Specifically P_{n_d} is the discrete probability fit to the data on how many minigames are played by day excluding values $< n_d$. That captures the probability of observing a certain number of games played knowing how many already have been played.

Putting it together, the probability of a change in delays w can be calculated from s as defined in Eq. 3.14.

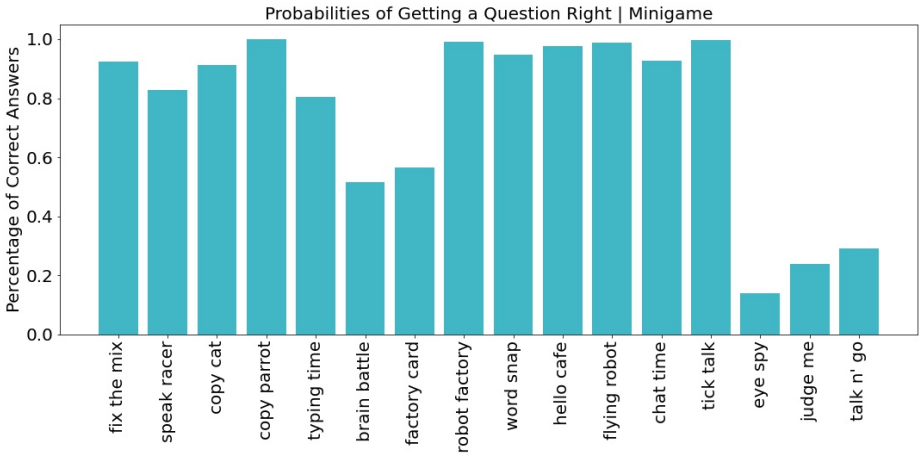


Figure 3.5: Success Probabilities per minigame

$$P(w|s) = \begin{cases} 1, & \text{with } p = 1 - P_{n_d}(N_d = n_d) \\ P(b), & \text{with } p = P_{n_d}(N_d = n_d) \end{cases} \tag{3.14}$$

$$P(s'|s,a) = P(n_q|a) * P(z|x,a,n_q) * P(w|s) \tag{3.15}$$

3.3.6 Reward Function

In order to construct an appropriate reward function (see Eq. 3.7) for DoC, an estimate of the effort required to complete the minigames is needed. As an available measure of effort, the times needed for completion were extracted from the data provided by Solve Education. An overview is provided in Figure 3.8.

For each action a, a log-normal distribution is fit to the data, from which rewards can be sampled.

$$P(r_a) = p(\frac{1}{r_a \sigma_a \sqrt{2\pi}} \exp\left(-\frac{(\ln r_a - \mu_a)^2}{2\sigma_a^2}\right)) \tag{3.16}$$

The reward r_a is transformed to represent the action cost by multiplying it with -1. When an action leads to a state that satisfies the current learning goal g, a positive reward is granted. As it would be out of scope of this project to find a way to realistically express the usefulness of reaching a certain skill level in English numerically, an arbitrarily high positive reward $r_g = 2000$ was set for all g.

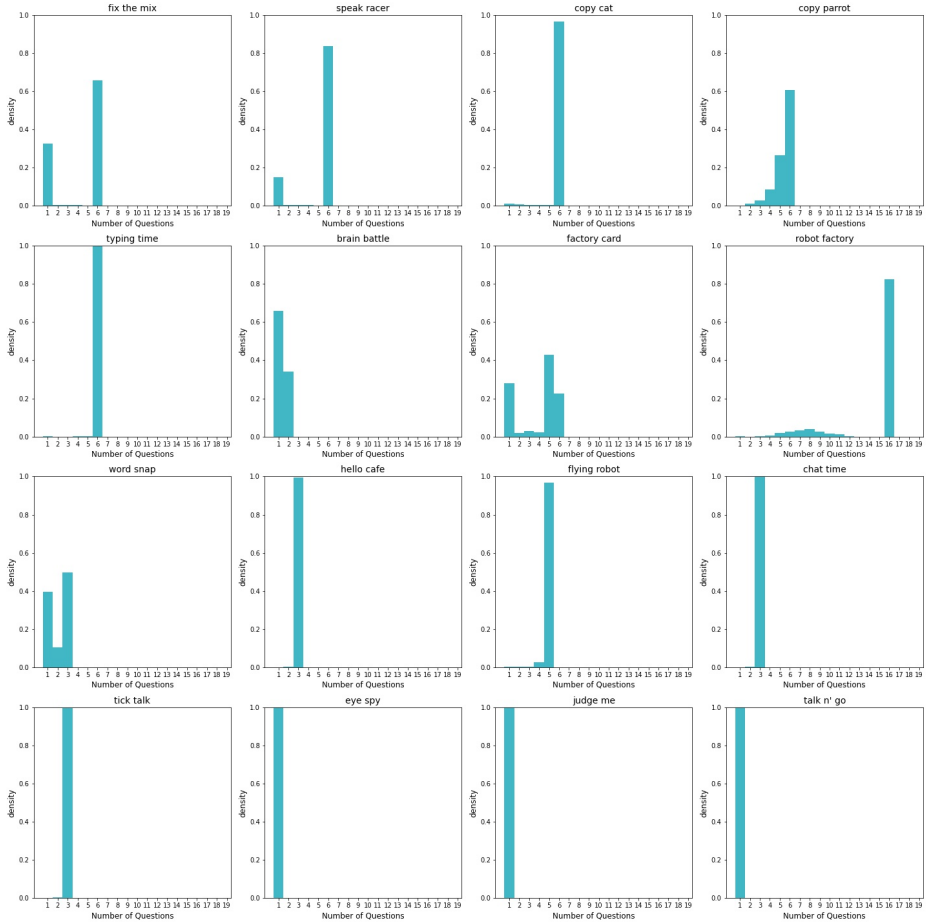
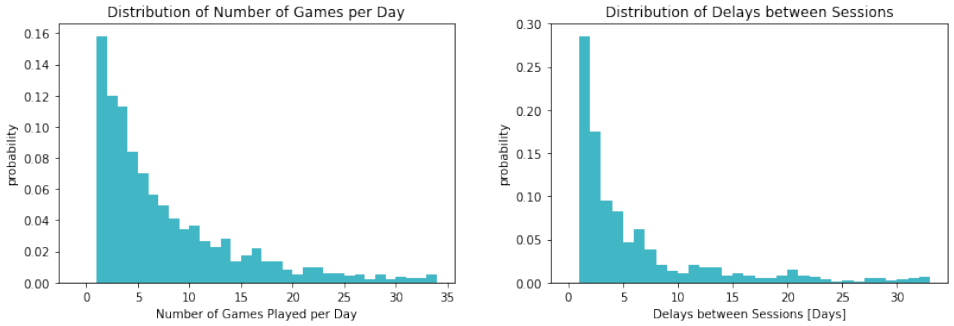


Figure 3.6: Number of questions presented per minigame



(a) Minigames played per day

(b) Delays between game days

Figure 3.7: Parameters for including time in transition function

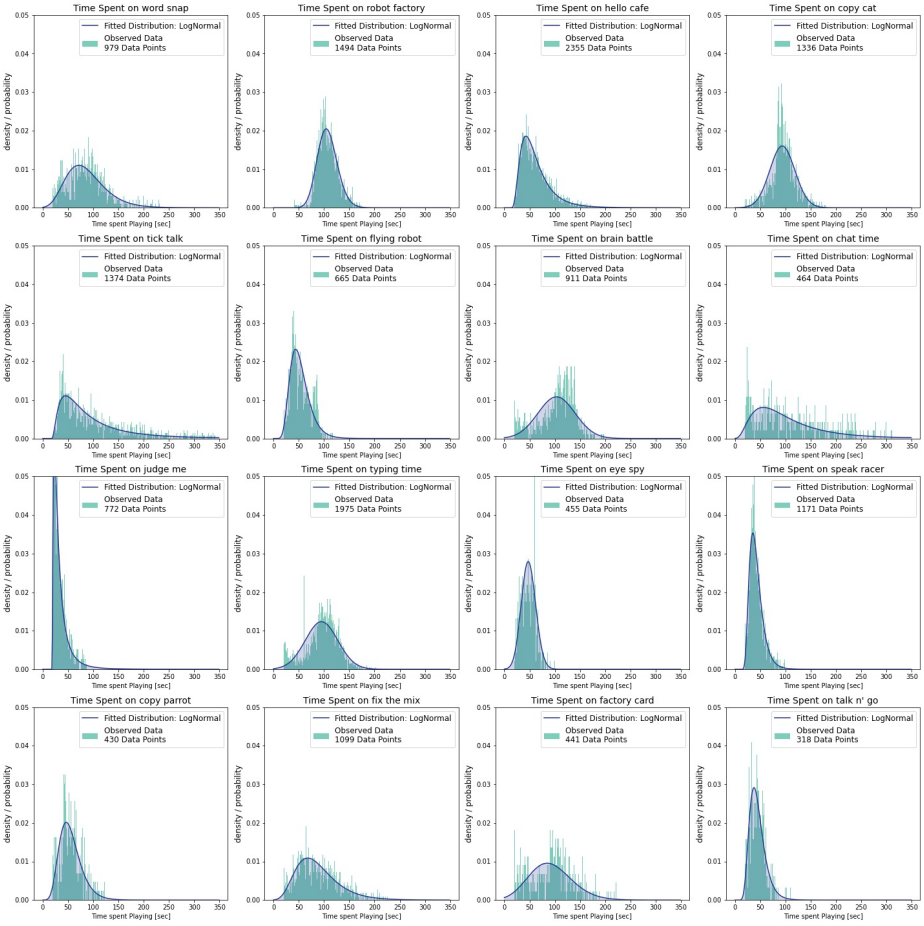


Figure 3.8: Time spent on each mini game: Add Description

Consequently, with r_a as defined in Eq. 3.16 and $\boldsymbol{c}$ calculated from s as specified in Eq. 3.10, the reward function of the MDP constructed to represent DoC is shown in Eq. 3.17.

$$R(s', a, s) = \begin{cases} 2000 - r_a, & \text{if } d(\boldsymbol{c}', \boldsymbol{g}) \wedge \neg d(\boldsymbol{c}, \boldsymbol{g}) \\ -r_a, & \text{otherwise} \end{cases} \tag{3.17}$$

This concludes applying the model for choosing educational activities to Dawn of Civilisations. All specific parameter values that have only been shown visually in the figures throughout this section can be found in the repository linked to in Appendix A, along with the implementation of the described formulas. In the next section, the path from the model to the brain points method is described in detail.

3.4 Deriving a Method for Calculating Brain Points for Dawn of Civilisations

The reason for modeling the choice of minigame in DoC as a MDP was to be able to define a pseudo-reward function as Eq. 3.9 such that a myopic learner maximizing short-term reward maximizes their long-term learning progress as well. In terms of Dawn of Civilisations, that means assigning brain points in a way that a learner purely motivated by gaining rewards for their city development still chooses minigames in a way that optimally benefits their English progress as well. Because brain points are meant to incentivize choices and are therefore displayed before the action is carried out, they need to be a function of the state s and the action a and cannot depend on the next state s'.

3.4.1 Defining and Solving a Simplified Model for Benchmarking

The state action space of the MDP modeling choosing between minigames in Dawn of Civilisations is too large to determine the optimal policy π^* analytically. Therefore, as a first step, a small and simplified subset of the game was modeled to serve as a benchmark throughout the development of the approach. Specifically, the number of minigames to choose from was limited, as were the number of questions. Furthermore, only the first two levels were included and the assumption was made that users play one minigame a day everyday. An overview of the changes made in order to obtain the simplified model is given in Table 3.1.

Parameter \ Model	Original	Simplified
Number of Questions	8480	23
Number of Competence Levels N_c	16	2
Number of minigames |A|	16	3
$P(N_d = 1)$	<1	1
$P(b = 1)$	<1	1
|s|	4861	126
|S|	unknown	94657
$P(n_q = x\|a)$	<1	1

Table 3.1: Changes in parameters for simplified model

The simplifications allowed the MDP to be solved using value iteration, an algorithm for determining the optimal state-value function $V^*(s)$, which has been described in Section 2.2.2. The discount factor γ was set to 0.9 to account for the possibility that a learner might loose interest and stop interacting with the minigames before reaching their learning goal. The convergence threshold δ was set to $1e-7$. Having obtained $V^*(s)$, it was possible to calculate optimal brain points $OBP(s,a)$ as definded in Eq. 2.16.

Figure 3.9 shows the reward achieved on average by a myopic agent receiving optimal brain points in comparison with an agent choosing random actions and a myopic agent without pseudo-rewards.

With that, a benchmark had been established which was used to guide the attempts to develop a method to approximate brain points for the original model and thereby for the original game. Those attempts are described in the next section.

3.4.2 Trying to Approximate a Useful State-Action Value Function

As a starting point, a handcrafted policy (HCP) was developed based on the author's insights into the game's mechanics. Its average return in the simplified model is shown in Figure 3.9 and its procedure is described in Algorithm 3. It can be seen that it performs close to optimal in the simplified MDP, leading to the following idea: Approximate the state-action-value function of the handcrafted policy $Q_{HCP}(s,a)$ and calculate brain points as described in Eq. 3.18.

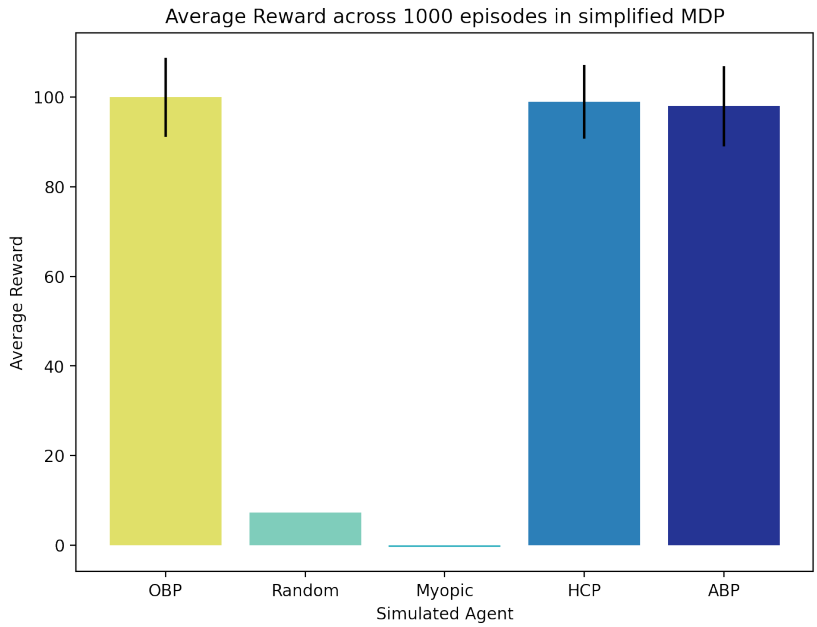


Figure 3.9: Comparison of Simulated Agents in simplified DoC-MDP: Depicted are the rewards obtained by simulated agents averaged across 1000 episodes. The error bars represent standard deviation. OBP denotes a myopic agent provided with optimal pseudo-rewards (Eq. 2.16). The myopic agent selects actions greedily while looking one step ahead without any pseudo-rewards. The random agent selects actions randomly and the HCP-agent follows the handcrafted policy defined in Alg 3. ABP labels a myopic agent provided with approximate pseudo-rewards (Eq. 3.20).

Feature	Description	β_x
$F_1(s,a)$	Number of questions with expired delay available for a	8
$F_2(s,a)$	1 if F1(s,a) median number of questions presented by a , 0 otherwise	$7 * 10^2$
$F_3(s,a)$	1 if any skill trained by a has not reached current learning goal, 0 otherwise	$3 * 10^{-12}$
$F_4(s,a)$	1 if any skill trained by a is (one of the) least developed skills, 0 otherwise	$-1 * 10^{-12}$
$F_5(s,a)$	Number of questions necessary to complete for skills trained by a to reach next skill level	$-9 * 10^1$
$F_6(s,a)$	Number of completed questions for current skill level for skills trained by a	$3 * 10^2$
$F_7(s,a)$	Mean QR-level of questions for current skill level for skills trained by a	9

Table 3.2: Overview of handcrafted features used to approximate $Q(s,a)$. The rounded coefficients resulted from linearly regressing the discounted sum of future rewards

$$BP(s,a) = Q_{HCP}(s,a) - \max_a \; Q_{HCP}(s,a) \tag{3.18}$$

To that end, we defined features to capture essential characteristics of the state and action pairs, inspired by the information that the handcrafted policy uses. An overview of the used features is given in Table 3.2. As the next step, a large number of samples were obtained from the interaction of an agent following the hand-crafted policy and the MDP. To diversify the samples, starting states were randomly selected by letting a randomly choosing agent interact with environment for a randomly chosen number of time steps prior to commencing sample collection with the hand crafted policy. Those samples consisted of the feature vector extracted from the states and actions and the discounted sum of future rewards obtained after taking the action in that state. Subsequently, a linear regression using an ordinary least squares estimator was fit to the obtained data set, using the feature vectors as the predictive variables and the discounted sum of future rewards as the regressand [45]. The resulting coefficients were used to approximate $Q(s,a)$. That means that the features were extracted from the state s and the action a and then multiplied with the coefficients to get a prediction of the future sum of rewards, representing the value of taking action a in state s. From there, brain points were calculated as defined in Eq. 3.18.

For the simplified model, approximating $Q_{HCP}(s,a)$ linearly with handcrafted features led to satisfactory results, meaning that a myopic agent receiving the brain points derived trough linear approximation performed as well as one receiving the optimal brain points. The coefficients for the simplified MDP can be found in rightmost column in Table 3.2. The intercept was -490.

Unfortunately, this approach did not scale to the original MDP, meaning that brain

Feature	Description
$F_8(s,a)$	Maximal Skill level - average skill level for skills trained by a
$F_9(s,a)$	Number of questions available at current level for a
$F_{10}(s,a)$	Number of questions with expired delay and QR-level = 0 for a
$F_{11}(s,a)$	Number of questions with expired delay and QR-level = 1 for a
$F_{12}(s,a)$	Number of questions with expired delay and QR-level = 2 for a
$F_{13}(s,a)$	Number of questions with expired delay and QR-level = 3 for a
$F_{14}(s,a)$	Number of questions with expired delay and QR-level = 4 for a
$F_{15}(s,a)$	Number of questions with expired delay and QR-level = 5 for a

Table 3.3: Overview of additional handcrafted features used to approximate $Q(s,a)$

points calculated via the same approach for the complete DoC MDP led a myopic agent to not reach the learning goal. Therefore, the feature list was expanded as shown in Table 3.3 and it was tried to capture characteristics better by including interactions between selected features. Specifically, the products of each of the feature values and $F_3(s,a)$ as well as $F_9(s,a)$ were included, leading to a feature vector of length 43. These interactions were chosen because if either all skills trained by a have already reached the learning goal levels or a offers no questions ready to review, it is clear that taking action a will not increase the competence values. For such (s,a) pairs, those products become zero, expressing that insight numerically.

Additionally, distributional shift was identified as a potential issue hindering a sufficiently precise approximation of $Q(s,a)$ for parts of the state-action space not visited by the hand-crafted policy albeit the efforts to vary the starting state [46]. Therefore, samples collected from random and myopic action selection were included in the data set as well.

The regression of this new data set showed that the extended features could predict some of the variability of the sums of discounted future rewards, with a coefficient of determination $R^2 = 0.37$. The adjusted R^2 was used to avoid inflation due to the large number of features. An evaluation of the resulting brain points showed that this was definitely not a sufficiently precise approximation, as a myopic agent acting upon those brain points did not reach any learning goals.

In order to more appropriately address the complexity of the state-action space and potential non-linearities in $Q(s,a)$, a deep Q-learning (DQN) approach as described in Section 2.2.4 was employed next [21]. Based on encouraging findings regarding the combination of handcrafted features with deep learning in different domains [47, 48], both the raw state representation and the handcrafted features were used as

input to the DQN. Additionally, transitions (s, a, r, s') gathered by HCP, as well as randomly and myopically, were used to populate the DQN's replay memory with the goal to facilitate learning by guiding exploration. This approach was subsequently abandoned due to two reasons. First, learning was not initially successful and the time frame for this thesis did not allow for more extensive parameter tweaking. Second and more importantly, one of Dawn of Civilisations' key characteristics is its ability to run on low-end smartphones. Even if increasing the network's size would have possibly led to a successful approximation of $Q_{HCP}(s, a)$ or even $Q^*(s, a)$, the resulting method would not have a practical use.

Hence, the focus was shifted away from approximating a useful state-action value function to calculate brain point from and towards directly approximating the brain points themselves.

Algorithm 3 Hand Crafted Policy

```
procedure HCP(s, A)
    c = (c_1, c_2, ... , c_{N_S}) = f(s)
    sort c
    for skill in c do
        for a ∈ A do
            if a trains c and a has priority 1 or 2 questions then return a
    for skill in c do
        for a ∈ A do
            if a trains c then return a
```

3.4.3 Directly Approximating Brain Points

The key element for approximating brain points directly is a potential function $\phi(s)$ [32]. The chosen potential function (Eq. 3.19) expresses the progress made towards the learning goal g. This is achieved by including the ratio of the summed QR-levels for all questions relevant to reaching the goal competence level for a skill and the summed maximal QR-levels for the same set of questions. The level value $g_i - 1$ is added to account for between-level progress in addition to the within-level progress captured by the ratio. The resulting values are summed across all skills in order to get an overall progress estimate for the state. The higher that sum, the closer the state is

to a state fulfilling g.

$$\begin{aligned} \phi(s) &= \sum_{i=1}^{N_s} \left[(g_i - 1) + \frac{\sum_{q \in \mathcal{Q}_{g_i - 1}} l(q)}{\sum_{q \in \mathcal{Q}_{g_i - 1}} m(q)} \right] \\ &\text{with} \\ \mathcal{Q}_{k_i} &= \text{Set of questions training skill } i \text{ at level } k \\ l(q) &= \text{QR-level of q} \\ m(q) &= \text{maximal QR-level of q} \end{aligned} \tag{3.19}$$

Having defined $\phi(s)$, approximated brain points $ABP(s, a)$ can be derived as shown in Eq. 3.20.

$$ABP(s, a) = \sum_{s'} P(s'|s, a) R(s, a, s') + \left(\max_{s'} \phi(s') - \phi(s) \right) \tag{3.20}$$

Noticeably, we made a significant change to the established way of defining the shaping function $F(s, a, s') = \gamma\phi(s') - (s)$ by taking the maximal value of $\phi(s')$ instead of its expected value across all possible next states [1, 32]. Due to this modification, the maximal possible benefit of taking action a in state s is taken into account while possible disadvantageous outcomes are not. This decision was made out of consideration for the context the model and the approximated brain points are to be applied to. Specifically, we deemed it important to not penalize choices in favor of hard minigames which might result in a decrease in QR-levels and therefore a less beneficial state. An ideal model could capture learning from failures by adapting $P(\text{correct}|a)$ to $P(\text{correct}|a, \omega)$ and for example include the number of previous encounters with the learning material in ω. Since the data the model was built upon did not allow for such fine-grained estimations, the concept is integrated into the approximated brain points instead. This was not a trivial decision, because the discussed modifications entails that the approximated brain points do not adhere to the shaping theorem and therefore a policy maximizing them is not guaranteed to also maximize long-term rewards obtained from the original MDP [32]. To reiterate, while a more sophisticated model would allow us to approximate brain points in agreement with the shaping theorem, under the given circumstances it was deemed a satisfactory solution to instead rely on using the maximal possible progress in the approximation in order to come to a solution. An encouraging piece of evidence for this decision is the comparable performance of the handcrafted policy and a myopic agent acting

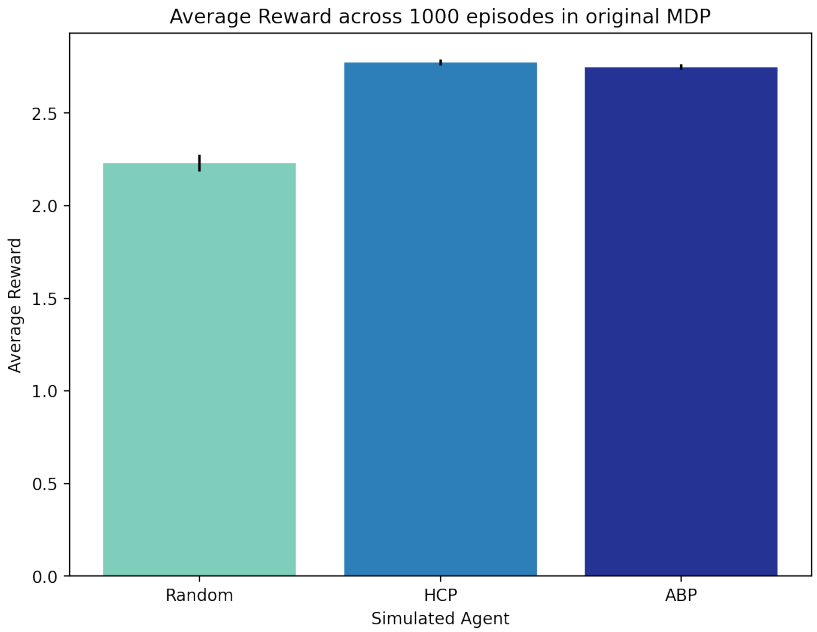


Figure 3.10: Comparison of Simulated Agents in original DoC-MDP: Depicted are the rewards obtained by simulated agents averaged across 1000 episodes. The error bars represent standard deviation. The random agent selects actions randomly and the HCP-agent follows the handcrafted policy defined in Alg 3. ABP labels a myopic agent provided with approximate pseudo-rewards (Eq. 3.20).

upon the approximated brain points in both the simplified and the original MDP shown in Figures 3.9 and 3.10. In the end, the evaluation with real learners has to show whether the brain points are beneficial to the learners even if they are not optimal.

4 Evaluation of the Brain Points Method in a Controlled Online Experiment

In this chapter, the evaluation of the approximated brain points derived in the previous chapter is described. That evaluation entailed the construction of an experimental learning environment with the help of two pilot experiments, described in Section 4.1. Following that, the design of the experiment with the environment and the analysis of the obtained data are presented in Section 4.2.

4.1 The Experimental Learning Environment

In order to evaluate the effect of presenting brain points to learners choosing between different learning tasks in an experiment with limited duration, an experimental learning environment had to be designed. The goal of this process was to strike a balance between applying the necessary simplifications and keeping some fundamental characteristics of Dawn of Civilisations well represented.

First of all, a learning environment needs learning material. A paired associative recognition task was chosen due to its widespread application in researching mechanisms of learning [49, 50, 51]. That means that participants are tasked to memorize associations between stimuli. As the task was designed as a recognition and not a cued-recall paradigm, participants had to recognize whether two stimuli presented to them were associated or not. In order to better represent learning a new language and to avoid confounding effects of participants language skills, non-words were chosen as the stimuli [52].

Besides the learning material itself, another aspect to be considered is how to construct different tasks from it. It is important that learners are not indifferent towards the choice between tasks, because that would hinder the evaluation the effectiveness of the incentives in helping them overcome short-sighted decision strategies. In Dawn

of Civilisations, users may prefer some minigames over others due to a large array of reason, be it their graphics, their game mechanics or their content. All these factors are difficult to replicate in a short and simple experiment. One crucial factor likely contributing to the appeal of a minigame is its difficulty. As elaborated in Section 2.1.5, people's tendency to feel aversion towards failure often leads them to seek out easy tasks [1, 2]. Therefore, the decision was made to vary the difficulty of the learning tasks between minigames in the experimental learning environment. Easier said than done, as the literature on paired-associate learning of non-words typically asks and answers other types of research questions than if increasing the number of letters in a non-word facilitates or impedes memorizing it.

Consequently, potential learning stimuli were tested beforehand to find sets of stimuli of varying difficulty. It is important to note that the purpose of these experiments was to identify a suitable set of stimuli and specifically not to test hypothesis regarding what factors make paired associate learning more or less difficult. On top of finding a way to manipulate difficulty between learning tasks, the pretests can also provide an estimate of $P(\text{correct}|a)$, which is crucial to calculating and approximating brain points for the experiment.

In the following, the two experiments run in order to test potential learning stimuli are described in detail.

4.1.1 Testing Learning Stimuli: Experiment 1

As a first step, three potential factors influencing difficulty were tested against a baseline. One idea was that increasing the number of non-word pairs to be memorized might hinder performance. The second idea was that people might find shapes more difficult to memorize than non-words. The third idea was to increase the difficulty by decreasing the distinguishability of valid and invalid pairs.

Methods

Design Each of the three experimental conditions (see overview in Table 4.1) investigated one possible difficulty manipulation in a within-subject design. The stimuli list used as the baseline consisted of 6 pairs of 4-letter non-words. The baseline list was the same across all three conditions. The first experimental manipulation, labeled "Set Size" concerned the size of the list to be learned and therefore encompassed 10

pairs of 4-letter non-words. The second condition, "Visual", featured 6 pairs consisting each of an abstract shape and a 4-letter non-word. The third condition differed from the baseline in the way the 6 pairs were constructed. Instead of pairing one non-word of each associated pair with a new non-word to form the non-associated pairs, the non-associated pairs were re-combinations of the same 6 non-words used in the associated pairs. This condition was labeled "Distractors". Across all conditions, half of the stimuli were assigned to be associated and the other half to be non-associated pairings.

The dependent variable was the percentage of correct classification achieved by participants.

Participants 57 (70.7% female; *mean age* = 23.9 years, *SD* = 6.32 years) participants were recruited via the online recruitment platform Prolific [53]. All participants provided written informed consent and received a base compensation of 0.75 £ for the 8 minute experiment. Additionally, they were awarded a bonus payment based on their performance, with the average bonus set to be 0.25 £. All participants were older than 18 years and fluent in English.

One participant was excluded according to the predefined exclusion criteria. This means that they failed both of the included attention checks. Each attention check consists of the message "It is important that you stay attentive throughout the experiment. Please press p to continue" and were considered to have failed if the participant corresponded with one of the keys used to communicate their response to the regular trials, indicating trying to skip trough the trials without paying attention to what was written on the screen.

This and the following experiment were covered by the ethics approval from the IEC of the University of Tübingen under IRB protocol number 667/2018BO2.

Materials and Procedure The experiment was programmed using jsPsych [54], a framework for running behavioural online experiments. The implementation was based off a "starter pack" provided by [55]. The non-words used as stimuli were randomly sampled from the ARC Nonword Database [52]. The shapes used as stimuli in the "Visual" condition were adapted from Clayton et al. (2018) [49]. Table 4.1 details the exact stimuli pairs used in each condition.

Participants were randomly assigned to an experimental condition. The experiment was divided into two blocks - the baseline block and the experimental block, which

	Baseline	Set Size	Visual	Distractors
Valid	dwor - zuik prus - ceaf gheg - kump	clee - vafe smar - cilp ulch - grov cauv - urbe fusk - tarb	 - clee - vafe - smar	clee - vafe smar - cilp ulch - grov
Invalid	dwor - chom prus - pefe gheg - skra	clee - demb smar - soys ulch . tovs cauv - gyte fusk - kilv	 - cilp - ulch - grov	clee - grov smar - vafe ulch - cilp

Table 4.1: Stimuli First Pretest

were counterbalanced and had a break in between. Each block consisted of 5 repetitions of the corresponding stimuli list. Within each repetition, the order of stimuli pairs was randomized.

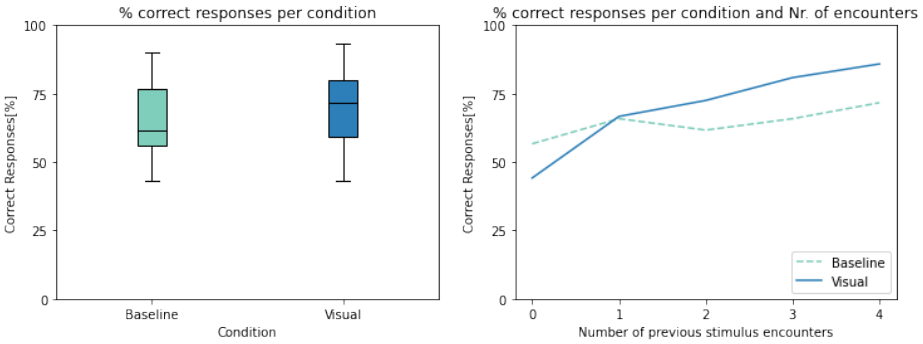
The stimuli pairs were presented to the participants on the screen and they reacted to them with a key press indicating whether they thought it was a valid or invalid pairing. After the reaction, corresponding feedback was shown.

Results

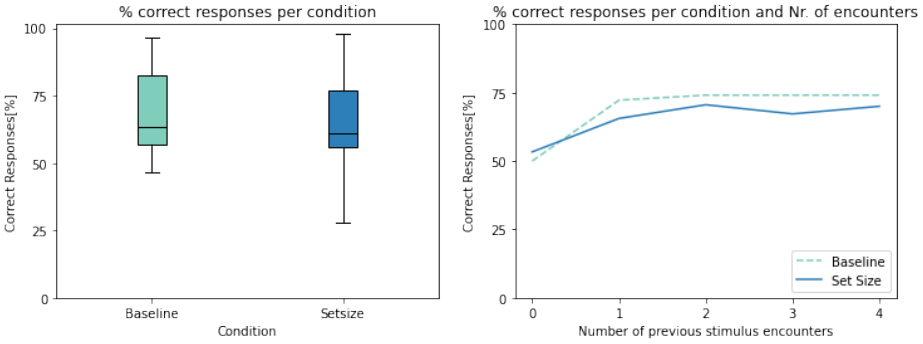
The within-subject design used resulted in paired samples for each conditions. Therefore, the Wilcoxon signed-rank test [56] was used to test for differences of the percentage of correct classifications between each of experimental conditions and the baseline. In order to address the multiple comparison problem, α was adjusted following the Bonferroni correction: $\alpha = \frac{0.05}{3} = 0.017$. [57]. The results are reported in Table 4.2 and Figure 4.1. None of the experimental manipulations of difficulty elicited a significant difference in the percentage of correct classifications. Consequently, a second experiment was designed so as to find a working manipulation of difficulty for constructing the experimental learning environment.

Contrast	Z-value	p-value	Cohen's d
Baseline - Visual	70.0	0.314	-0.12
Baseline - Set Size	82.0	0.879	0.00
Baseline - Distractor	83.0	0.913	0.08

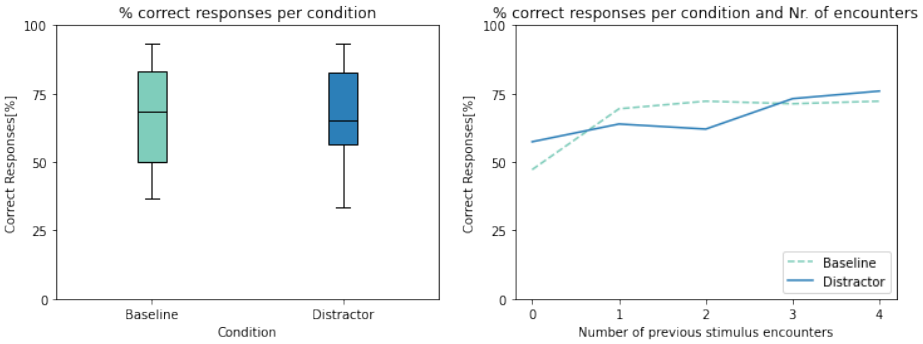
Table 4.2: Results First Pretest



(a) Visual Condition



(b) Set Size Condition



(c) Distractor Condition

Figure 4.1: Results of first Stimulus Pretest: The left columns shows the percentages of correct classification across all trials. The right columns shows the percentages depending on how often participants had already seen the stimuli pair.

4.1.2 Testing Learning Stimuli: Experiment 2

Seeing the results of the first experiment, the size of the baseline list was decreased in the hope of making it easier to master. Out of the same considerations, one set with more pairs than the maximum of ten in the previous experiment was tested. As a second idea, more complex stimuli consisting of paired non-words which in turn were combined with other paired non-words into the valid and invalid pairs were used. Thirdly, it was tested whether making the non-words themselves more similar to each other might impact difficulty.

Methods

Design As in the first experiment, each of the three experimental conditions (see overview in table 4.3) investigated one possible difficulty manipulation in a within-subject design. The stimuli list used as the baseline was reduced to 4 pairs of 4-letter non-words. The size of the list to be learned in the "Set Size" condition was increased to 16 pairs of 4-letter non-words. For the "Doubles" condition, each of the 8 stimuli pairs itself consisted of pairs of non-words. In order to construct the invalid pairs, the second non-words of the double stimuli were swapped. The third condition, termed "Similarity", comprised 2 sets of 4 stimuli pairs. Each of those sets was created by starting off with a pair of non-words from which an additional one was created by changing one consonant in each of its components. Recombining the resulting non-words yielded the two corresponding invalid pairs.

The dependent variable was again the percentage of correct classifications achieved by participants.

Participants 60 participants (67.8% female; *mean age* = 25.5 years, *SD* = 6.96 years) were recruited via the online recruitment platform Prolific [53]. All participants provided written informed consent and received a base compensation of 0.50 £for the 6 minute experiment. Additionally, they were awarded a bonus payment based on their performance, with the average bonus set to be 0.25 £. All participants were older than 18 years and fluent in English. With the use of Prolific's prescreening tool, it was ensured that no participants from the first version of the experiment took part again. Two participants were excluded according to the predefined exclusion criteria, which were the same as for the first experiment.

	Baseline	Set Size	Doubles	Similarity
Valid	fipt - bonk shec - rukt	zonz - nylk fubb - cwob veav - knyz murt - yoes fost - zict wope - filf nooc - twes nels - spyc	preg hilv - zubs qued coaz rert - deec nibe fubb cwop - fost zict nyln klaz - soag yesc	vomp - ancs vonp - anzs feph - gwug fegh - grug
Invalid	fipt - dynk shev - alvs	zonz - zict fubb - filf veav - twes murt - spyc fost - nylk wope - cwop nooc - knyz nels - yoes	preg qued - zubs hilv coaz nibe - deec rert fubb zict - fost cwob nyln yesc - soag klaz	vonp - ancs vomp - anzs feph - grug fegh - gwug

Table 4.3: Stimuli Second Pretest

Materials and Procedure The experiment was again programmed based on [55] using jsPsych [54] and the non-words randomly sampled from the ARC Nonword Database [52]. For the "Similarity" condition, 4 non-words were sampled from the database and manually altered to create stimuli differing only in a single consonant. Table 4.3 details the exact stimuli pairs used in each condition.

The basic procedure was the same as for the first experiment. Two changes were made to address previous oversights. Firstly, the order in which the non-words forming a pair were presented was randomized to ensure that the participant had to learn the association and not just the second stimulus to produce correct responses. Secondly, a response time limit of 3.5 seconds was imposed and the feedback shown for 2 seconds. These changes were meant to prevent participants from noting down the stimuli pairs in order to achieve a higher bonus payment or to skip the feedback for the sake of finishing the experiment more quickly.

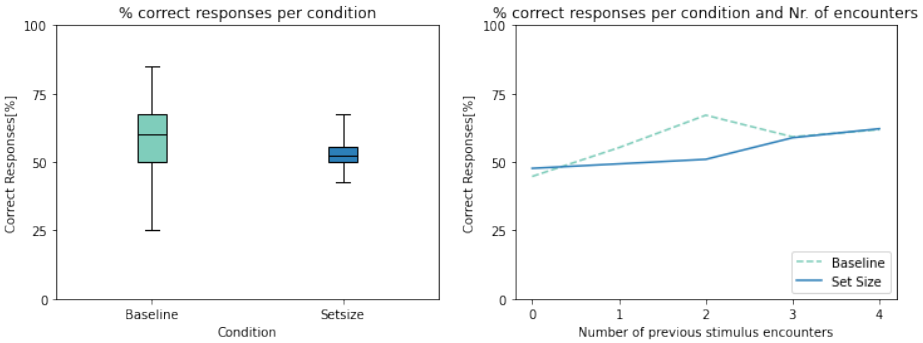
Results

The results were obtained in the same manner as described for the first experiment and are reported in Table 4.4 and Figure 4.2. The list of more similar non-word pairs

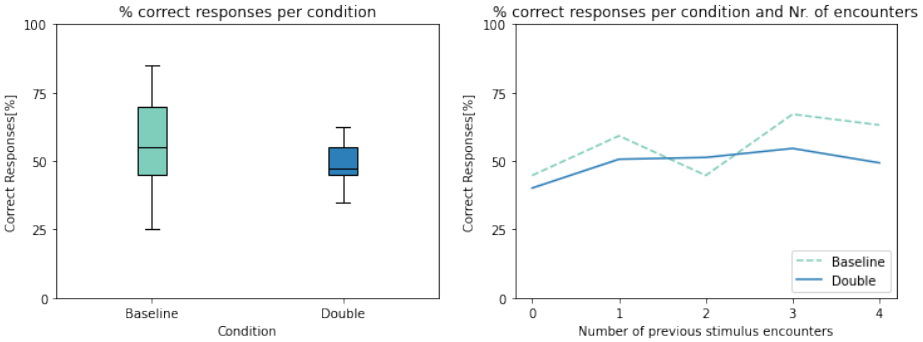
Contrast	Z-value	p	Cohen's d
Baseline - Set Size	48.5	0.171	0.07
Baseline - Doubles	61.0	0.107	0.13
Baseline - Similarity	**17.0**	**0.003**	**0.31**

Table 4.4: Results Second Pretest

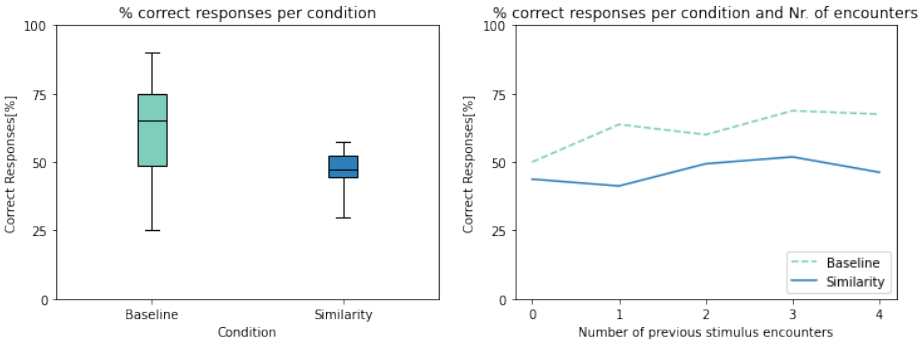
was shown to be significantly more difficult to learn than the baseline list, with a moderate effect size ($Z = 17, p = 0.003, d = 0.31$).



(a) Set Size Condition



(b) Doubles Condition



(c) Similarity Condition

Figure 4.2: Results of second Stimulus Pretest: The left columns shows the percentages of correct classification across all trials. The right columns shows the percentages depending on how often participants had already seen the stimuli pair.

4.1.3 The Finalized Learning Environment

Based on the findings derived from the pre-tests, the learning environment's design was finalized in the following way. Three skills are to be trained, $N_s = 3$. Those are the "Base" skill, the "Medium Similarity" skill and the "Similarity" skill. It was deemed appropriate to include an intermediate set of stimuli due to the medium effect size of the difference between the baseline and the similarity stimuli. That set was created by randomly sampling non-words from the database [52] and changing two letters (instead of one as for the similarity stimuli) in order to obtain pairings containing stimuli with medium similarity to one another. Each skill translates to the ability to memorize associations between stimuli from the corresponding lists. The baseline and similarity stimuli lists are equal to those detailed in Table 4.3. The newly created medium similarity stimuli are shown in Table 4.5.

A learner's competence in each of these skills is measured in 2 levels, $N_{C_i} = 2$ $\forall\, i$. The QR-system employed in Dawn of Civilisations was slightly adapted to better fit the framework of the experiment. Most notably, the spaced repetition was completely omitted due to the fact that the experiment only spans a very limited amount of time. Consequently, the state s consists only of the number of questions at each QR-level for each skill. The maximal QR-level was reduced to 4. After the first encounter of a question - or stimulus pair - the QR-level only increased to 1 regardless whether a correct answer was given because that answer was considered to be random. The rule that a skill level is considered as completed if at least 80% of the pertaining questions reached their maximal QR-level was kept as is. That translates to all 4 questions for the baseline skill and 7 question for both the medium similarity and the similarity skill. The learning goal was set to the completion of the first and

	Medium Similarity
Valid	zonz - nylk
	zamz - nirk
	murt - cwob
	mokt - cvab
Invalid	zonz - nirk
	zamz - nylk
	murt - cvab
	mokt - cwob

Table 4.5: Medium Similarity Stimuli

only level for all three skills, $\boldsymbol{g} = (1, 1, 1)$. The environment comprises three actions, $|\mathcal{A}| = 3$. Each action trains exactly one skill. $P(\text{correct}|a)$ was taken directly from the results of the pre-test for the baseline and the similarity stimuli. For the medium similarity stimuli, the average of the former two was taken as an estimate. Consequently, $P(\text{correct}|a = \text{baseline}) = 0.585$, $P(\text{correct}|a = \text{similarity}) = 0.465$ and $P(\text{correct}|a = \text{medium similarity}) = 0.525$. Each action always presents exactly 4 questions, therefore the action costs were uniformly set to -1 while reaching the learning goal elicited a positive reward. The optimal state-value function $V^*(s)$ for the sketched-out experimental learning environment MDP was calculated by use of the value iteration algorithm (see Section 2.2.2 and Algorithm 1). From there, optimal brain points were calculated according to Eq. 2.16. Approximated brain points $ABP(s, a)$ were derived according to Eq. 3.20. Subsequently, both kinds of brain points were scaled to integers between 0 and 5. This range came to be out of considerations made for Dawn of Civilisations, where brain points would be awarded in form of 0 - 5 additional reward cards.

As a first evaluation step, the performance of simulated myopic agents receiving either optimal, approximated or rounded approximated pseudo-rewards were contrasted between one another and with those of a random and myopic agents The simulation results are depicted in Figure 4.3. They allow us to conclude that for the simulated myopic agents, the approximated brain points are as beneficial as the optimal brain points and that scaling and rounding them to fit the constrains of displaying them to human learners does not elicit negative effects. Therefore, the second evaluation step commenced and the approach tested with human learners, as described in the following section.

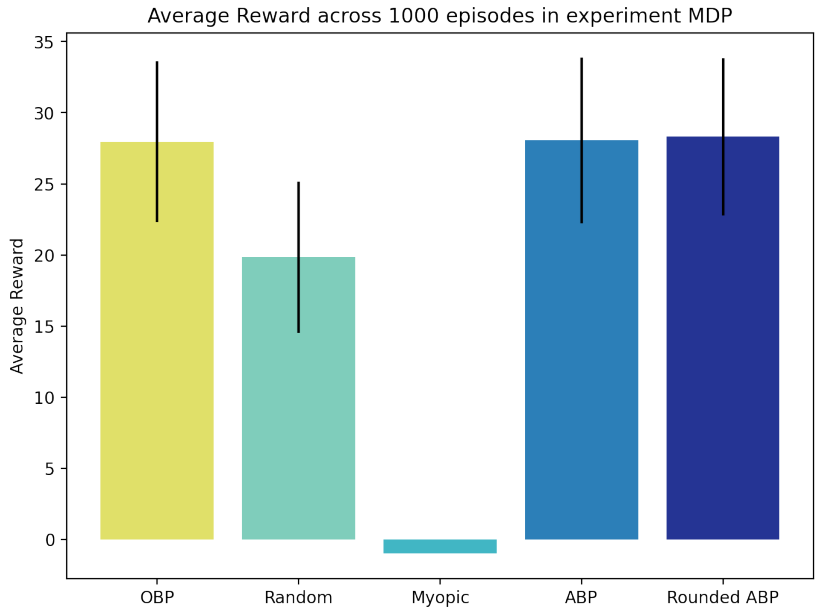


Figure 4.3: Comparison of Simulated Agents in Experimental MDP: Reward obtained by different simulated agents averaged over 1000 episodes. The error bars represent standard deviation. OBP means a myopic agent receiving optimal brain points or pseudo-rewards. Random means an agent that chooses its actions randomly, while the myopic agent selects actions greedily while looking one step ahead without any pseudo-rewards. ABP labels the approximated pseudo-rewards (Eq. 3.20) and Rounded ABP are those same pseudo-rewards scaled and rounded to integers between 0 and 5.

4.2 The Evaluation Experiment

4.2.1 Hypotheses

Based on the findings in the literature summarized in Chapter 2 and the results obtained with simulated agents reported in Sections 3.4 and 4.1.3, the following hypotheses are postulated: First, it is hypothesized that brain points can improve user's learning choice behaviour during self-directed interaction with a digital learning environment by reducing the tendency to exploit strong skills for gathering more rewards. That means that we hypothesize that participants viewing brain points choose more challenging games more often and games they can no longer progress in less often. Second, it is hypothesized that brain points can improve user's learning outcome during self-directed interaction with a digital learning environment. That means that we hypothesize that participants viewing brain points complete more word pairs and overall achieve higher QR-levels throughout the experiment. Third, it is hypothesized that the effects of approximated brain points do not differ from the effects of optimal brain points.

The following sections detail the experiment conducted in order to test this hypotheses.

4.2.2 Methods

Design The experiment was a between-subjects design with five conditions. In all conditions, participants saw a score based on the number of questions they had answered correctly and knew that their bonus payment depended on that score. In the control condition, participants chose between learning activities without any further incentives shown to them. In the second condition (OBP), participants were shown rounded optimal brain points when choosing between learning activities. In the third condition (ABP), participants were shown rounded approximate brain points when choosing between learning activities. In order to be able to detangle possible effects of the different incentive schemes on learners' choice behaviour and their learning outcomes, two forced-choice conditions were included. In these, participants did not get to choose which learning task they wanted to engage in. Instead, they were forced to play the game assigned the highest number of optimal brain points (OBP-FC) or approximate brain points (ABP-FC). This design allowed to find effects on learning

outcomes even if there were no effects on choice behaviour. That is especially important as the mechanisms of gamification are difficult to incorporate in the minimalist experiment and the shown brain points were arguably less incentivizing than sophisticated gamification elements displayed in an overall more engaging context. Including the brain points in the bonus payments, on the other hand, would have been too strong of an incentive to meaningfully interpret the results.

In order to test the hypotheses, learning choice behaviour and learning outcome had to be measured. Three dependent variables were chosen to measure aspects of choice behaviour. First, the percentage of choices made in favour of the game that yielded the highest score on average up to the time the choice is made was measured. This variable captured the concept of exploiting strong skills in order to increase rewards. Second, the percentage of choices made in favour of a game with the highest number of optimal brain points was measured to quantify the quality of participants' choices. Third, the percentage of choices made in favour of the easiest game was used to assess how much participants preferred the easy task over the harder tasks.

Learning outcome was operationalized in two different ways. The first was the number of completed word pairs at the end of the experiment as per the rules of Dawn of Civilisation, meaning having reached the maximal QR-level of 4. Because the measure dichotomizes learning progress to some extend, additionally the sum of all the pairs' QR-levels at the end of the experiment was measured as a more continuous indicator of made progress.

Having specified the dependent variables allows to also concretize the hypotheses formulated in Section 4.2.1: Participants receiving either optimal or approximate brain points were expected to choose the baseline game and their highest-scoring game (which can, but do not have to be the same) with a lower percentage and an optimal game with a higher percentage than participants in the control condition. Further, participants receiving either optimal or approximate brain points or being forced to choose according to the optimal or approximate policy were expected to complete a higher number of word pairs and achieve a higher sum of QR-levels by the end of the experiment than participants in the control condition. Lastly, the type of points was not expected to elicit differences in any of the dependent variables.

Participants The required sample size was determined based on the effect sizes reported by Xu, Wirzberger & Lieder (2019) [1] and with the help of G*Power [58]. Assuming a effect size of $\eta^2 = 0.218$, a Type-I error probability of 0.05 and striving for

a statistical power of 95% for 5 groups, a required sample size of 272 was calculated. Allowing for buffer, 60 participants were recruited per condition and with that 300 participants overall.

Consequently, 300 participants (57.5% female; *mean age* = 24.9 years, *SD* = 6.16 years) were recruited via the online recruitment platform Prolific [53]. All participants provided written informed consent and received a base compensation of 1.70 £ for the 16 minute experiment. Additionally, they were awarded a bonus payment based on their performance, with the average bonus set to be 0.40 £. All participants were older than 18 years and fluent in English.

36 participants were excluded according to the predefined exclusion criteria. This means that they failed two or more of the included attention checks, which were identical to those used in the previously reported stimuli tests (see Section 4.1.1). New participants were recruited to fill their places.

Participants were randomly assigned to conditions. In the end, 67 participants completed the control condition, 57 participants each completed the optimal brain points, approximated brain points and optimal brain points with forced-choice conditions and 62 participants completed the approximate brain points with forced-choice condition.

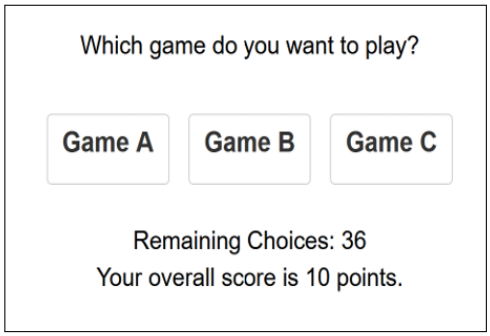
This experiment was covered by the ethics approval from the IEC of the University of Tübingen under IRB protocol number 667/2018BO2. Funding for participants' compensations was granted by the Rationality Enhancement Group at the Max-Planck-Institute for Intelligent Systems in Tübingen.

Materials and Procedure The experiment was programmed with jsPsych [54, 55].

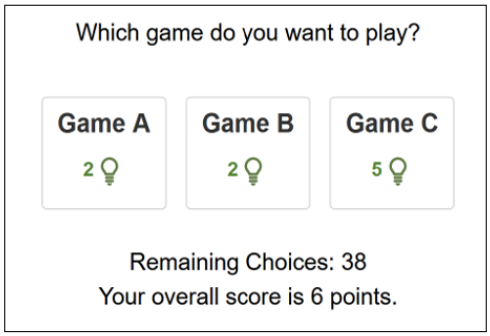
After giving their consent to participate in the experiment, participants were provided with instructions.

The main part of the experiment consisted of 40 mini-blocks. Each mini-block in turn consisted of one choice trial, four learning trials and one screen message informing participants of their score. If a participant reached the learning goal, the experiment ended early and they did not have to complete the full 40 mini-blocks.

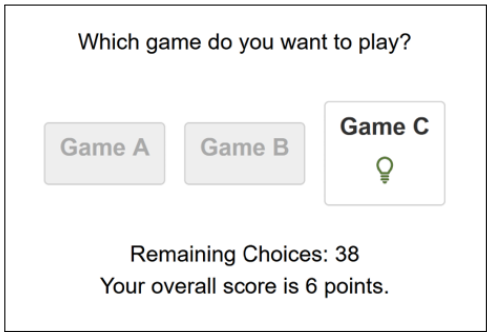
In each choice trial (see Figures 4.4a, 4.4b and 4.4c), participants clicked on the button corresponding to the game they wanted to play. In the forced-choice conditions, only one button was enabled. The assignment of the base, medium similarity and similarity games to Game A, B and C was randomized to rule out the possibility of an effect of the order. The learning trials consisted of two parts. First participants are



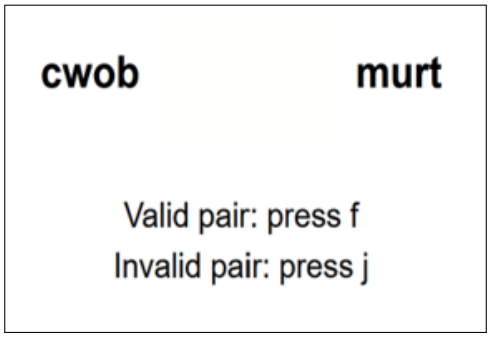
(a) Choice trial in control condition



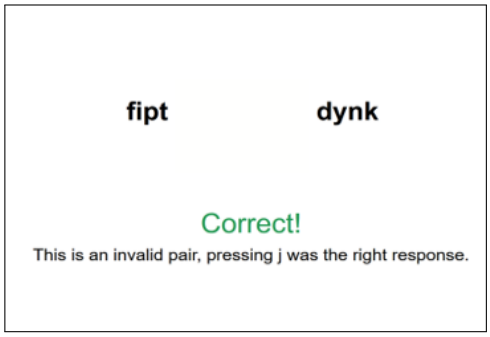
(b) Choice trial in free choice conditions



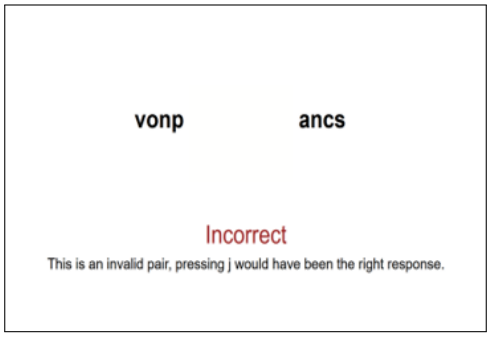
(c) Choice trial in forced-choice conditions



(d) Example learning trial



(e) Feedback after correct response



(f) Feedback after incorrect response

Figure 4.4: Materials of the evaluation experiment:
Figures 4.4a, 4.4b and 4.4c show exemplary choice trials illustrating the difference between the conditions. Figure 4.4d shows how a learning trials looked like. Figures 4.4e and 4.4f show the presentation of feedback following a correct or incorrect response, respectively.

presented with the word pair and a reaction prompt (Figure 4.4d). There was a time limit of 3.5 seconds for responses. Then, the appropriate feedback to the response was shown for 2 seconds (Figures 4.4e and 4.4f). The stimuli to be presented were selected according to the QR-level based priorities described in Section 3.3.5, which translates into showing the stimuli with the lowest QR-levels for the simplified environment. The order in which the selected stimuli were shown within one game was randomized. After the 4 learning trials, participants were informed of the score they achieved, which equates to the number of correct answers given.

This concludes the description of the experimental methods. In the next section, the process of data analysis and its reasoning and the results are presented.

4.2.3 Results

The analyses reported in this section have been pre-registered with aspredicted.org (see Appendix D). All presented figures were created with Matplotlib [59]. The analyses were performed with the help of pandas [60], scipy [61], statsmodels [62] and pingouin [63].

Descriptive Analyses

First of all, a look was taken on how well participants generally coped with the learning task. As shown in Figure 4.5, over half (55%) of the participants managed to master at least one skill. Specifically, 39.2% of participants completed one skill. 75.4% of that group mastered the baseline skill, 8.5% completed the medium similarity skill and 16.1% completed the similarity skill. 11.2% of participants completed two skills. Of those, 23.5% mastered the baseline and the medium similarity skill, 64.7% the baseline and similarity skill and 11.8% the medium similarity and similarity skill. Overall, 4.4% of participants managed to complete all three skills.

These findings indicate that the learning task was neither too easy nor too hard, which forms a good basis for further evaluation.

Secondly, a manipulation check was performed by testing for differences in the obtained scores per games with a Kruskal-Wallis H-test [64]. Based on the results $(H = 181.1, p < 0.001)$ indicating a significant difference between the three games, pairwise comparisons were made using the Mann-Whitney U-test [56], for which the false discovery rate was controlled with the Bejamini-Hochberg procedure [65]. The

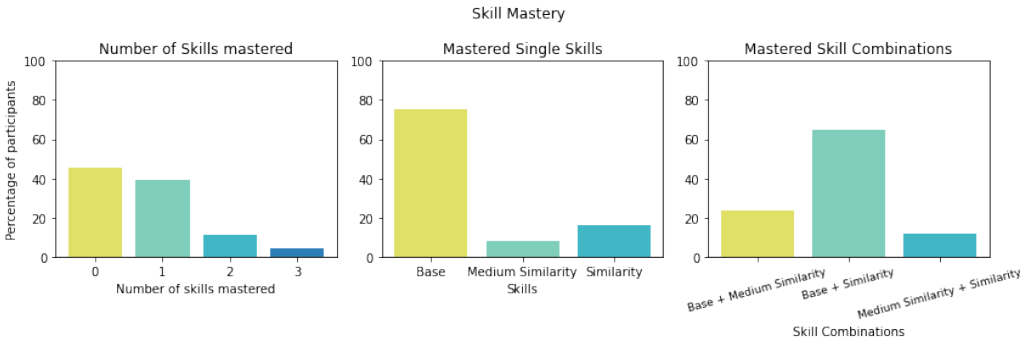
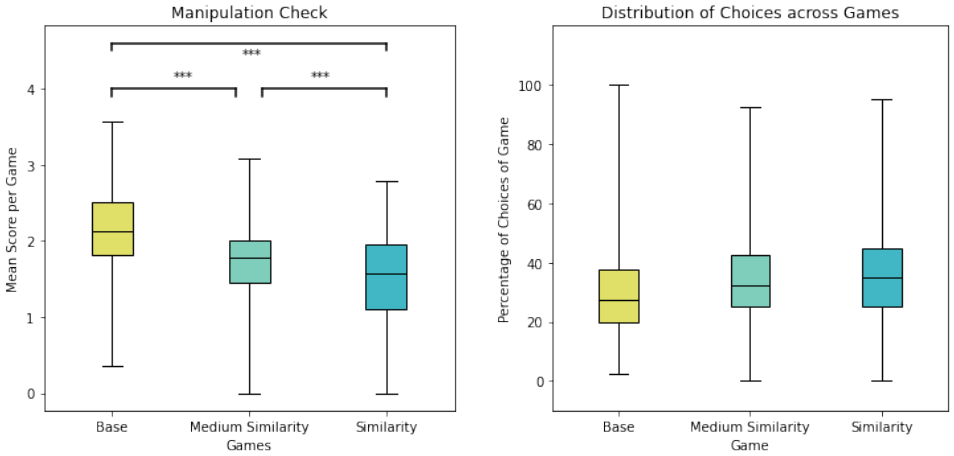


Figure 4.5: Overview of Skill Mastery: The left tile shows the percentage of participants who completed none, one, two or all three skills. The middle tile shows which percentage of the 39.2% participants who mastered one skill completed each of the skills. The rightmost tile shows which percentage of the 11.2% participants who mastered two skills completed each of the possible two-skills combinations.

Contrast	U	p	η^2
Baseline - Medium Similarity	65714.0	<0.001	0.15
Baseline - Similarity	72159.5	0.001	0.26
Medium Similarity - Similarity	55441.5	<0.001	0.04

Table 4.6: Results of Score Comparisons

results are summarized in Table 4.6 and Figure 4.6a. It can be seen that the difficulty of the three learning games differed as intended by the experimental design as participants gave more correct answers to the baseline stimuli than to those meant to be more difficult to learn. The medium similarity stimuli, which had not been tested before, elicited significantly more correct answers than the similarity stimuli while eliciting significantly less correct answers the baseline stimuli. Therefore, the introduction of a medium difficulty stimuli set can be viewed to have been successful, even though the effect sizes reported in Table 4.6 lead to the conclusion that the medium similarity stimuli are much closer in difficulty to the similarity stimuli than they are to the baseline stimuli. The data depicted in Figure 4.6b shows that overall the participants chose to interact with the three different games at comparable rates. Thereby, the evidence of a successful difficulty manipulation is supported because it allows to conclude that the observed difference in difficulty is not just an effect of increased practice of the baseline game.



(a) Manipulation Check

(b) Distribution of Choices across Games

Figure 4.6: Manipulation Check - Scores and Choices: Figure 4.6a shows the scores obtained by participants by games. The score can range from 0 for no correct answer to 4 for 4 correct answers, as each game presents 4 stimuli to react to. On the right hand side, in Figure 4.6b, the percentages of choices made by participants for each of the three games is shown. The percentage is in reference to the 40 choices made by each participant throughout the experiment. In both subplots, the boxes represent the interquartile range. The black lines inside the boxes indicate the median. The whiskers extend to the minimal and maximal observed values. *** indicates a significant difference with $p < .001$.

Variable	W	p
Choice by Highest Score	0.93	0.001
Choice for Optimal Action	0.46	0.001
Choice for Base Game	0.87	0.001

Table 4.7: Test for normality of choice variables

After this familiarization with the experimental data, the hypotheses tests were performed as described in the following sections.

Choice Behaviour

As described in Section 4.2.2, three aspects of the learner's choice behaviour have been measured. The first is the percentage of choices made in favour of the game that up to that point led to the highest average score. The second is the percentage of choices made in favour of any game assigned the maximal number of optimal brain points. And the third is the percentage of choices made in favour of the easiest game, presenting the baseline stimuli. Due to the fact that choosing any game training a skill which has not yet been completed is an optimal action, choices made after the first skill acquisition were analysed in addition to testing for effects across all choices made throughout the experiment. Because the choices made are the variable of interest, data from participants in the two forced-choice conditions are not included in these analyses. The Shapiro-Wilk test was used to test for normality on account of its statistical power [66]. Table 4.7 summarizes the results for the three choice variables. It can be concluded that the data is not well-modelled by normal distributions. Therefore, non-parametric tests were deemed the appropriate tool to test the hypothesis formulated in Section 4.2.1.

Choice by Highest Score Participants in the control condition showed the highest percentage of choices in favour of their highest-scoring game (median = 40%, IQR = 20%), followed by participants presented with optimal brain points (median = 37.5%, IQR = 12.5%) and participants presented with approximate brain points (median = 32.5%, IQR = 15%), see Figure 4.7a.

The Kruskal-Wallis H-test [64] showed that at least one of the samples stochastically dominates another (H = 8.76, p = 0.012, $\eta^2 = 0.038$). Hence, pairwise comparisons

Contrast	U	p (corrected)	η^2
Control - OBP	2216.5	0.167	0.015
Control - ABP	**2527.5**	**0.010**	**0.068**
OBP - ABP	1900.0	0.167	0.021

Table 4.8: Pairwise Comparison of Choice by Highest Score

Contrast	U	p (corrected)	η^2
Control - OBP	**39**	**0.030**	**0.709**
Control - ABP	**18**	**0.035**	**0.725**
OBP - ABP	44	0.691	0.704

Table 4.9: Pairwise Comparison of Choice by Highest Score after first Skill Acquisition

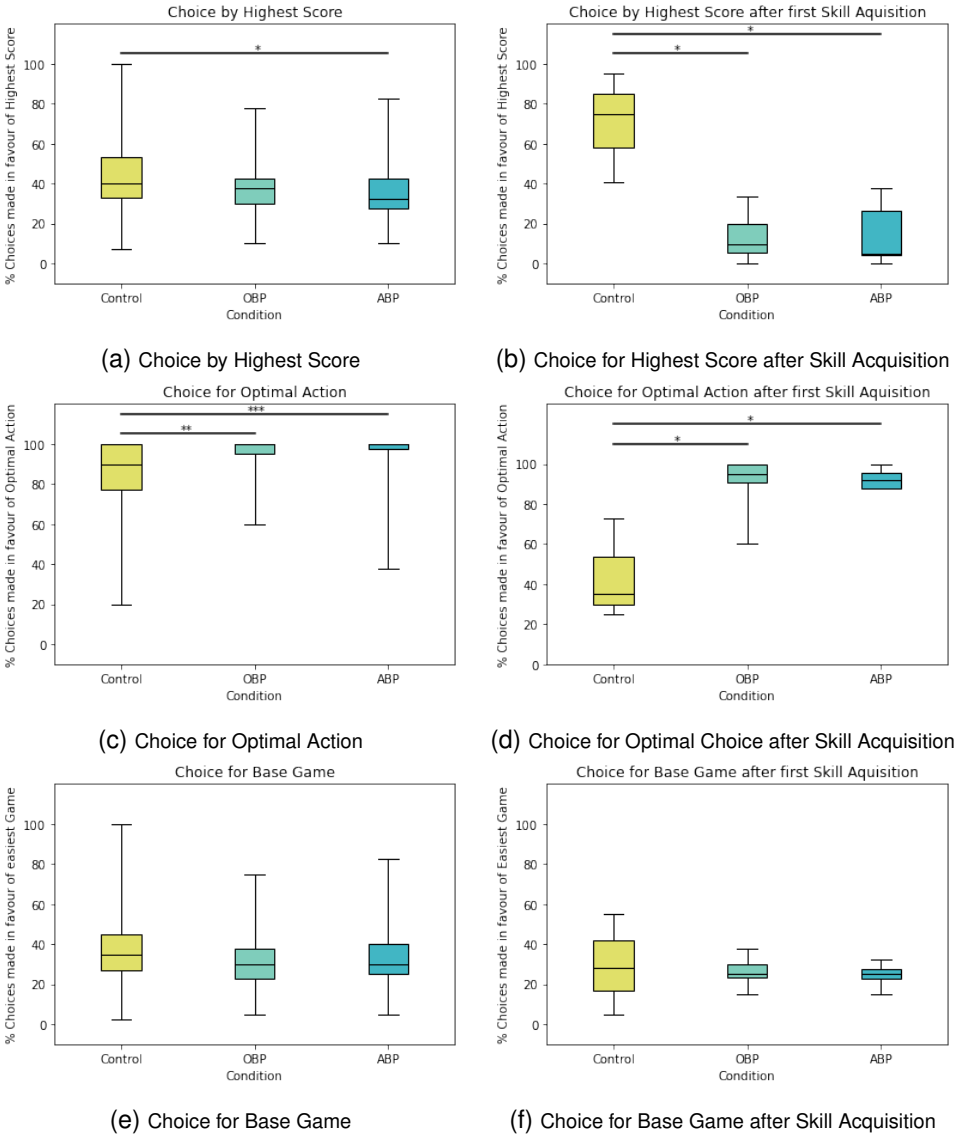
were performed with the Mann-Whitney U-test [56], the results of which are summarized in Table 4.8. The false discovery rate was controlled with the Benjamini-Hochberg procedure [65].

It was shown that participants presented with approximated brain points chose the game they performed best in so far at significantly lower rates than participants in the control condition, with a medium effect size.

When considering only the choices made after the first skill acquisition (Figure 4.7b), the following pattern presents itself: Participants in the control condition chose the game they performed best in so far at the highest rate (median = 75%, IQR = 27%). Participants in the OBP-condition (median = 9.5%, IQR = 14.4%) and in the ABP-condition (median = 4.7%, IQR = 22.2%) chose their highest-scoring game at lower rates.

A significant difference between at least two groups was found with the Kruskal-Wallis H-test (H = 7.62, p = 0.022, η^2 = 0.031). The subsequently performed pairwise comparisons reported in Table 4.9 showed that the choice rates in the control condition significantly differ from those in both the optimal brain points and the approximate brain points condition, with large effect sizes.

Choice for Optimal Action The percentage with which an optimal action was chosen (see Figure 4.7c) was generally high with varying spread: The participants in the control condition (median = 90%, IQR = 23.1%) showed a higher variance in their choices than those presented with either optimal (median = 100%, IQR = 5%) or approximate (median = 100%, IQR = 2.5%) brain points. The Kruskal-Wallis H-test [64]



(a) Choice by Highest Score

(b) Choice for Highest Score after Skill Acquisition

(c) Choice for Optimal Action

(d) Choice for Optimal Choice after Skill Acquisition

(e) Choice for Base Game

(f) Choice for Base Game after Skill Acquisition

Figure 4.7: Effect of Type of Points on Choice Behaviour:
Depicted are different aspects of choice behaviour by condition: Control, Optimal Brain Points (OBP) and Approximate Brain Points (ABP). The left column shows the effect of the type of points on all choices made by participants. The right columns shows the effect on choices made after participants had completed at least one skill. The uppermost row shows the percentage of choices in favour of the game that up to that choice led to the highest average score. The middle row shows the percentage of choices made in favour of a game assigned the highest number of optimal brain points. The lower row shows the percentage of choices made in favour of the base game and with that the easiest game.
In all subplots, the boxes represent the interquartile range. The black lines inside the boxes indicate the median. The whiskers extend to the minimal and maximal observed values. *** indicates a significant difference with $p < .001$, ** indicates a significant difference with $p < .01$ and * indicates a significant difference with $p < .05$.

Contrast	U	p (corrected)	eta
Control - OBP	**1344.0**	**0.003**	**0.069**
Control - ABP	**1167.0**	**<0.001**	**0.117**
OBP - ABP	1409.5	0.017	0.013

Table 4.10: Pairwise Comparison of Choice for Optimal Action

Contrast	U	p (corrected)	η^2
Control - OBP	**2.0**	**0.041**	**0.737**
Control - ABP	**0.0**	**0.041**	**0.738**
OBP - ABP	44.5	0.655	0.705

Table 4.11: Pairwise Comparison of Choice for Optimal Action after first Skill Acquisition

showed that at least one difference between the groups (H = 19.67, p = < 0.001, η^2 = 0.099). Pairwise Mann-Whitney U-test with Benjamini-Hochberg correction (results reported in Table 4.10) showed that the medium differences between the control and both brain points conditions are statistically significant.

As completing one skill removes the corresponding game for the set of optimal actions, it is especially interesting to consider the choices made after the first skill acquisition for this variable, which are depicted in Figure 4.7d. After completing their first skill, participants who were shown optimal brain points choose an optimal action with the highest rate (median = 95%, IQR = 9.5%), followed by participants who saw approximate brain points (median = 91.8%, IQR = 8.7%). Participants in the control condition exhibit the largest decrease in choices in favour of an optimal action after acquiring a skill (median = 35%, IQR = 23.8%).

An effect of the condition was found using the Kruskal-Wallis H-test (H = 6.70, p = 0.035, η^2 = 0.027). The subsequent pairwise comparisons reported in Table 4.11 show that the participants in the control condition chose an optimal action at significantly lower rates than participants presented with either kind of brain points.

Choice for the easiest Game Lastly, the percentage of choices in favour of the easiest game was examined (see Figure 4.7e). Participants in the control condition (median = 35%, IQR = 18.1%), the OBP-condition (median = 30%, IQR = 15%) and the ABP-condition (median = 30%, IQR = 15%) chose the easiest game at comparable rates. No difference between conditions was found (H = 4.904, p = 0.086. $\eta^2 = 0.016$).

Variable	W	p
Learned Word Pairs	0.938	<0.001
Sum of QR-Levels	0.966	<0.001

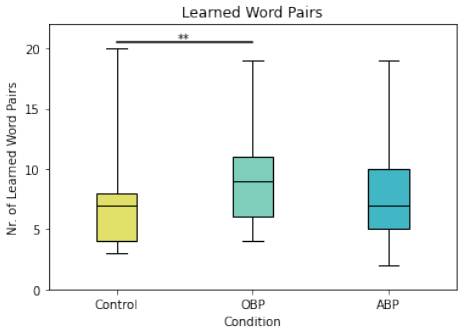
Table 4.12: Test for normality of learning outcome variables

A similar pattern can be observed for choices made after the completion of one skill (Figure 4.7f), with the percentages of choices in favour of the easiest game being in a similar range for participants in the control condition (median = 28.2%, IQR = 25%), and both the optimal (median = 25%, IQR = 6.4%) and the approximate (median = 25.1%, IQR = 4.9%) brain points condition. Again, no difference between conditions was found (H = 0.431, p = 0.009).

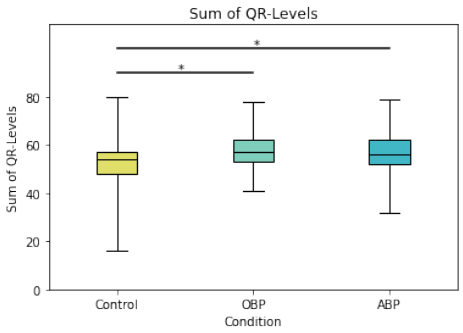
Learning Outcomes

As already explained in Section 4.2.2, the learning outcome has been operationalized by two different measures. Firstly the number of word pairs classified as learned by the QR-System mimicking Dawn of Civilisations'. Secondly, the sum of the QR-Levels, serving as a more continuous measure of learning progress. The Shapiro-Wilk test was employed to test whether these variables are well-modelled by a normal distribution [66]. Seeing the results in Table 4.12, it was apparent that this was not the case. Hence, non-parametric test were used in subsequence.

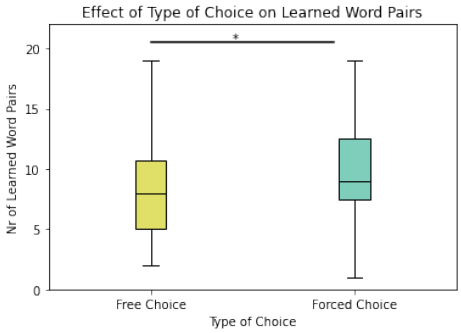
In order to adequately deal with the not fully factorial experimental design, Kruskal-Wallis H-tests were used to test for differences between the three free choice conditions. Two-way ANOVAs were planned to be used to evaluate the main effects and possible interactions of the factors type of points (optimal vs approximate) and type of choice (free vs forced) for both dependent variables. Based on a literature review [67, 68], the decision was made to not use MANOVAs because in spite of dealing with two dependent variables, the research questions asked are univariate in nature. However, the data violate the necessary assumption of normality. For this case, the preregistered analysis intended to rely on the Friedman test [56] as a non-parametric alternative. That was an unfortunate mistake as the Friedman test procedure is only suitable for the analysis of matched samples. As a first alternative, performing the two-way ANOVA on rank-transformed data was considered, but it has been argued that this procedure suffers from inflated Type-1 error rates [69]. Instead, the analyses of interactions between the factors is forgone for the sake of statistical robustness and the factors will be looked at separately with Mann- Whitney U-tests.



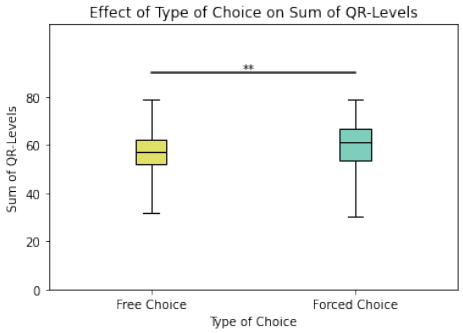
(a) Learned Word Pairs in the free choice conditions



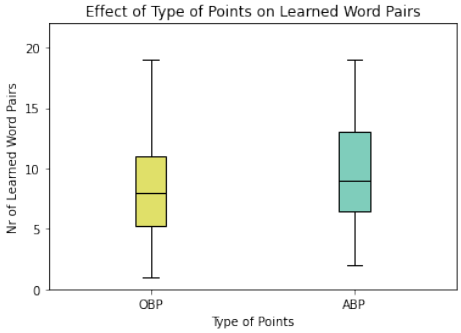
(b) Sum of QR-Levels in the free choice conditions



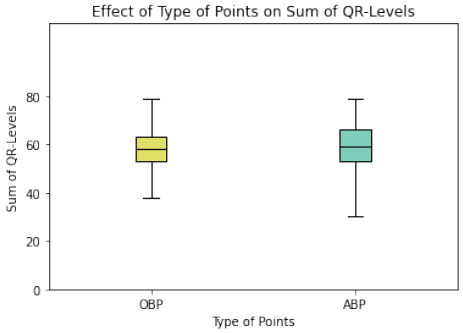
(c) Learned Word Pairs by type of choice



(d) Sum of QR-Levels by type of choice



(e) Learned Word Pairs by type of points



(f) Sum of QR-Levels by type of points

Figure 4.8: Learning Outcomes: Depicted are the two operationalisations of learning outcomes, learned word pairs in the left column and sum of QR-levels in the right columns. The maximal possible value for learned word pairs is 20. The maximal possible sum of QR-levels is 80. The uppermost row shows the outcomes by the three types of points in the conditions allowing participants to choose freely which learning activity to engage in. The lower two rows show the effects of type of choice and type of points on the outcome measures, respectively. These plots do not contain data from the control condition, as an aggregation of data by the factor if type of choice is not possible.
In all subplots, the boxes represent the interquartile range. The black lines inside the boxes indicate the median. The whiskers extend to the minimal and maximal observed values. ** indicates a significant difference with $p < .01$ and * indicates a significant difference with $p < .05$.

Contrast	U	p (corrected)	eta
Control - OBP	**1321.0**	**0.006**	**0.075**
Control - ABP	1687.5	0.211	0.012
OBP - ABP	1883.0	0.211	0.019

Table 4.13: Pairwise Comparison of the Number of Learned Word Pairs

Number of Learned Word Pairs Participants receiving optimal brain points completed a higher number of questions (median = 9, IQR = 5) than those receiving either approximate (median = 7, IQR = 4) or no (median = 7, IQR = 5) brain points, as can be observed in Figure4.8a. An effect of the type of points was found using the Kruskal-Wallis H-test (H = 8.96, p = 0.011, $\eta^2 = 0.039$). Pairwise Mann-Whitney U-tests (see results in Table 4.13) showed that effect to be driven by a significant difference between the number of word pairs completed by participants in optimal brain points condition and by those in the control condition.

Participants choosing freely between games completed less word pairs (median = 8, IQR = 5.75) than those forced to interact with the game with the highest incentive assigned to it (median = 9, IQR = 5). This is a small but significant of the type of choice on the number of learned word pairs (U = 5340.5, p = 0.005, $\eta^2 = 0.034$).

The type of incentive scheme (approximate: median = 9, IQR = 6.5, optimal: median = 8, IQR = 5.75) did not have an effect of the number of learned word pairs (U = 5954.0, p = 0.106, $\eta^2 = 0.011$).

Sum of QR-Levels As can be seen in Figure 4.8b the sum of QR-levels were highest in the OPB-condition (median = 57, IQR = 9), followed by the ABP-condition (median = 56, IQR = 10) and then the control condition (median = 54, IQR = 9.25). A significant difference between at least two groups was found with the Kruskal-Wallis H-test (H = 8.456, p = 0.015, $\eta^2 = 0.036$). The subsequently performed pairwise comparisons reported in Table 4.14 showed that participants in both brain points conditions achieved significantly higher QR-sums than those in the control condition with medium effect sizes.

Participants which were forced to play either the game with the highest number of optimal or approximate brain points achieved higher sums of overall QR-levels at the end of the experiment (median = 61, IQR = 13) than participants who saw those points but were free to choose which game to interact with (median = 57, IQR = 10). This is

Contrast	U	p (corrected)	eta
Control - OBP	**1393.5**	**0.020**	**0.058**
Control - ABP	**1503.0**	**0.046**	**0.037**
OBP - ABP	1729.5	0.553	0.003

Table 4.14: Pairwise Comparison of the Sum of QR-Levels

depicted in Figure 4.8d. This small effect of type of choice on sum of QR-levels was found to be significant (U = 5403.5, p = 0.007, η^2 = 0.031).

Figure 4.8f shows that there was no effect of whether approximate (median = 58, IQR = 10) or optimal (median = 59, IQR = 13) brain points were used as incentives or as choice directive on the sum of QR-levels found (U = 6359.0, p = 0.409, η^2 = 0.003).

4.2.4 Discussion

The purpose of this experiment was to evaluate the effects of presenting brain points to learners choosing between different educational activities. Specifically, the effects on learners' choice behaviour and their learning outcomes were examined. It was hypothesized that brain points could reduce the rate at which learners choose the game they performed best in so far. The results show that to be the case. Especially, learners were shown to switch from choosing their highest-scoring game to another after having completed the corresponding skill. That means that brain points encouraged the learners to challenge themselves rather than to exploit a mastered skill either for increasing their score or for obtaining a feeling of achievement, a tendency often found in previous studies on gamifying learning [1]. Furthermore, even though the experiment did not include specific narratives promoting a growth mindset, this pattern of seeking out challenges can be interpreted as a behaviour reflecting a growth rather than a fixed mindset [7].

In line with these findings, brain points were further shown to impact the quality of learners' action selection positively. Contrary to the predicted effects, no difference in the rate at which learners chose the easiest game was found. Nevertheless, as the effect of brain points on the choice of the highest-scoring game can be viewed as a more personalized estimate of the learners' tendency to exploit the task easiest to them, the following conclusion can be made: The display of brain points has positive effects on the way learners choose educational activities, independent of the type of brain points. This is very encouraging, since participants could have reasonably

opted to exploit their strongest skill in order to obtain more monetary reward but the simple display of the incentives impacted their behaviour positively. Therefore, brain points are expected to unfold an even stronger behavioural impact when applied in a non-artificial learning environment and tied to a relevant in-game currency.

Regarding the learning outcome measures, it was found that optimal brain points increase the number of word pairs learned by freely choosing participants compared to showing no incentives. For the more continuous outcome measure, the sum of QR-levels, both types of brain points led to a significant improvement compared to the control condition. Considering the effect of the two incentive schemes across the respective free and forced-choice conditions, no difference in effectiveness was found. In hindsight, the experiment was maybe too short to effectively use the number of learned word pairs as an outcome measure. It can be interpreted to be an evaluation of an theoretically ongoing learning progress after 40 choices. A more powerful alternative could have been to phrase the experimental MDP as a finite horizon MDP [70] and calculate the brain points in a way that explicitly includes the knowledge that only a limited amount of time is available to the learner. At the same time, the sum of QR-levels offer a more continuous measure of progress but also a less stable one. What is meant by that is that, according to the learning model behind the QR-system, uncompleted materials are at a higher risk for being forgotten again [41, 42]. In summary, measuring learning outcomes for artificial tasks in a 16 minute experiment is not completely straightforward. Nevertheless, the results allow to conclude brain points - both optimal and approximate - improve rather than hinder learning progress. Therefore, they can be responsibly further evaluated in a real-world context such as Dawn of Civilisations without having to fear participants experiencing harm in the sense of diminished learning outcomes.

A small effect of the type of choice on the learning outcome was found, with participants choosing freely achieving slightly lower outcomes than those not allowed to choose. In Dawn of Civilisations, implementing forced choices is not an option, as that would entail the risk of users getting frustrated and reducing their learning time. Therefore, the small magnitude of the effect is encouraging in the sense that displaying brain points can have almost as beneficial effects as imposing the choice deemed optimal on the learner. On top of that, as has been already argued, brain points are expected to have stronger motivational effects in Dawn of Civilisations since they translate to the in-game currencies. Thereby, the difference could be further mitigated.

All in all, it can be concluded that brain points can improve learners' choice behaviour

and their learning outcomes. Moreover, approximate brain points were shown to have as beneficial effects on learner's choice behaviour and learning outcomes as optimal brain points and can therefore be further evaluated within the context of Dawn of Civilisations.

Further implications and avenues for future research are discussed in the next chapter.

5 Outlook

In line with the findings on the application of optimal gamification to learning contexts [1] discussed in Chapter 2, we have found the approach to incentivizing learning choices to produce promising results on learners' choice behaviour and learning outcomes in an artificial learning task. The more interesting question, however, is which effects can be observed when actual learners engaging with real-world educational material are incentivized with the approximated brain points. To that end, the next step is to evaluate the approach in cooperation with Solve Education. Users in the experimental condition will see the approximated brain points when selecting minigames as illustrated in Figure 5.1. After completing a minigame, learners receive the corresponding number of reward cards in addition to those derived from the current reward system explained in Section 3.2. The cooperation with Solve Education offers the immensely valuable opportunity to evaluate the approach directly with its target audience. Thereby, the planned evaluation study will have considerable ecological validity and allow a more meaningful evaluation of the approach developed in this thesis.

Nevertheless, some important limitations have already become apparent during the first evaluation step presented here. First of all, let us revisit the way the brain points are derived in Section 3.4. Going against the principles of the shaping theorem [32], the basis for approximating brain points is the maximal possible progress instead of the expected progress to be made by selecting an action. That means that it cannot be guaranteed that a learner maximizing the brain points they receive will at the same time maximize their potential learning progress. The empirical evaluation of the approach does not hint at a misalignment of rewards and objective, but deriving brain points in a way that adheres to the shaping theorem has to be considered a superior solution. One possibility to that end would be to refine the learning model. Currently, we estimate the probability of a learner getting a question right purely based on which minigame they are playing. Remaining in the data-driven approach, it would be useful to condition that probability on several and more meaningful factors. A more sophisticated approach could be to treat the probability derived from the data as a prior and update it online with the observed responses of the individual interacting with the

Figure 5.1: Presenting Brain Points in Dawn of Civilisation: Brain points are communicated by the purple annotations by the minigames. Screenshot is a courtesy by Solve Education

minigames according to the Bayes' rule [71]. Further improvements could be made by supporting the data-driven approach with a sophisticated cognitive learning model, such as the multiple trace memory model [72].

Another promising field of research that could inform improvements to the approach is optimal spaced repetition [40, 73]. It has produced and successfully evaluated approaches that treat spaced repetition as an optimal control problem with the goal to maximize recall probability while considering the cost of reviewing [40]. It could replace the simple heuristics used to determine how much time should pass between recalls of the learning materials.

Lastly, the approach as is has been presented here has an important implicit assumption, which is that the learner will carry out one of the offered educational activities. In the strict sense, the brain points can be considered a decision aid but not necessarily a motivational device to start learning in the first place. This could be addressed by adding an action to the model's action space that represents not choosing any minigame and entails no action cost. Adding the action of choosing not to engage with any minigame would entail further changes to the model, specifically how the passage of time can still be reasonably incorporated. One would probably have to define the action along the lines of choosing not to play any minigame on that particular day in order to update delay periods as described in Section 3.3.5.

Generally, the effectiveness of the incentives derived from the proposed approach

will always depend on the quality of the learning model fed into it. Apart from its effectiveness, two additional expectations towards the developed incentive approach were formulated. The first claim to be fulfilled was scalability. As discussed in Section 3.4.2, scalability posed a serious challenge during the development, leading to a failed attempt to approximate the state-action value function for the model of Dawn of Civilisation. In the end, directly approximating brain points proved to be a functional way of dealing with high complexity of real-world learning environments. However, that approach of course depends on a suitable potential function, which leads to the second claim to be addressed, which was generality. In principle, we set out to develop a method to compute incentives for any gamified educational environment allowing learners to learn in a self-directed manner and choose between educational activities. During the formulation of the general model of choosing education activities in Section 3.1, the list of requirements grew. Specifically, the educational environment must provide a way to quantify competence values and progress, a detailed model of skill improvement, a way to estimate the effort each action requires and a way to express learning goals as competence values. As argued above, the quality of the incentives derived by this approach depends heavily on the quality of the learning model it is to be applied to. The improvements proposed above mostly apply to language learning and therefore do not support the generalizability of the approach to other learning tasks. That leaves a lot of variable factors potentially impacting the efficacy of the approach. Therefore, claiming that the approach developed and evaluated throughout this thesis generalizes would be a bit of a stretch. Rather it should be seen as a first step towards that goal. To draw a comparison, the developed approach can be viewed as a recipe which requires a range of spices and cooking appliances not typically found in every kitchen. Evaluating the approach with different learning environments could allow to enhance it by providing principles for formulating the environment's learning model and mechanisms in a suitable way to other potential applicators.

Returning to the research question posed in Section 2.3, whether the combination of core findings in motivational psychology and methods from reinforcement learning allows to develop a scalable method for deriving incentives for real-world educational environments the following conclusion can be made: Yes, but there is still a long way to go.

6 Conclusion

We have learned that gamification is a trending tool in education, but not always brought to use with the necessary carefulness and theoretical foundation [1, 3, 4, 6, 17]. A key issue that has been identified is the misalignment of the gamified incentives and the target behaviour. Building on research demonstrating the successful combination of psychological research on the benefits of fostering a growth mindset and modeling approaches taken from reinforcement learning for incentivizing learners optimally [1], we aimed to develop a principled and scalable approach for designing gamification for digital learning environments.

The developed approach is based on a formal model of choosing between different educational activities. Its application allows to predict which choice will lead to the maximal learning progress. Its effectiveness was first demonstrated with simulated learners.

A first real evaluation of the developed approach within a controlled online experiment has resulted in promising findings regarding the approach's efficacy: Learners incentivized with the points derived from the approach showed a higher tendency to challenge themselves rather than exploiting what they already had learned and thereby learned more during the experiment. It will be very interesting to see the results of the planned evaluation with Dawn of Civilisations.

However, it has also become apparent that both generalizability and scalability pose serious challenges that make further improvements to the approach necessary.

A more general learning from this project is that unlocking the full potential of gamification in educational is an interdisciplinary challenge. The most scalable computational method is useless if it optimizes an ill-chosen objective, while handcrafted incentives build on psychologically sound reasoning can elicit unforeseen effects if they are gameable. Therefore, we can conclude that the answer to the question "How to incentivize efficient learning choices in digital learning environments?" is by drawing on a diverse set of academic resources and by continuing research in a interdisciplinary fashion.

Bibliography

[1] L. Xu, M. Wirzberger, and F. Lieder, “How should we incentivize learning? An optimal feedback mechanism for educational games and online courses.”, in *CogSci*, 2019, pp. 3136–3142.

[2] F. Lieder, O. X. Chen, P. M. Krueger, and T. L. Griffiths, “Cognitive prostheses for goal achievement”, *Nature Human Behaviour*, vol. 3, no. 10, pp. 1096–1106, Oct. 2019. DOI: 10.1038/s41562-019-0672-9.

[3] A. M. Toda, P. H. D. Valle, and S. Isotani, “The dark side of gamification: An overview of negative effects of gamification in education”, in *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions*, A. I. Cristea, I. I. Bittencourt, and F. Lima, Eds., vol. 832, Series Title: Communications in Computer and Information Science, Cham: Springer International Publishing, 2018, pp. 143–156. DOI: 10.1007/978-3-319-97934-2_9.

[4] R. C. Callan, K. N. Bauer, and R. N. Landers, “How to avoid the dark side of gamification: Ten business scenarios and their unintended consequences”, in *Gamification in Education and Business*, T. Reiners and L. C. Wood, Eds., Cham: Springer International Publishing, 2015, pp. 553–568. DOI: 10.1007/978-3-319-10208-5_28.

[5] E. O’Rourke, K. Haimovitz, C. Ballweber, C. Dweck, and Z. Popović, “Brain points: A growth mindset incentive structure boosts persistence in an educational game”, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 3339–3348. DOI: 10.1145/2556288.2557157.

[6] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, “Off-task behavior in the cognitive tutor classroom: When students "game the system"”, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004, pp. 383–390. DOI: 10.1145/985692.985741.

[7] C. S. Dweck and D. S. Yeager, “Mindsets: A view from two eras”, *Perspectives on Psychological Science*, vol. 14, no. 3, pp. 481–496, 2019. DOI: 10.1177/1745691618804166.

[8] S. Claro, D. Paunesku, and C. S. Dweck, "Growth mindset tempers the effects of poverty on academic achievement", *Proceedings of the National Academy of Sciences*, vol. 113, no. 31, pp. 8664–8668, 2016. DOI: 10.1073/pnas.1608207113.

[9] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? – A literature review of empirical studies on gamification", in *2014 47th Hawaii International Conference on System Sciences*, 2014, pp. 3025–3034. DOI: 10.1109/HICSS.2014.377.

[10] R. A. Oxarart and J. D. Houghton, "A spoonful of sugar: Gamification as means for enhancing employee self-leadership and self-concordance at work", *Administrative Sciences*, vol. 11, no. 2, p. 35, 2021. DOI: 10.3390/admsci11020035.

[11] D. N. Karagiorgas and S. Niemann, "Gamification and game-based learning", *Journal of Educational Technology Systems*, vol. 45, no. 4, pp. 499–519, 2017. DOI: 10.1177/0047239516665105.

[12] P. Buckley and E. Doyle, "Gamification and student motivation", *Interactive learning environments*, vol. 24, no. 6, pp. 1162–1175, 2016. DOI: 10.1080/10494820.2014.964263.

[13] M. D. Hanus and J. Fox, "Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance", *Computers & Education*, vol. 80, pp. 152–161, 2015. DOI: 10.1016/j.compedu.2014.08.019.

[14] L. Hakulinen, T. Auvinen, and A. Korhonen, "Empirical study on the effect of achievement badges in TRAKLA2 online learning environment", in *2013 Learning and Teaching in Computing and Engineering*, IEEE, 2013, pp. 47–54. DOI: 10.1109/LaTiCE.2013.34.

[15] R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett, and K. Koedinger, "Why students engage in "gaming the system" behavior in interactive learning environments", *Journal of Interactive Learning Research*, vol. 19, no. 2, pp. 185–224, 2008.

[16] A. Domínguez, J. Saenz-de Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes", *Computers & Education*, vol. 63, pp. 380–392, 2013. DOI: 10.1016/j.compedu.2012.12.020.

[17] A. Palmquist and J. Linderoth, “Gamification does not belong at a university”, in *2020 DiGRA International Conference: Play Everywhere*, Digital Games Research Association (DiGRA), 2020.

[18] C. Romero, A. Master, D. Paunesku, C. S. Dweck, and J. J. Gross, “Academic and emotional functioning in middle school: The role of implicit theories.”, *Emotion*, vol. 14, no. 2, pp. 227–234, 2014. DOI: 10.1037/a0035490.

[19] E. O’Rourke, E. Peach, C. S. Dweck, and Z. Popovic, “Brain points: A deeper look at a growth mindset incentive structure for an educational game”, in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, ACM, 2016, pp. 41–50. DOI: 10.1145/2876034.2876040.

[20] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, 2nd ed. MIT Press, 2018.

[21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, “Human-level control through deep reinforcement learning”, *Nature*, vol. 518, no. 7540, pp. 529–533, 2015. DOI: 10.1038/nature14236.

[22] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of Go with deep neural networks and tree search”, *Nature*, vol. 529, no. 7587, pp. 484–489, 2016. DOI: 10.1038/nature16961.

[23] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey”, *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013. DOI: 10.1177/0278364913495721.

[24] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, “Reinforcement learning-based multi-agent system for network traffic signal control”, *IET Intelligent Transport Systems*, vol. 4, no. 2, pp. 128–135, 2010. DOI: 10.1049/iet-its.2009.0070.

[25] Z. Zhou, X. Li, and R. N. Zare, “Optimizing chemical reactions with deep reinforcement learning”, *ACS central science*, vol. 3, no. 12, pp. 1337–1344, 2017. DOI: 10.1021/acscentsci.7b00492.

[26] S. Russell and P. Norvig, *Artificial intelligence: A modern approach*. Prentice Hall Upper Saddle River, NJ, USA: 2002.

[27] O. Berger-Tal, J. Nathan, E. Meron, and D. Saltz, "The exploration-exploitation dilemma: A multidisciplinary framework", *PLoS One*, vol. 9, no. 4, e95693, 2014. DOI: 10.1371/journal.pone.0095693.

[28] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10295.

[29] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching", *Machine Learning*, vol. 8, no. 3-4, pp. 293–321, 1992.

[30] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay", in *International Conference on Learning Representations*, Puerto Rico, 2016.

[31] M. Grzes and D. Kudenko, "Theoretical and empirical analysis of reward shaping in reinforcement learning", in *2009 International Conference on Machine Learning and Applications*, IEEE, 2009, pp. 337–344. DOI: 10.1109/ICMLA.2009.33.

[32] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping", in *ICML*, 1999, pp. 278–287.

[33] P. Krueger, T. Griffiths, and S. J. Russell, "Shaping model-free reinforcement learning with model-based pseudorewards", in *Conference on Cognitive Computational Neuroscience.*, 2018, pp. 1–5. DOI: 10.32470/CCN.2018.1191-0.

[34] T. Amalia, *Core Curriculum*. Solve Education! Foundation, 2018. [Online]. Available: https://solveeducation.org/research/se-core-curriculum/ (visited on 05/05/2021).

[35] Solve Education!, *Solve Education Homepage*, 2021. [Online]. Available: https://solveeducation.org/ (visited on 05/03/2021).

[36] ——, *Dawn of Civilisation Homepage*, 2021. [Online]. Available: https://dawnofcivilization.net/ (visited on 05/03/2021).

[37] R. Pinon and J. Haydon, "English language quantitative indicators: Cameroon, Nigeria, Rwanda, Bangladesh and Pakistan", *A custom report compiled by Euromonitor International for the British Council*, 2010.

[38] C. M. Bishop, "Pattern recognition", *Machine Learning*, vol. 128, no. 9, 2006.

[39] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001.

[40] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf, and M. Gomez-Rodriguez, "Enhancing human learning via spaced repetition optimization", *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 3988–3993, 2019. DOI: 10.1073/pnas.1815156116.

[41] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis.", *Psychological Bulletin*, vol. 132, no. 3, pp. 354–380, 2006. [Online]. Available: https://escholarship.org/uc/item/3rr6q10c.

[42] J. M. Murre and J. Dros, "Replication and analysis of Ebbinghaus' forgetting curve", *PLoS One*, vol. 10, no. 7, pp. 1–23, 2015. DOI: 10.1371/journal.pone.0120644.

[43] S. Leitner, *So lernt man lernen: angewandte Lernpsychologie - ein Weg zum Erfolg*. Herder, 1991.

[44] N. Gopalan, M. L. Littman, J. MacGlashan, S. Squire, S. Tellex, J. Winder, L. L. Wong, *et al.*, "Planning with abstract Markov decision processes", in *Twenty-Seventh International Conference on Automated Planning and Scheduling*, 2017.

[45] S. Weisberg, *Applied linear regression*. John Wiley & Sons, 2013.

[46] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning", 2020. arXiv: 2006.04779 [cs.LG].

[47] W. Lin, K. Hasenstab, G. M. Cunha, and A. Schwartzman, "Comparison of handcrafted features and convolutional neural networks for liver mr image adequacy assessment", *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020. DOI: 10.1038/s41598-020-77264-y.

[48] P.-H. Hsu and Z.-Y. Zhuang, "Incorporating handcrafted features into deep learning for point cloud classification", *Remote Sensing*, vol. 12, no. 22, pp. 1–28, 2020. DOI: 10.3390/rs12223713.

[49] F. J. Clayton, C. Sears, A. Davis, and C. Hulme, "Verbal task demands are key in explaining the relationship between paired-associate learning and reading ability", *Journal of Experimental Child Psychology*, vol. 171, pp. 46–54, 2018. DOI: 10.1016/j.jecp.2018.01.004.

[50] K. L. Windfuhr and M. J. Snowling, "The relationship between paired associate learning and phonological skills in normally developing readers", *Journal of Experimental Child Psychology*, vol. 80, no. 2, pp. 160–173, 2001. DOI: 10.1006/jecp.2000.2625.

[51] C. Hulme, K. Goetz, D. Gooch, J. Adams, and M. J. Snowling, "Paired-associate learning, phoneme awareness, and learning to read", *Journal of Experimental Child Psychology*, vol. 96, no. 2, pp. 150–166, 2007. DOI: 10.1016/j.jecp.2006.09.002.

[52] K. Rastle, J. Harrington, and M. Coltheart, "358,534 nonwords: The ARC nonword database", *The Quarterly Journal of Experimental Psychology Section A*, vol. 55, no. 4, pp. 1339–1362, 2002. DOI: 10.1080/02724980244000099.

[53] S. Palan and C. Schitter, "Prolific. ac—A subject pool for online experiments", *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018. DOI: 10.1016/j.jbef.2017.12.004.

[54] J. R. De Leeuw, "JsPsych: A JavaScript library for creating behavioral experiments in a web browser", *Behavior Research Methods*, vol. 47, no. 1, pp. 1–12, 2015. DOI: 10.3758/s13428-014-0458-y.

[55] F. Callaway and C. Correa, *Psirokuturk*, https://github.com/fredcallaway/psirokuturk, 2019.

[56] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. CRC Press, 2020.

[57] J. Bortz and C. Schuster, "Kontraste und Mehrfachvergleiche für einfaktorielle Versuchspläne", in *Statistik für Human-und Sozialwissenschaftler*, Springer, 2010, pp. 221–236.

[58] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences", *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, 2007. DOI: 10.3758/BF03193146.

[59] J. D. Hunter, "Matplotlib: A 2D graphics environment", *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[60] Wes McKinney, "Data structures for statistical computing in Python", in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 –61. DOI: 10.25080/Majora-92bf1922-00a.

[61] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python", *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[62] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python", in *9th Python in Science Conference*, 2010.

[63] R. Vallat, "Pingouin: Statistics in Python", *The Journal of Open Source Software*, vol. 3, no. 31, p. 1026, Nov. 2018. DOI: 10.21105/joss.01026.

[64] W. Kruskal, "Kruskall–Wallis one way Analysis of Variance", *J Am Stat Assoc*, vol. 47, no. 260, pp. 583–621, 1952.

[65] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: https://www.jstor.org/stable/2346101.

[66] N. M. Razali, Y. B. Wah, *et al.*, "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests", *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[67] C. J. Huberty and J. D. Morris, "Multivariate analysis versus multiple univariate analyses.", *Psychological Bulletin*, vol. 105, no. 2, pp. 302 –308, 1992. DOI: 10.1037/0033-2909.105.2.302.

[68] F. L. Huang, "Manova: A procedure whose time has passed?", *Gifted Child Quarterly*, vol. 64, no. 1, pp. 56–60, 2020. DOI: 0.1177/0016986219887200.

[69] T. C. Headrick, *Type I error and power of the rank transform analysis of covariance (ANCOVA) in a 3 x 4 factorial layout*. Wayne State University, 1997.

[70] M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender, "Complexity of finite-horizon Markov decision process problems", *Journal of the ACM (JACM)*, vol. 47, no. 4, pp. 681–720, 2000. DOI: 10.1145/347476.347480.

[71] J. V. Stone, "Bayes' rule: A tutorial introduction to Bayesian analysis", 2013. DOI: 10.13140/2.1.1371.6801.

[72] D. L. Hintzman, "Judgments of frequency and recognition memory in a multiple-trace memory model.", *Psychological review*, vol. 95, no. 4, pp. 528 –551, 1988. DOI: 10.1037/0033-295X.95.4.528.

[73] S. Reddy, I. Labutov, S. Banerjee, and T. Joachims, “Unbounded human learning: Optimal scheduling for spaced repetition”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1815–1824. DOI: 10.1145/2939672.2939850.

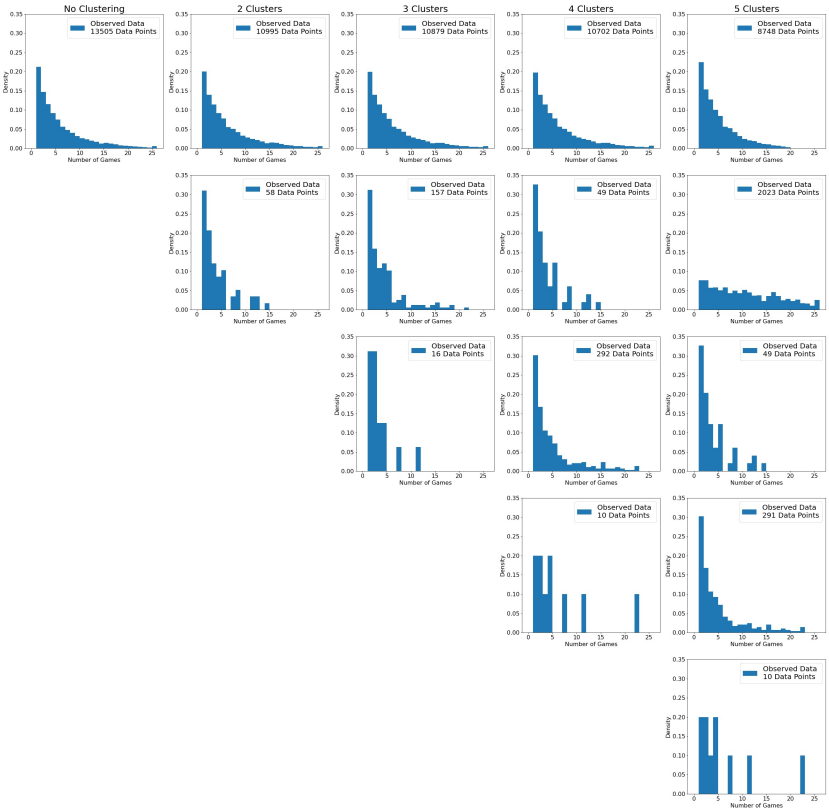
Appendix

A

The code used to derive the results presented in this thesis can be accessed here: https://github.com/pauly-rc/How-to-incentivize-efficient-learning-choices-

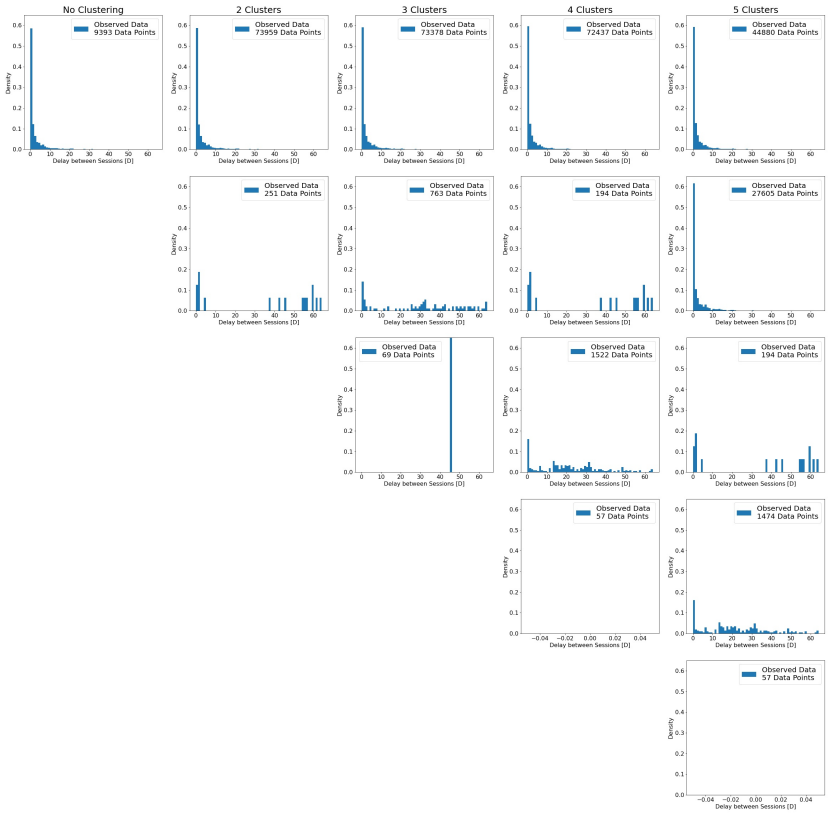
B

Distributions of Games Played per Session, with different numbers of clusters [k-means]



C

Distributions of Delays between Sessions, with different numbers of clusters [k-means]



D

AS PREDICTED

Wharton CREDIBILITY LAB
University of Pennsylvania

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Brain Points: Evaluation Experiment (#78820)

Created: 11/03/2021 06:51 AM (PT)

This is an anonymized copy (without author names) of the pre-registration. It was created by the author(s) to use during peer-review.
A non-anonymized version (containing author names) should be made available by the authors when the work it supports is made public.

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Can incentives based on a computational model improve user's learning behaviour during self-directed interaction with a digital learning environment?
Can incentives based on a computational model can improve user's learning outcome during self-directed interaction with a digital learning environment?
Do approximated incentives have the same effect as optimal incentives?

3) Describe the key dependent variable(s) specifying how they will be measured.
Number of Words learned
Word levels of all Questions
Percentage of choices in which the hardest / easiest task was chosen
Percentage of Choices made in favor of the game with the highest score
Percentage of choices made for the game with the highest / second highest /lowest number of optimal brain pints (for all conditions)
Percentage of choices made for the game with the highest / second highest /lowest number of shown brain pints (for conditions 1 and 3)

4) How many and which conditions will participants be assigned to?
There are 5 conditions.
Control Condition: Participants are not shown any choice incentives.
Optimal Brain Points Condition: Participants are shown optimal choice incentives.
Optimal Forced Choice Condition: Participants do not choose, but interact with the task computed to be optimal.
Approximated Brain Points Condition: Participants are shown approximated choice incentives.
Approximated Forced Choice Condition: Participants do not choose, but interact with the task approximated to be optimal.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Test for an effect of condition on Learning Progress and Choice behaviour:
As a first step, it will be assessed whether the collected data is well-modeled by a normal distribution by using visual expectation and the Shapiro-Wilk Test. Depending on the results, either the One-way ANOVAs or the Kruskall-Wallis H test will be used to test for statistically significant differences between the 3 levels of the factor incentive points (None, approximated, optimal) with regard to the dependent variables. Depending on the results, post-hoc tests (Benjamini-Hochberg Procedure with either Mann-Whitney or paired t-tests) will be performed.
Two-way ANOVAs or Friedman Test with the factors incentive points (approximated, pütimal) and choice (free, forced) will be used to evaluate the main effect of choice and possible interactions between points and choice on the dependent variables. Depending on the results, post-hoc tests (Benjamini-Hochberg Procedure with either Mann-Whitney or paired t-tests) will be performed.
As the effect of the experimental conditions can be expected to be more impactful once a participant masters one of the skills (= meets the criteria of having learnt 80% of the word pairs pertaining to the skill) , the above described analyses will be conducted both for all choice trials and for choice trials meeting that condition.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Participants who failed at least two of the four attention checks will be excluded. Participants who failed to respond within the time limit in more than ¼ of the 160 trials will be excluded.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We plan to recruit 300 participants (distributed evenly across conditions) through prolific.co. Due to random assignment, exclusions, and incomplete responses, the exact number of participants per condition can vary.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)