Real-World Testing Matters in Reinforcement Learning for Education

Anna Riedmann
Socially Interactive Agents, University of Würzburg
Würzburg, Germany
anna.riedmann@uni-wuerzburg.de

Carlo D'Eramo
Center for Artificial Intelligence and Data Science, University of Würzburg
Würzburg, Germany
Technical University of Darmstadt
Darmstadt, Germany
carlo.deramo@uni-wuerzburg.de

Birgit Lugrin
Socially Interactive Agents, University of Würzburg
Würzburg, Germany
birgit.lugrin@uni-wuerzburg.de

ABSTRACT

Deep Reinforcement Learning (DRL) has proven its usefulness across various fields, sparking growing interest in applying it to education. However, most research on DRL in educational applications utilizes methods in simulation, with little evaluation involving real learners, resulting in limited evidence of their effectiveness in real-world contexts. Arguably, we consider real-world applications and in-situ experiments with users as essential for a thorough evaluation. We thus propose *ResUli-RL*, a novel DRL approach rooted in educational psychology, designed to provide adaptive feedback to young learners in the form of a pedagogical agent in a mobile educational app. To investigate its effectiveness, we conducted a five-week real-world evaluation with 56 primary school students, comparing *ResUli-RL* to an expert-designed baseline. Both groups significantly improved in reading competence, with no significant differences between them and a notable decrease in motivation in both conditions. In our aim to further improve the children's reading competence using DRL, our approach did, however, not yield the expected results. Our findings provide guidance for future work and highlight the need for real-world evaluations in education to assess the value of an educational DRL approach.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → **Field studies**; • **Computing methodologies** → **Reinforcement learning**.

KEYWORDS

deep reinforcement learning; residual learning; technology-supported learning; pedagogical agent

ACM Reference Format:
Anna Riedmann, Carlo D'Eramo, and Birgit Lugrin. 2025. Real-World Testing Matters in Reinforcement Learning for Education. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Reinforcement Learning (RL), a method of learning an optimal mapping of state-action pairs to optimize a numerical reward [49], has emerged as a powerful method for addressing complex sequential decision-making tasks in several domains due to its applicability to different scientific and engineering disciplines [47]. Deep Reinforcement Learning (DRL) further leverages neural network models to enhance traditional RL algorithms [51], thus allowing these to handle high-dimensional learning problems [47], and applications, such as natural language processing (e.g., [54]), or robot control tasks (e.g., [1]). We have also been witnessing a growing interest in applying (D)RL in education, demonstrating promising results in the area of, for example, personalized learning [20]. However, despite the promising potential demonstrated in existing literature [18], these approaches are often not evaluated with real learners, resulting in a gap when it comes to translating simulation results into practice, providing limited evidence of their effectiveness in addressing real-world problems in education. Further, RL in education appears to be most effective when informed by principles and theories from cognitive psychology and the learning sciences, and requires approaches to be tested in real-world settings against sophisticated baselines [18].

To address this, we introduce a new approach for incorporating DRL into an empirically validated digital reading application for primary school students, guided by knowledge from the cognitive sciences, implemented as an adaptive Pedagogical Agent (PA). PAs are virtual characters aiming to support the learning process [30], allowing for peer-like interaction and a personalized learning experience for children [9]. We employ DRL to adapt the agent's feedback behavior to each child's individual learning process and evaluate the model in a real-world experiment. For this purpose, we compare it to an expert-designed baseline in the form of an empirically validated reading app for primary students, allowing them to practice various reading skills on different difficulty levels. Thus, our contribution is twofold: 1) We present a new approach to the Deep Q-Network (DQN) algorithm [35] (*ResUli-RL*), where the components of the Markov-Decision-Process (MDP) are inferred from educational psychology, and 2) conducted a five-week intervention within the app's target group to assess its effects on the children's reading competence and motivation in the long-term, comparing the *ResUli-RL* model to an expert-designed baseline. In the following, we describe our novel concept and the results and implications of the real-world evaluation.

2 RELATED WORK

2.1 (Deep) Reinforcement Learning

Reinforcement Learning involves determining the optimal actions to take in situations in a trial-and-error process in order to maximize a numerical reward signal [49]. Deep Reinforcement Learning further incorporates neural network models into conventional RL algorithms and has proven its usefulness in various application areas, for example, end-to-end control, robotics, investment or recommender systems [22, 51]. RL is commonly framed as a Markov Decision Process, where an agent acts as the decision-maker, taking an action at each time step within an environment, entailing changes in the environment's state with the agent ultimately aiming to maximize a reward [49]. MDPs are characterized by a tuple $\langle S, A, p, R, \gamma \rangle$, where S represents the set of states the agent can be in, and A denotes the set of available actions at each time step t within a given state, receiving a numerical reward R in the subsequent time step [18, 19, 49]. The probability distribution function $p(s'|s, a)$ specifies the likelihood of transitioning from state s to state s' after executing action a and receiving a reward r discounted by $\gamma \in [0, 1]$ [16, 39, 49]. The interaction process within an environment unfolds as follows: Starting in a specific state $s_0 \in S$, the agent makes an initial observation $\omega_0 \in \Omega$ (with Ω as the set of possible observations). At each time step t, the agent then chooses an action $a_t \in A$, leading to a reward $r_t \in R$, a transition to the next state $s_{t+1} \in S$, and a subsequent observation $\omega_{t+1} \in \Omega$ [22].

2.2 Reinforcement Learning in Education

One field of application that seems to be increasingly researched is the application of RL in education. Recent reviews by Doroudi et al. [18] and Fahad Mon et al. [20] indicated promising learning-related outcomes in educational settings. Half of the studies (21/36) that applied RL for instructional sequencing in education and evaluated it in a real-world experiment showed that at least one RL policy significantly outperformed all baselines, with the majority reporting a substantial Cohen's d effect size of 0.8 [18]. Recent examples involve using RL for scheduling linear algebra course activities, with learners in the RL condition achieving significantly higher learning gains compared to an expert-designed baseline [4], or RL-based personalized story selection in literacy education for children, resulting in significantly higher engagement and learning outcomes in comparison to a non-adaptive condition [37].

RL has also been applied in education to induce pedagogical policies, through providing adaptive feedback and hints. For example, Chi et al. [13] used RL to select suitable tutorial decisions (e.g., either prompting the student for the next problem-solving step or providing it directly) and evaluated them in a college-level physics course, where the proposed method significantly surpassed both the RL and random baselines in terms of students' learning performance. Further, Zhou et al. [55] demonstrated the effectiveness of applying a pedagogical decision policy using RL, with the proposed method significantly outperforming the baseline in terms of students' post-test score. Focusing on primary education, Chen et al. [12] leveraged RL to adapt an agent's interactive behavior to the child's knowledge level. The RL-adapted agent helped children learn the most words and evoked positive emotions significantly more often.

While there exists a growing number of research conducting real-world experiments as presented above, reviewing related work shows that out of 80 publications applying RL in education, half of them (37/80) rely on interaction datasets or simulated data [44]. Considering that real-world experiments are crucial to accurately evaluate the effectiveness of RL on students' learning outcomes [18], this highlights the need for further research regarding real-world RL applications. Real-world evaluations entail several ethical and practical challenges (e.g., vulnerable target groups, constraints of real-world classrooms). The complexity of real-world environments thus poses challenges for implementing RL systems in practical settings [19]. Similar to the challenges highlighted by Dulac-Arnold et al. [19] regarding recommender systems, RL-driven educational applications face numerous and diverse optimization goals (e.g., enhancing motivation and learning gain, reducing dropout rates, and improving algorithm performance). The need for real-time interaction further complicates matters, as the system must make immediate decisions. Particularly in education, extensive offline datasets are often scarce, thus RL-based systems depend heavily on such data logs, as online experimentation tends to be costly.

Additionally, approaches deeming the RL policy to be significantly superior often apply weak baselines (e.g., random or poorly designed models), whereas expert-crafted baselines are used less often [18]. In the context of instructional sequencing, Doroudi et al. [18] summarized that 71% of approaches that demonstrated significant beneficial effects of their adaptive RL policies applied random baselines or other RL policies, which have not demonstrated robust performance, instead of being compared to state-of-the-art baselines. This implies that although a personalized learning process might be helpful, it does not clarify whether RL-based approaches lead to significantly better outcomes compared to relying on expert-driven strategies [18]. Moreover, RL in education seems to be most successful when framed by concepts and theories from psychology and the learning sciences, indicating that RL approaches should more frequently draw on insights from educational research extensively explored by psychologists, rather than relying solely on a data-driven approach [18].

2.3 Pedagogical Agents

Educational applications often integrate a virtual character, a so called pedagogical agent (PA), that "[...] seeks to promote learning, enhance motivation, and provide support to engage in an educational activity" [30, p. 307]. It can take on different roles in educational applications, comprising the role of 1) an expert agent demonstrating expertise or deep knowledge of the domain, 2) a peer-like motivator agent aiming to engage the learner with the learning tasks, 3) a mentor agent as a hybrid version offering both information and encouragement, and 4) a student-like learner agent interacting with an expert or mentor agent [5, 15], with the expert agent being the most commonly used role [15].

PAs seem to have an overall small positive effect on students' performance compared to learning environments without an agent [10, 46], with 2D agents being slightly more effective than 3D agents [10]. They can also be beneficial for student motivation [45]. Combining these agents with methods from Artificial Intelligence further enables the introduction of adaptivity in the behavior of PAs. This

could enhance the learning process by allowing students to advance through the learning environment at their own pace, facilitated by individually scheduled learning activities and personalized learning pathways [2]. They also allow for peer-like interaction, promoting children's development and learning [9].

3 *RESULI-RL*: CONCEPT AND IMPLEMENTATION

In previous work, we proposed the initial framework for our DRL-powered PA providing automated feedback behavior [41] and evaluated it in its target group of primary students, demonstrating promising results on real-world feasibility and a long-term motivational potential [40]. Building on this, we present our approach *ResUli-RL*, a **Res**idual **RL**-powered PA named "**Uli**" that provides individually adapted feedback within the framework of a validated reading intervention. Thus, we aim to account for individual differences in children's learning behaviors and abilities, while also improving motivation and reading competence. In the following, we describe our baseline learning environment, the proposed novel DRL approach, and model pre-training.

3.1 The *MobiLe* Reading App (Expert Baseline)

The environment used for implementing and testing our approach is built on an empirically validated digital reading application [24, 25, 42], which is itself based on a validated analogue reading intervention for second graders with reading difficulties [36]. The digital training is a mobile app and was developed in close collaboration with experts from pedagogical psychology. It has been validated within the target group in a 20 session wait-list-control group design where children using the app demonstrated significant gains in general word recognition and phonological recoding processes compared to similarly low-skilled children who received no intervention [24].

The app comprises 21 games designed to enhance various reading skills, focusing on phonological and orthographic comparison processes, as well as reading comprehension. Games involve different reading-related activities, for example separating word items syllable by syllable or reading short stories. Several games have also been evaluated within the target group regarding its usability and enjoyment [42]. All games are wrapped in a training structure, allowing for different difficulty levels. The app also includes a PA named "Uli, the owl" (see Figure 1) that provides feedback and support after a set number of incorrect attempts, following a predetermined behavioral sequence defined by experts from pedagogical psychology.

The PA either remarks on the accuracy of the child's response, provides the correct solution, or offers help by directing the child to a tutorial video. In earlier work, we highlighted the benefits of incorporating adaptive mechanisms rooted in educational psychology into both the learning environment and the behavior of the PA [43]. Our previous method relied on a fixed feedback system, which we now seek to improve by leveraging DRL to automate the type of the agent's feedback behavior. This automation enables the PA to provide real-time, personalized feedback tailored to each child's learning process. In this context, personalized feedback means that the type of feedback on mistakes varies for each learner (either

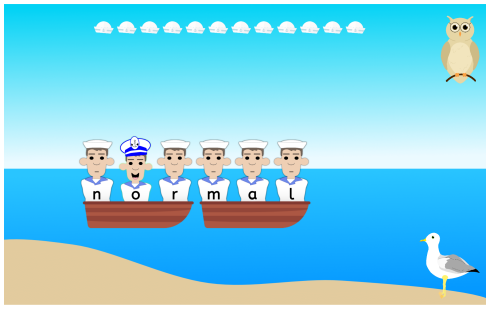

Figure 1: The *Sailor Game* with the pedagogical agent "Uli, the owl" in the top right corner

motivating feedback, the solution, or a hint), depending on their learning behavior. The original digital reading intervention used as our learning environment comprises 20 sessions, which exceeded the scope of our planned real-world experiment. For our approach, we thus selected a subset of games that cover all central aspects of the digital reading training while maintaining some variation. We adopted a reduced form of the original reading training structure, resulting in 13 different games (namely Anthill, Balloon Ride, Bee Flight, Butterfly Flight, Picnic, Sailor Game, Sea Game, Syllable Drum, Syllable Salad, Syllable Soup, Uschis's Post, Word Bakery, Word Signals) that are played multiple times on different difficulty levels while progressing through a fixed training structure. A brief description of all games can be found in Heß et al. [24].

3.2 *ResUli-RL*

We consider our environment as a MDP, consisting of states, actions, rewards, and probabilities that capture the dynamics of the decision-making process [49]. State feature variables are used to represent the state, encapsulating the history of the learning session and preserving all relevant details about both past and current interactions. The state features are selected based on key factors that influence the perception and effectiveness of feedback in educational science (see [7, 23, 53]).

They comprise the difficulty of the game currently played ($1-5$) and the support given by the PA (i.e., referring to the tutorial video), both for the word item currently worked on and the entire game session. Further, the likeability (i.e., how motivated children currently are to engage with the app) is queried regularly (every five tasks) within the app. A pop-up is displayed and children are prompted to adjust a slider between $0-1$ (with 1 as highest likeability), aiming to include motivational aspects into the state space, as suggested by Doroudi et al. [18]. Additionally, the number of mistakes made by the learner in the current game session is monitored, along with the reaction time for each interaction. Together, all specified features form the current state $s = \{GameDifficulty, SupportPerWord, SupportPerSession, Likeability, Mistakes, ReactionTime\}$.

For each incorrect answer given by the child during app interaction, the DRL-powered PA chooses one out of three feedback behaviors, thus the set of actions is defined as $A = \{GiveFeedback, ProvideHint, ShowSolution\}$. For selecting a respective action, we used a residual approach to the Deep Q-Network (DQN) algorithm [35]. Residual RL is a concept in which a RL agent uses a

residual function to improve or adjust an existing control policy or value function [52]. It is built on the premise that an agent doesn't need to learn the full policy or value function from scratch, but can instead adjust or improve an already existing, often heuristic-based or pre-trained Q-function. Residual RL is frequently used to improve stability and enhance generalization [52] and has been applied in various domains, such as robotics [27, 52], value function factorizing in Multi-Agent RL [38], or other control tasks [31]. Our proposed approach adjusts the concept of residual RL with a focus on error correction and bias reduction. For *ResUli-RL*, the residual Q-value represents the difference between the actual Q-value and a pre-trained Q-value (estimated as described in Section 3.3). We expect the pre-trained Q-value to provide a strong initial estimate, and the goal of the residual network is to learn the necessary adjustments to this estimate to refine the Q-function further. In traditional DQN, as described by Mnih et al. [35], the Q-value function $Q(s, a)$ is directly approximated by a neural network, which maps states s and actions a to their corresponding expected cumulative rewards. However, this approach can suffer from high variance and slow convergence, especially in environments where a rough prior of the Q-values can be established. To address these challenges, our approach decomposes the Q-value function into two components: A fixed, pre-trained Q-function $Q_{pre}(s, a)$ and a target Q-function $Q_{target}(s, a)$. The residual Q-value $Q_{residual}(s, a)$ is computed as the difference between the target Q-value and the fixed component:

$$Q_{residual}(s, a) = Q_{target}(s, a) - Q_{pre}(s, a) \tag{1}$$

Our approach allows the residual network to focus on learning the correction to the pre-trained baseline provided by $Q_{pre}(s, a)$, rather than learning the entire Q-function from scratch. This approach allows for bias adjustment and error correction between pre-trained and actual target values to avoid compounding errors. This method is particularly advantageous in this context since $Q_{pre}(s, a)$ is derived from pre-training on a dataset collected in the actual application context (see Section 3.3), thus serving as an informative yet biased estimate that requires refinement. Subtracting the pre-trained Q-value $Q_{pre}(s, a)$ can help highlight where the pre-trained value is overestimated or biased, mitigating overfitting to the fixed Q-value. The pre-trained value provides an initial estimate of the Q-values, capturing general trends and relationships within the environment, and remains static during subsequent training.

The target Q-value $Q_{target}(s, a)$ is approximated by a deep neural network. As in DQN, we used two neural networks with identical architectures, with both the Q-network and target network being structured as a two-layer sequential neural network (64, 64), utilizing the $Tanh$ function for activation, with the final layer corresponding to the dimensions of the action space. Input feature values are normalized using min-max scaling, and the algorithm is fine-tuned with the Adam stochastic optimization method [28]. We estimate $Q_{target}(s, a)$ following the Bellman Equation in that:

$$Q(s, a) = r + \gamma \max_{a'} Q(s', a') \tag{2}$$

In this context, r represents the expected immediate reward for choosing action a in state s, γ is the discount factor, and $Q(s', a')$ is the expected value of $Q_{target}(s, a)$ in the subsequent state s' when taking an action a' and then adhering to policy π afterwards. The Q-Network is then fine-tuned during online learning through live interactions with the target group (see Figure 2), where it updates the residual Q-function by learning the discrepancy between the fixed Q-values and the actual returns observed during interaction with the environment, see Equation 1.

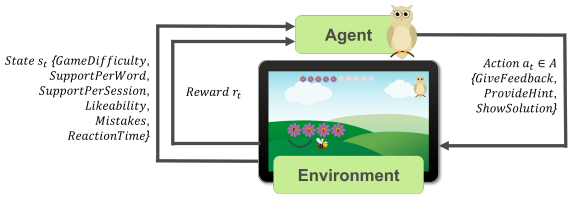

Figure 2: Agent-environment interaction at time-step t

Further, the target network is updated based on a child's next interaction after making a mistake and receiving feedback from the DRL-powered PA. We opted for the Huber loss function as a more robust alternative to MSELoss. To balance exploration and exploitation, the DRL-based PA makes use of the epsilon-greedy policy, selecting the action with the highest estimated reward. Initially, we set $\varepsilon_{init} = 4.5e-1$, with linear epsilon decrement for each interaction with $\varepsilon_{decay} = 2e-4$, to end at $\varepsilon_{end} = 8e-2$. The learning rate is set as $lr = 1e-1$. Algorithm 1 shows the pseudo-code for training *ResUli-RL* online.

Algorithm 1 *ResUli-RL*

Input Live training data D, $Q_{pre}(s, a)$, current residual predictions $Q(s, a; \theta)$ with parameters θ
for $k = 1, 2, 3, \ldots$ **do**
$Q_{target}(s, a) = r + \gamma \max_{a'} Q_{target}(s', a')$
$Q_{residual}(s, a) = Q_{target}(s, a) - Q_{pre}(s, a)$
$\theta \leftarrow \theta - lr(Q(s, a; \theta) - Q_{residual}(s, a))\nabla_\theta Q(s, a; \theta)$
end for

Rewards are assigned based on the accuracy of the child's answer, either 1 for correct or 0 for incorrect answers. A discount $\xi \in [0, 1]$ is applied for previously given incorrect answers, calculated as the accuracy divided by the total number of answers given in the current game. Thus, we denote $r = \xi * accuracy$.

3.3 Offline Training to Estimate Q_{pre}

To estimate Q_{pre}, we trained offline on a previously collected dataset of the app's target group which contains interaction data of 144 children, working through the reading training in the app in 20 sessions. They played the selected 13 games multiple times, involving approximately $40-60$ interactions per game. The data collected encompasses accuracy, reaction time, difficulty level, and specific word characteristics. All other necessary information for training, such as actions and states, was inferred from the dataset. We used Boosted Fitted Q-Iteration (BFQI) [50] to estimate the Q-function based on the given dataset, accounting for the continuous state space. BFQI is an approximated value iteration algorithm which estimates the action-value function in RL problems utilizing a boosting technique. Given a pre-collected dataset of transitions, BFQI

approximates the optimal action-value function by aggregating the approximations of the Bellman residuals over multiple iterations, see Algorithm 2.

Algorithm 2 Boosted Fitted Q-Iteration

Input Training dataset D, $Q_0 = 0$
  for $k = 0, ..., K$ **do**
    $\tilde{\varrho}_k \leftarrow \hat{T}^* Q_k - Q_k$ (w.r.t. $D^{(k)}$)
    $Q_{k+1} \leftarrow Q_k + \tilde{\varrho}_k$
  end for
    return $\bar{\pi}(s) = \arg\max_a Q_{K+1}(s, a) \quad \forall_s \in S$

Accordingly, as noted by Tosatto et al. [50, p. 3435] "the complexity (e.g., supremum norm) of fitting the Bellman residual should decrease as the estimated value function approaches the optimal one [...], thus allowing to use simpler function approximators and requiring less samples". With it, we aimed to compensate for both the continuous state space as well as the small amount of historical learner data available for training, since simulating young learners' behavior is not possible and learning applications have to be ultimatively evaluated in real-world scenarios, requiring in-situ studies with children. We adopted the BFQI implementation of the MushroomRL library [17] and adapted it for our setting, using the GradientBoostingRegressor to approximate the Q-function. On the dataset, we achieved an average cumulative reward of $\mu_{G_{pre}} = 51.88$ and Q_{pre} was estimated as $Q_{pre} = 0.75$.

4 REAL-WORLD EVALUATION

We assessed the effectiveness of the DRL-powered PA in a real-world setting, examining whether the DRL-induced policy outperformed the expert baseline in a school environment regarding students' reading competence and motivation. For our experiment, we connected the DRL model with the existing reading app using the Chaquopy plugin [11], enabling the PA to determine the appropriate feedback behavior whenever a child made a mistake while using the app. We expected a higher increase in reading competence scores (H1) and post-test motivation (H2) in the group using the app with the DRL-powered PA compared to the baseline condition. Further, we investigated the feasibility and performance of the *ResUli-RL* approach (RQ1).

4.1 Method

4.1.1 Measures. To measure the children's reading competence, we administered a standardized screening test (Stolperwörter-Lesetest (STOLLE) [34]) that measures reading fluency and -ability at the sentence level. As noted by Metze [34], the test implicitly includes the assessment of meaning and syntactic coherence. In 60 sentences, the children's task is to identify the word that doesn't align with the overall meaning of the sentence, such as the word "young" in the example: "My friend is eight young years old.". A percentile rank can then be derived for each child's raw score (number of correctly solved sentences), indicating the percentage of children in the tested sample who performed the same or worse. Percentile ranks $25 - 74$ correspond to an average performance in reading. We assessed the children's reading competence before and after using the app for five weeks. Three children had to be excluded due to missing pre-test values. The test–retest reliability, calculated based on pre- and post-measures in both conditions, was computed as the intraclass correlation of the STOLLE percentile ranks for a total of 53 children (those with complete data sets). For calculating the intraclass correlation coefficient (ICC), a two-way mixed effects model for mean rating and consistency was used [33, 48]. According to Koo and Li [29], the estimated test–retest reliability was good with $ICC(3, k) = .87$, 95% $CI[0.78, 0.93]$.

Motivation was assessed after the initial use of the app and again on the final day of the intervention after five weeks. Five children had to be excluded due to missing values. We used the Reduced Instructional Materials Motivation Survey (RIMMS) [32]. The questionnaire contains four subscales, namely *Attention*, e.g., "I liked that there were different tasks.", *Relevance*, e.g., "It helps me if I can read well.", *Confidence*, e.g., "I was sure from the beginning that I would manage the tasks.", and *Satisfaction*, e.g., "I had a lot of fun playing.". Each scale comprises three items measured on a five-point Likert scale, with the wording slightly adapted to the target group as described in Riedmann et al. [43]. An overall score across all scales measures motivation. The test–retest reliability was computed as the intraclass correlation of the overall RIMMS score (pre- and post-measures in both conditions) for all children with complete datasets ($n = 51$). For calculating the ICC, a two-way mixed effects model for mean rating and consistency was used [33, 48] and according to Koo and Li [29], the estimated test–retest reliability was moderate with $ICC(3, k) = .62$, 95% $CI[0.35, 0.78]$.

To assess the perception of the pedagogical agent, we employed the Agent Persona Instrument (API) [6] at the final day of the intervention, containing four subscales. They comprise 10 items for *Facilitating Learning*, e.g., "The agent made the instruction interesting." ($\alpha = .93$), and five items each for *Credibility*, e.g., "The agent was intelligent." ($\alpha = .77$), *Human-Likeness*, e.g., "The agent was human-like." ($\alpha = .86$), and *Engagement*, e.g., "The agent was friendly." ($\alpha = .90$), each measured on a five-point Likert scale.

Additionally, we tracked the children's performance and number of attempts in all games through the reading intervention to derive their accuracy ratio $\phi \in [0, 1]$, calculated as the number of correct answers given in relation to the overall number of interactions, used as discount ξ for reward calculation. As input for the DRL model, we logged the difficulty, support per word and session given by the PA, likeability, reaction time, and mistakes made by the child. Further, cumulative reward and loss values for each network time step were logged to calculate the average cumulative reward (ACR) and average Rate of Change (ARoC) as indicators of how well our model is learning. Data of five children of the *ResUli-RL* condition had to be excluded from this analysis due to technical difficulties.

4.1.2 Participants and Procedure. The real-world evaluation of *ResUli-RL* was conducted as a five-week intervention study in a German primary school, involving two second grade and one third grade school classes, in close consultation with responsible teachers, school administration, and the education authority. Sixty-one primary school students voluntarily agreed to participate in the experiment. Parents were informed about the study in advance through an information letter and provided written informed consent for all participants. One child had to be excluded due to technical difficulties and four children did not participate in post-testing, resulting in

$N = 56$ children, aged 7–10 years ($M = 8.45$, $SD = 0.71$). Thirty-one girls and 25 boys participated, none identified as diverse. They were randomly assigned to either the condition with the adaptive PA (*ResUli-RL*, $n = 29$) or the control group using the expert-designed version with the non-adaptive PA (ExpertApp, $n = 27$).

All participating children used the app regularly once or twice a week for 45 minutes for in total five weeks, supervised by an experimenter during school time. The first and last session included filling out the respective paper-based questionnaires and the STOLLE. The children played a subset of the original reading training on a 10.5-inch Android tablet, involving 13 games that were repeated several times in different difficulties through the course of the training, with varying focus on different reading aspects. The sequence, game types, and the content of the PA's feedback in the learning app were the same for both conditions, while the feedback selection differed in being either on a fixed schedule in the ExpertApp condition or individually adapted to the child by the DRL model in the *ResUli-RL* condition. Our experiment was approved by the responsible institutional ethics committee (application number 150524).

4.2 Results

All analyses were conducted with JASP [26] and alpha was set at .05. We used mixed-design analysis of variance (ANOVA) to investigate the children's reading competence and motivation, with time as the within-subjects factor and condition as between-subjects factor. As indicated by Levene's test, equality of variances was not given with $p < .05$, however, repeated-measures ANOVA (and thus mixed ANOVA) can be considered robust to non-normality [8]. An a priori power analysis was conducted using G*Power [21] to estimate the sample size needed for a mixed-design ANOVA. The analysis aimed to detect a medium effect size ($d = 0.50$) [14], using an alpha level of .05 and a desired power of .95, resulting in a suggested minimum sample size of 54 participants. The sample size of 56 participants exceeded this minimum value, suggesting adequate statistical power for detecting group differences.

To compare the children's perception of the agent between conditions, a one-way multivariate analysis of variance (MANOVA) was used. The Shapiro-Wilk test for multivariate normality indicated significance ($p < .001$), however, MANOVA is considered relatively robust against violations of equality of variance [3]. Homogeneity of covariances was given, as assessed by Box's test ($p = .117$). Table 1 displays all descriptive values.

4.2.1 Reading Competence. There were no significant differences among the children's STOLLE pre-test scores ($U = 320.00$, $p = .587$, $d = -0.09$), thus we assume a similar range of reading skills in both groups. We investigated whether children in the *ResUli-RL* condition improved their reading skills more from pre- to post-test compared to the ExpertApp group. The mixed-design ANOVA revealed a significant main effect for time on reading competence, $F(1, 51) = 50.03$, $p < .001$, $\eta^2 = 0.10$. Overall, children improved their reading ability from pre-test ($M = 35.91$, $SD = 23.21$) to post-test ($M = 52.02$, $SD = 25.76$). Post-hoc tests showed that this effect applied for both the *ResUli-RL* condition ($t(26) = -4.55$, $p_{holm} < .001$, $d = -0.59$), and the control group ($t(25) = -5.44$, $p_{holm} < .001$, $d = -0.72$), see Figure 3. However, there was no significant difference between the two conditions ($F(1, 51) = 0.241$, $p = .626$, $\eta^2 = 0.00$). We also found no significant interaction effect between time and condition ($F(1, 51) = 0.484$, $p = .490$, $\eta^2 = 0.00$).

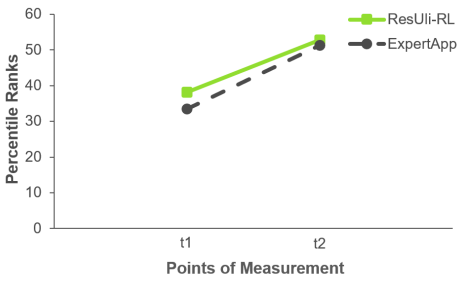

Figure 3: Pre-/post-test improvement of the children's reading competence for both conditions

4.2.2 Motivation. We further investigated whether the children's motivation differed from first to last app usage between the two conditions. A mixed-design ANOVA showed a significant effect for time on motivation, $F(1, 49) = 23.10$, $p < .001$, $\eta^2 = 0.11$. Overall, motivation significantly decreased from pre-test ($M = 4.55$, $SD = 0.49$) to post-test ($M = 4.12$, $SD = 0.71$). Post-hoc tests revealed significantly reduced motivation in the *ResUli-RL* condition from pre- to post-test ($t(25) = 4.30$, $p_{holm} < .001$, $d = 0.87$), but not for the ExpertApp condition ($t(24) = 2.51$, $p_{holm} = .062$, $d = 0.52$). There was neither a significant difference between the two conditions ($F(1, 49) = 0.01$, $p = .921$, $\eta^2 = 0.00$) nor a significant interaction effect between time and condition ($F(1, 49) = 1.49$, $p = .228$, $\eta^2 = 0.01$).

4.2.3 Perception of the PA. We additionally compared the perception of the PA between the two conditions, subdivided according to its ability to facilitate learning, credibility, human likeness, and engagement. Using these aspects as dependent variables and condition as independent variable, we conducted a one-way MANOVA. We found no significant main effect for condition, $Wilks' \Lambda = .92$, $F(4, 42) = 0.95$, $p = .445$.

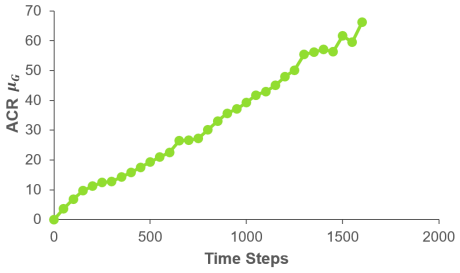

Figure 4: Average Cumulative Reward (ACR) μ_G for time steps reached in app interactions by at least 50% of the children using the app version with the DRL-powered PA

Table 1: Means for both conditions with SD in parentheses

	Scale	*ResUli-RL*			*ExpertApp*		
		Pre	Post	*n*	Pre	Post	*n*
Reading competence	0-100	38.22(26.32)	52.78(28.49)	27	33.50(19.72)	51.23(23.14)	26
Motivation	1-5	4.61(0.37)	4.08(0.70)	26	4.49(0.60)	4.17(0.72)	25
Facilitating learning	1-5		3.70(1.01)	28		4.03(0.88)	25
Credibility	1-5		4.10(0.84)	29		4.21(0.75)	26
Human likeness	1-5		3.64(1.07)	28		3.78(1.06)	24
Engagement	1-5		3.85(1.22)	27		3.95(1.02)	26

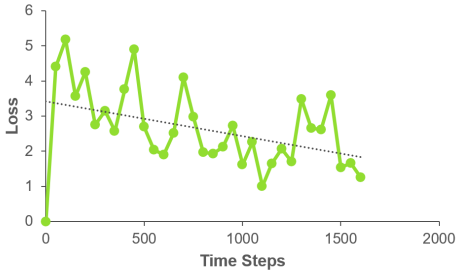

Figure 5: Average loss for time steps reached in app interactions by at least 50% of the children using the app version with the DRL-powered PA

Figure 6: Average Rate of Change (ARoC) $\overline{\Delta L}$ for time steps reached in app interactions by at least 50% of the children using the app version with the DRL-powered PA

4.2.4 Additional Measurements. In the *ResUli-RL* condition, the children ($n = 29$) played 27 games on average during the five-week intervention, resulting in approximately 67 interactions per game for the DRL agent. These children further made averagely 544 mistakes throughout the entire app interaction, which is equal to the mean number of actions selected by the PA per child and thus the training episodes. To capture both the agent's effectiveness across multiple users and its ability to learn over time, we calculated the ACR. As our DRL model was trained live in a continuous interaction setting, we calculated the ACR μ_G as $\frac{1}{n}\sum_{i=1}^{n} G_i$, averaged by the number of children ($n = 29$). G_i is computed as $\sum_{t=0}^{T} r_{i,t}$, with $r_{i,t}$ as the reward received at time step t for child i. For our context, this resulted in $\mu_G = 73.43$, which is an increase by 41.54% compared to the ACR achieved when trained on historical data. The ACR values for *ResUli-RL* during live training are displayed in Figure 4, showing that while rewards continuously increased, the policy was not able to fully converge.

Figure 5 shows a general downward trend of the Huber loss, indicating that the DRL model was learning and improving during the live training, but the learning process was not very smooth as visible by the large oscillations. This continued fluctuation is reflected in the ARoC (see Figure 6), suggesting that the policy was not converging effectively. The ARoC $\overline{\Delta L}$ represents the difference between the current and the previous loss value for each child at each time step, thus $\overline{\Delta L} = \frac{1}{n-1}\sum_{i=1}^{n-1} \frac{L(t_{i+1})-L(t_i)}{t_{i+1}-t_i}$, averaged over the children, resulting in an overall small, but positive value $\overline{\Delta L} = 0.06$. Please note that Figures 4 to 6 display only results of a subset of children in the *ResUli-RL* condition to account for the high variation in time steps between children while being able to show a gradient. This variation arises from differences in the speed at which children progressed through the app, necessitating a cut-off at 1600 time steps — the maximum number reached by at least 50% of the children during app interactions.

5 DISCUSSION

With our work, we aimed to address the gap in real-world research on applying RL in education. We presented our novel approach *ResUli-RL*, a DRL-powered PA in a reading app for primary students, and demonstrated its feasibility in a real-world experiment. We leveraged knowledge from educational psychology to infer suitable state features and actions, and compared the DRL-powered application to an expert-designed baseline. In a five-week primary school intervention, we collected data on the children's motivation and reading competence to investigate long-term effects.

Children improved reading competence in both conditions. We did not find differences in the increased reading competence scores between the two conditions, thus rejecting H1. Interestingly, children significantly improved their reading competence from pre- to post-test in both groups, suggesting that the reading training achieved the desired effect regardless of whether the feedback was personalized or not. This aligns with previous work [24], demonstrating the reading training's general effectiveness. Further, the high standard deviation in both groups' pre- and post-test means indicates the children's high diversity in terms of their prior reading abilities, although being equally distributed across conditions.

High post-test motivation with an overall decrease. There was also no significant difference between the children's motivation in the post-test between conditions and we reject H2. Both groups demonstrated descriptively lower means in motivation in the post-test compared to the pre-test, with a significant difference for the *ResUli-RL* condition, suggesting a significant decrease in motivation. The fixed feedback request intervals in the *ResUli-RL* condition might have negatively impacted the children's experience. However, overall post-test motivation can still be considered high in both conditions with a mean of 4.12 on a five-point Likert scale, considering the overall long time children interacted with the app. The lack of significant differences between the conditions might be due to the initially high motivational potential of the expert-designed baseline and further stretches the need to compare new educational RL approaches to sophisticated baselines. Further, adjusting only the feedback type might not have sufficiently personalized the experience, probably requiring a greater variety of feedback types (i.e., actions), however, for comparability we aimed to align with the original reading training as close as possible.

***ResUli-RL* provided personalized feedback, though not fully converged.** Regarding the performance of the DRL model (RQ1), it seems that our *ResUli-RL* approach, while demonstrating increasing rewards and a small downwards trend of the average loss values, was not able to fully converge within the given real-world setting. Looking at the ACR values, the agent progressively improved its performance over time, however, the fluctuations near the end suggest that the agent was still fine-tuning its policy and had not fully converged yet. However, considering the relatively high learning rate $lr = 1e{-}1$, this smooth increase suggests that the agent was managing to adapt effectively, even with limited interaction steps (544 episodes per child on average). This is reflected in the *ResUli-RL* group interaction data, as children that struggled more often received more solutions by the PA, while those previously performing well were given more opportunities to try again.

There is also a general decrease in the loss over time, indicating that, overall, the model is learning and improving, but the large oscillations, especially early on, suggest that the high learning rate might have contributed to instability. While this rate helped the model adapt quickly, which we considered important due to real-world setting, it is also preventing smooth convergence. Our model probably struggled to stabilize due to placing too much weight on each update during interactions. Additionally, the general limited amount of training steps due to the real-world constraints might have hindered model convergence, with the model requiring more training time to fully converge. This aligns with Park et al. [37], reporting failed convergence after pursuing a similar number of 423 training episodes per child. The significant decrease in the children's motivation in the *ResUli-RL* condition could also be a sign that the DRL-powered agent's feedback was not well-aligned with maintaining student motivation, potentially explaining why the policy did not converge effectively. However, overall children managed to improve their reading abilities significantly in both conditions, indicating that the DRL-induced feedback provided by the PA could keep up with the high standard of expert-designed feedback. While this suggests that the policy successfully adapted to the students, it was not able to fully converge within the given real-world setting, which limits our findings.

RL in education has potential, but requires real-world evaluation. In summary, our results demonstrate that applying RL in real-world settings is equally important as it is challenging. While RL-based approaches demonstrated to outperform baselines in various contexts (and already did so several times within educational applications), simply applying RL in education might not be adequate or effective in every context, particularly when benchmarked against sophisticated baselines. Still, despite some technical and methodological limitations, our proposed approach *ResUli-RL* demonstrated to be feasible for running on an average tablet to be used in an applied school setting, demonstrating technical progress with steadily increasing cumulative rewards while decreasing the average loss. However, this did not translate into significantly better reading performance or motivation compared to the expert-designed baseline, with both groups achieving comparable reading gains. While this suggests that the DRL-powered PA did not offer added value in this particular context, it should be carefully considered that the model's performance not reaching convergence limit our results. Thus, while our findings highlight the importance of real-world evaluations, they also provide guidance for future work, stressing the need to evaluate new approaches in realistic real-world settings, comparing them to sophisticated baselines. Future research should focus on educational application contexts entailing a larger and diverse action space allowing a more fine-grained adaptation to the learner, as well as exploring hybrid approaches that combine expert-driven and data-driven methods to support both performance and motivation effectively.

6 CONCLUSION

In our work, we proposed a novel residual approach for RL in primary education, *ResUli-RL*, where the components of the MDP are inferred from educational psychology. We further address the gap in real-world evaluations of educational RL systems by comparing the DRL-powered pedagogical agent with an expert-designed baseline in an in-situ five-week intervention, assessing long-term effects on the children's reading competence and motivation. While our RL agent demonstrated technical improvements, as reflected by increasing cumulative rewards during the interaction and reduced average loss, the policy did not seem to fully converge and the post-test results revealed no significant differences in increased reading competence between the two conditions, but a significant improvement from pre- to post-test in both groups. Further, motivation decreased in both groups, with a significantly larger decline in the *ResUli-RL* group. Notably, our DRL model was not able to reach full convergence, limiting the resulting implications and highlighting the need to evaluate DRL policies with real learners to better understand their practical impact in such complex environments. However, successfully applying RL in education requires overcoming multiple hurdles (limited training data, vulnerable target group) while at the same time outperforming sophisticated baselines, deeming our method a feasible approach that can inform future work.

ACKNOWLEDGMENTS

Partly funded by the German Federal Ministry of Education and Research (BMBF) under the grant agreement MobiLe (03VP07080).

REFERENCES

[1] Raid Rafi Omar Al-Nima, Tingting Han, and Taolue Chen. 2020. Road Tracking Using Deep Reinforcement Learning for Self-driving Car Applications. In *Progress in Computer Recognition Systems*, Robert Burduk, Marek Kurzynski, and Michał Wozniak (Eds.). Springer, Cham. https://doi.org/10.1007/978-3-030-19738-4{_}12

[2] Ufuoma Chima Apoki, Aqeel M. Ali Hussein, Humam K. Majeed Al-Chalabi, Costin Badica, and Mihai L. Mocanu. 2022. The Role of Pedagogical Agents in Personalised Adaptive Learning: A Review. *Sustainability* 14, 11 (2022), 6442. https://doi.org/10.3390/su14116442

[3] Can Ateş, Özlem Kaymaz, H. Emre Kale, and Mustafa Agah Tekindal. 2019. Comparison of Test Statistics of Nonnormal and Unbalanced Samples for Multivariate Analysis of Variance in terms of Type-I Error Rates. *Computational and mathematical methods in medicine* 2019 (2019), 2173638. https://doi.org/10.1155/2019/2173638

[4] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C. Mitchell. 2020. Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (ACM Digital Library)*, Regina Bernhaupt (Ed.). Association for Computing Machinery, New York,NY,United States, 1–12. https://doi.org/10.1145/3313831.3376518

[5] Amy L. Baylor and Yanghee Kim. 2005. Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education* 15, 2 (2005), 95–115.

[6] Amy L. Baylor and Jeeheon Ryu. 2003. The API (Agent Persona Instrument) for Assessing Pedagogical Agent Persona. In *Proceedings of EdMedia + Innovate Learning 2003*, David Lassner and Carmel McNaught (Eds.). Association for the Advancement of Computing in Education (AACE), Honolulu, Hawaii, USA, 448–451.

[7] Andrew Thomas Bimba, Norisma Idris, Ahmed Al-Hunaiyyan, Rohana Binti Mahmud, and Nor Liyana Bt Mohd Shuib. 2017. Adaptive feedback in computer-based learning environments: a review. *Adaptive Behavior* 25, 5 (2017), 217–234. https://doi.org/10.1177/1059712317727590

[8] María J. Blanca, Jaume Arnau, F. Javier García-Castro, Rafael Alarcón, and Roser Bono. 2023. Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power. *Psicothema* 35, 1 (2023), 21–29. https://doi.org/10.7334/psicothema2022.292

[9] Justine Cassell. 2022. Socially Interactive Agents as Peers. In *The Handbook on Socially Interactive Agents*, Birgit Lugrin, Catherine Pelachaud, and David Traum (Eds.). ACM Press, New York.

[10] Juan C. Castro-Alonso, Rachel M. Wong, Olusola O. Adesope, and Fred Paas. 2021. Effectiveness of Multimedia Pedagogical Agents Predicted by Diverse Theories: a Meta-Analysis. *Educational Psychology Review* 33, 3 (2021), 989–1015. https://doi.org/10.1007/s10648-020-09587-1

[11] Chaquo Ltd. 2022. Chaquopy. https://chaquo.com/chaquopy/

[12] Huili Chen, Hae Won Park, and Cynthia Breazeal. 2020. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education* 150 (2020), 103836. https://doi.org/10.1016/j.compedu.2020.103836

[13] Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. 2010. Inducing Effective Pedagogical Strategies Using Learning Context Features. In *User Modeling, Adaptation, and Personalization*, Paul de Bra, Paul Del Brassey, Alfred Kobsa, and David Chin (Eds.). Lecture Notes in Computer Science / Information Systems and Applications, incl. Internet/Web, and HCI, Vol. 6075. Springer, Berlin and Heidelberg, 147–158. https://doi.org/10.1007/978-3-642-13470-8{_}15

[14] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2. ed. ed.). Erlbaum, Hillsdale, NJ. https://doi.org/10.4324/9780203771587

[15] Laduona Dai, Merel M. Jung, Marie Postma, and Max M. Louwerse. 2022. A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects. *Computers & Education* 190 (2022), 104607. https://doi.org/10.1016/j.compedu.2022.104607

[16] Floris den Hengst, Eoin Martino Grua, Ali el Hassouni, and Mark Hoogendoorn. 2020. Reinforcement learning for personalization: A systematic literature review. *Data Science* 3, 2 (2020), 107–147. https://doi.org/10.3233/DS-200028

[17] Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. 2021. MushroomRL: simplifying reinforcement learning research. *Journal of Machine Learning Research* 22, 1 (2021).

[18] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. 2019. Where's the Reward? *Int. J. Artif. Intell. Ed.* 29, 4 (2019), 568–620. https://doi.org/10.1007/s40593-019-00187-x

[19] Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. 2021. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning* 110, 9 (2021), 2419–2468. https://doi.org/10.1007/s10994-021-05961-4

[20] Bisni Fahad Mon, Asma Wasfi, Mohammad Hayajneh, Ahmad Slim, and Najah Abu Ali. 2023. Reinforcement Learning in Education: A Literature Review. *Informatics* 10, 3 (2023), 74. https://doi.org/10.3390/informatics10030074

[21] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191. https://doi.org/10.3758/bf03193146

[22] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. 2018. An Introduction to Deep Reinforcement Learning. *Foundations and Trends® in Machine Learning* 11, 3-4 (2018), 219–354. https://doi.org/10.1561/2200000071

[23] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. https://doi.org/10.3102/003465430298487

[24] Janina Heß, Panagiotis Karageorgos, Bettina Müller, Anna Riedmann, Philipp Schaper, Birgit Lugrin, and Tobias Richter. 2024. Improving word reading skills of low–skilled readers: An intervention combining a syllable–based approach with digital game–based features. *Journal of Computer Assisted Learning* 40, 5 (2024), 2306–2324. https://doi.org/10.1111/jcal.13021

[25] Janina Heß, Anna Riedmann, Panagiotis Karageorgos, Philipp Schaper, Birgit Lugrin, Tobias Richter, and Bettina Müller. 2024. MobiLe: Konzeption einer digitalen silbenbasierten Leseförderung für die Grundschule. *Psychologie in Erziehung und Unterricht* 71, 1 (2024), 41–51. https://doi.org/10.2378/peu2024.art05d

[26] JASP Team. 2021. JASP. https://jasp-stats.org/

[27] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. 2019. Residual Reinforcement Learning for Robot Control. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, Piscataway, NJ, 6023–6029. https://doi.org/10.1109/ICRA.2019.8794127

[28] Diederik P. Kingma and Jimmy Ba. [n.d.]. Adam: A Method for Stochastic Optimization. http://arxiv.org/pdf/1412.6980

[29] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

[30] H. Chad Lane and Noah L. Schroeder. 2022. Pedagogical Agents. In *The handbook on socially interactive agents*, Birgit Lugrin, Catherine Pelachaud, and David R. Traum (Eds.). Association for Computing Machinery, New York, NY, 307–330. https://doi.org/10.1145/3563659.3563669

[31] Yu Tang Liu, Eric Price, Michael J. Black, and Aamir Ahmad. 2022. Deep Residual Reinforcement Learning based Autonomous Blimp Control. In *IROS 2022 Kyōto - IEEE/RSJ International Conference on Intelligent Robots and Systems*, Zhidong Wang, Noriaki Ando, and Natsuki Yamanobe (Eds.). IEEE, Piscataway, NJ, 12566–12573. https://doi.org/10.1109/IROS47612.2022.9981182

[32] Nicole Loorbach, Oscar Peters, Joyce Karreman, and Michaël Steehouder. 2015. Validation of the Instructional Materials Motivation Survey (IMMS) in a self-directed instructional setting aimed at working with technology. *British Journal of Educational Technology* 46, 1 (2015), 204–218. https://doi.org/10.1111/bjet.12138

[33] Kenneth O. McGraw and S. P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1, 1 (1996), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

[34] Wilfried Metze. 2009. STOLLE (STOLperwörter-LEsetest). https://www.uni-potsdam.de/de/gsp-deutsch/forschung/stolle

[35] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533. https://doi.org/10.1038/nature14236

[36] Bettina Müller, Tobias Richter, and Panagiotis Karageorgos. 2020. Syllable-based reading improvement: Effects on word reading and reading comprehension in Grade 2. *Learning and Instruction* 66 (2020), 101304. https://doi.org/10.1016/j.learninstruc.2020.101304

[37] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019), 687–694. https://doi.org/10.1609/aaai.v33i01.3301687

[38] Rafael Pina, Varuna de Silva, Joosep Hook, and Ahmet Kondoz. 2024. Residual Q-Networks for Value Function Factorizing in Multiagent Reinforcement Learning. *IEEE Transactions on Neural Networks and Learning Systems* 35, 2 (2024), 1534–1544. https://doi.org/10.1109/TNNLS.2022.3183865

[39] Martin L. Puterman. 2005. *Markov decision processes: Discrete stochastic dynamic programming*. Wiley-Interscience, Hoboken, New Jersey. https://doi.org/10.1002/9780470316887

[40] Anna Riedmann, Julia Götz, Carlo D'Eramo, and Birgit Lugrin. 2024. Uli-RL: A Real-World Deep Reinforcement Learning Pedagogical Agent for Children. In *KI 2024: Advances in Artificial Intelligence*, Andreas Hotho and Sebastian Rudolph (Eds.). Lecture Notes in Artificial Intelligence, Vol. 1410. Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-031-70893-0{_}25

[41] Anna Riedmann and Birgit Lugrin. 2023. Towards an Adaptive Pedagogical Agent in a Reading Intervention Using Reinforcement Learning. In *Proceedings of the*

23rd ACM International Conference on Intelligent Virtual Agents (ACM Digital Library), Birgit Lugrin, Marc Latoschik, Sebastion von Mammen, Stefan Kopp, Florian Pécune, and Catherine Pelachaud (Eds.). Association for Computing Machinery, Erscheinungsort nicht ermittelbar, 1–3. https://doi.org/10.1145/3570945.3607320

[42] Anna Riedmann, Philipp Schaper, Melissa Donnermann, Martina Lein, Sophia C. Steinhaeusser, Panagiotis Karageorgos, Bettina Müller, Tobias Richter, and Birgit Lugrin. 2022. Iteratively Digitizing an Analogue Syllable-Based Reading Intervention. *Interacting with Computers* 33, 4 (2022), 411–425.

[43] Anna Riedmann, Philipp Schaper, Benedikt Jakob, and Birgit Lugrin. 2022. A Theory Based Adaptive Pedagogical Agent in a Reading App for Primary Students - A User Study. In *Intelligent Tutoring Systems*, Scott Crossley and Elvira Popescu (Eds.). Springer eBook Collection, Vol. 13284. Springer International Publishing and Imprint Springer, Cham, 276–292. https://doi.org/10.1007/978-3-031-09680-8{_}26

[44] Anna Riedmann, Philipp Schaper, and Birgit Lugrin. 2024. Reinforcement Learning in Education: A Systematic Literature Review. *Submitted (preprint available on request)* (2024).

[45] Noah L. Schroeder and Olusola O. Adesope. 2014. A Systematic Review of Pedagogical Agents' Persona, Motivation, and Cognitive Load Implications for Learners. *Journal of Research on Technology in Education* 46, 3 (2014), 229–251. https://doi.org/10.1080/15391523.2014.888265

[46] Noah L. Schroeder, Olusola O. Adesope, and Rachel Barouch Gilbert. 2013. How Effective are Pedagogical Agents for Learning? A Meta-Analytic Review. *Journal of Educational Computing Research* 49, 1 (2013), 1–39. https://doi.org/10.2190/EC.49.1.a

[47] Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. 2023. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications* 231 (2023), 120495. https://doi.org/10.1016/j.eswa.2023.120495

[48] P. E. Shrout and J. L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

[49] Richard S. Sutton and Andrew Barto. 2018. *Reinforcement learning: An introduction* (second edition ed.). The MIT Press, Cambridge, MA and London.

[50] Samuele Tosatto, Matteo Pirotta, Carlo D'Eramo, and Marcello Restelli. 2017. Boosted Fitted Q-Iteration. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3434–3443. https://proceedings.mlr.press/v70/tosatto17a.html

[51] Hao-nan Wang, Ning Liu, Yi-yun Zhang, Da-wei Feng, Feng Huang, Dong-sheng Li, and Yi-ming Zhang. 2020. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 21, 12 (2020), 1726–1744. https://doi.org/10.1631/FITEE.1900533

[52] Shuhuan Wen, Yili Shu, Ahmad Rad, Zeteng Wen, Zhengzheng Guo, and Simeng Gong. 2025. A deep residual reinforcement learning algorithm based on Soft Actor-Critic for autonomous navigation. *Expert Systems with Applications* 259 (2025), 125238. https://doi.org/10.1016/j.eswa.2024.125238

[53] Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2019. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in psychology* 10 (2019), 3087. https://doi.org/10.3389/fpsyg.2019.03087

[54] Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*. 820–828. http://arxiv.org/pdf/1611.00179

[55] Guojing Zhou, Hamoon Azizsoltani, Markel Sanz Ausin, Tiffany Barnes, and Min Chi. 2019. Hierarchical Reinforcement Learning for Pedagogical Policy Induction. In *Artificial intelligence in education*, Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (Eds.). LNCS sublibrary, Vol. 11625. Springer International Publishing, Cham, 544–556. https://doi.org/10.1007/978-3-030-23204-7{_}45