ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Reinforcement Learning-Based Algorithm Design for Enhancing Self-Efficacy

Pei Zhang[1,*]

1 Developmental and Educational Psychology, Xinjiang Teacher's College, Urumqi, Xinjiang 830043, China

SUMMARY: *Adaptive learning systems frequently carry out optimization on correctness and completion, meanwhile they neglect the changes of learners' self-efficacy. This research establishes a reinforcement learning frame which together models learning condition, mission hardness, feedback category, and support strength, and makes them align through a multi-goal reward that balances self-efficacy, achievement, and perseverance. In one 8-week Python small course which has 228 undergraduate students and 21,864 interactive records, the RL group displayed stronger results, which include a self-efficacy increase of 0.49, post-test correct rate of 87.6%, a finish rate of 94.2%, and a dropout rate that is 11.3% lower; The subgroup which has low level before SE obtained an effect size with the value of d=0.74. Further deeper analysis brings forward the idea that continuous challenge, motive-related feedback, and rule-based control can assist in the enhancement of successful experiences, self-confidence, and the stability of learning. These results give support to a route which has more interpretability and extensibility for intelligent education intervention.*

KEYWORDS: *Reinforcement learning; Self-efficacy; Adaptive learning system; Multi-objective reward; Personalized feedback*

1 Introduction

Digital study surroundings have made adaptive systems go past simple resource distribution and move toward teaching decision support. In this transformation, the reinforcement learning has become particularly important because educational interactions develop through time, rely on feedback from previous steps, and seldom bring about immediate results. Because of this reason, therefore, it is more and more utilized in question recommendation, tutoring work, learning path adjustment, and interaction support work. Recently past summaries indicate a wider change inside the domain: reinforcement learning is not any more researched only as a technical concept proof, but as a method to model true learning processes in real situations [1].

Even so, the majority of adaptive systems are still evaluated mainly by observable indices like accuracy, completion degree, click numbers, residence time, or path-matching working effect. These methods can obtain short-term results, but therefore they reflect far less information about whether learners still keep willingness to go on, how they make reaction to hardship, or whether they build up self-confidence through many repeated tasks. The authors Li and others has proven that deep reinforcement learning is able to express adaptive learning as a sequential decision question which has states, actions, and delayed rewards[2]. This viewpoint gives space for regarding self-efficacy as a component of the decision-making

*zhangpei850315@163.com
https://doi.org/10.65102/is2026733
Published: 30 April, 2026

process itself, instead of putting it as a variable which is only studied after teaching activities.

From the perspective of the learning process itself, a learner's performance is not determined by the outcome of a single test at a given point in time, but is jointly influenced by a series of continuous behaviors, including strategy selection, task adaptation, and feedback assimilation. Osakwe et al. applied reinforcement learning to identify effective strategies in self-regulated learning, demonstrating that learning behavior trajectories possess a learnable structure, and that the system can identify action patterns more conducive to learning progress through continuous interaction [3]. This conclusion provides direct inspiration for this paper: if educational algorithms track only immediate scores, they struggle to capture how learners form judgments about their own abilities through challenges, setbacks, adjustments, and retries; only by viewing the learning process as a dynamic system requiring long-term coordination can we discuss the mechanisms underlying the formation of self-efficacy and the pathways through which algorithms play a role in this process.

Among the elements of educational interventions, the form and timing of feedback have a significant impact on the learning experience. Sailer et al. found in simulated learning scenarios that adaptive feedback generated by artificial neural networks can enhance learners' diagnostic reasoning performance [4]. This result indicates that the role of feedback is not limited to indicating correctness or error; it also influences learners' perception of task difficulty, their attribution of errors, and their expectations for subsequent learning.When the feedback has good matching with the level of the task, learners have more probability to build a stable feeling of success through the continuous interaction; On the opposite side, when feedback does not match, prompting words have too much interference, or continuous failures are not handled properly, therefore the learning process has higher possibility to get broken, people quit, and have low level of involvement. Therefore, the handling of feedback ways and speed by adaptive systems should not only be utilized for promoting performance, but also be utilized for constructing study confidence.

In the research of today, an important gap still exists. As Memarian and Doleck point out, reinforcement learning within education still is directed mainly at path suggestion, resource distribution, and efficiency promotion, while long-time handling of psychological factors, explanation of intervention working principles, and proof of educational worth are much less developed[5]. In this circumstance, self-efficacy is often regarded as an outcome that is measured after teaching instead of as a goal that is embedded in policy study. As the consequence, one system may promote short-term effect without continuously consolidating learners' belief in their own ability, and thus it becomes more difficult to explain why one method is effective for certain students, has reduced influence as time passes, or generates non-uniform results among different groups.

The differences that exist among learners therefore make this problem have more obvious manifestation. The authors Ruan and others In mathematics tasks we have found that reinforcement learning tutoring has special helpfulness for learners who get lower scores[6], therefore this suggests that the algorithmic benefits are not shared in an even way. The students who have weaker basic knowledge, more lasting setbacks, or worse initial performance therefore often more rely on the system’s fine detailed support. Because of this point, the research about educational algorithms must go further than the average group comparisons, and therefore put more close attention on which people get benefits, how the change develops step by step, and how policies work under the different conditions of learners. If we do not have this step, the worth of reinforcement learning inside education thus still is only partially comprehended by people.

The reinforcement learning which is applied in education moreover faces more rigorous deployment requirements when it is compared with common try-and-error environments.

Decisions concerning task difficulty, feedback strength, and route alteration must stay instructionally rational instead of being pushed only by instant reward[7]. Yun and other scholars have provided proof that limiting methods can enhance the stable character and application range of learning-path strategy sets. This viewpoint is especially connected here: if self-efficacy is regarded as a goal together with study outcomes, the policy must hence avoid moving toward overly easy tasks, overly numerous prompts, or narrowed challenge ranges just because these bring short-term increases.

The correlative research about motivation and support has supplied the extra foundation for this research work. Zhang discovered that deep reinforcement learning is able to be utilized for the optimization of incentive mechanisms for online English learners[8], therefore suggesting that its function in education expands further than content sequencing into the construction of support structures. Even so, more powerful stimulus measures do not certainly bring about higher self-efficacy. Engagement-focused design mainly assists in keeping interest and participation, while self-efficacy has a closer connection with learners' evaluations about whether they are able to finish tasks, deal with hardship, and achieve progress. For this cause, the holding of engagement and the building of confidence ought to be regarded as related but separate goals in both reward design and effectiveness assessment.

From the angle of participation, self-efficacy bears stronger explanation weight than many surface type indicators. Getenet and other colleagues have made a report that students' digital attitudes, digital literacy and self-efficacy all have connections with online engagement, hence self-efficacy displays a particularly important connection with engagement behavior[9]. This indicates that self-efficacy assists in linking the technical environment to real learning behaviors, instead of acting as an edge psychological measurement. A system which can stably enhance learners' confidence on judging and handling their own performance is hence more possible to support continuance, interaction quality, and longer-term accomplishment.

Shin also discovered that curiosity has a contribution to self-regulated learning and academic achievement in online classes, but that this effect changes along with the levels of learners' self-efficacy[10]. Put together, these results show that self-efficacy influences not only final results but also the way learners give reaction to difficulty, feedback and exploration. On this foundation, the current research puts self-efficacy promotion in the core position of reinforcement learning design, and puts forward three questions: whether adaptive strategies can promote self-efficacy, whether such promotion goes together with increases in performance and persistence, and which groups of learners can get the most benefit. For answering these questions, we have constructed a reinforcement learning frame which takes self-efficacy as the core objective, we use a multi-objective reward to make a balance among performance, engagement and psychological change, and we conduct an examination of the policy by means of controlled experiments, ablation analysis and heterogeneity analysis.

2 Methods

2.1 Research Context, Participants, and Multi-source Data

This study defines personalized learning support as a continuous decision-making problem: after each learning interaction, the system reads the learner's current performance and behavioral characteristics to determine the next task difficulty, feedback method, and support intensity, then updates the strategy based on learning outcomes, sustained engagement, and changes in self-efficacy[11]. The research focuses not only on single-round response results but also on changes in learners' sense of competence, persistence, and performance evolution across consecutive tasks.

For the combination of stable policy study and actual teaching verification, this research uses a two-step data arrangement. The offline stage employs the Open University Learning Analytics Dataset (OULAD) to conduct feature screening, early behavior pattern recognition, and baseline comparison. Because this dataset possesses student background information, course sign-up records, assessment documents, and virtual learning environment behavior logs, therefore it supports the construction of variables which are related to previous knowledge foundation, learning participation, and dropout possibility. The network phase includes a 8-week university Python small class experiment which targets the first and second year undergraduate students. This curriculum contains six modules-variable and data type, conditional control, circulation, function, list and dictionary, and mistake correction-and arranges three compulsory assignments plus one optional strengthening practice every week, thus getting approximately 24–32 effective interactions for each learner.

A total of 236 students were initially recruited for the experiment. After excluding samples with missing pre-tests, missing post-tests, or those who dropped out midway, 228 valid participants were retained, comprising 114 in the experimental group and 114 in the control group.The platform recorded a total of 21,864 valid logs, with log granularity covering task start and submission times, hint requests, retry counts, answer accuracy, duration per round, and interruption status. Variables were categorized into four types: the first category comprised background variables, including grade level, major, prior programming performance, pre-course test scores, and prior online learning experience;the second category comprises psychological variables, including pre- and post-course self-efficacy measurements and their changes; the third category comprises behavioral variables, including response time, hint use, retry count, number of consecutive failures, and dropout; the fourth category comprises outcome variables, including single-round accuracy, stage quizzes, the course post-test, and the retention test two weeks later. The study participants and variable structure are shown in Table 1.

Table 1: Study Participants and Variable Collection Framework

Dimension	Indicator	Source	Purpose
Background Variables	Grade level, major category, prior programming performance, pre-course assessment	Collected once before the course begins	Control for differences in prior learning
Psychological Variables	Self-efficacy (pre-test, post-test, change)	Week 1/Week 8 questionnaire	Measuring changes in ability judgments
Behavioral Variables	Response time, hint use, retry count, consecutive failures, dropout	Platform automatically records	Constructing State and Assessing Persistence
Outcome variables	accuracy, stage test, post-test, retention test	Process logs and stage tests	Testing learning outcomes and retention

According to the above design, this paper puts forward four hypotheses which need to be tested[12]. H1: The personalized tactics which are based on reinforcement learning can make the promotion of self-efficacy obtain obvious enhancement. H2: For the post-test accuracy and task completion, the experiment group has better performance than the baseline group. H3: For learners who have low starting self-efficacy, they display bigger promotion. H4: The promotion of self-efficacy partially plays the mediating role in the connection between algorithm intervention and academic achievement. The decision circulation in Figure 1 and the change

structure in Table 1 together construct the base for following model study, controlled experiments, and intermediary analysis.

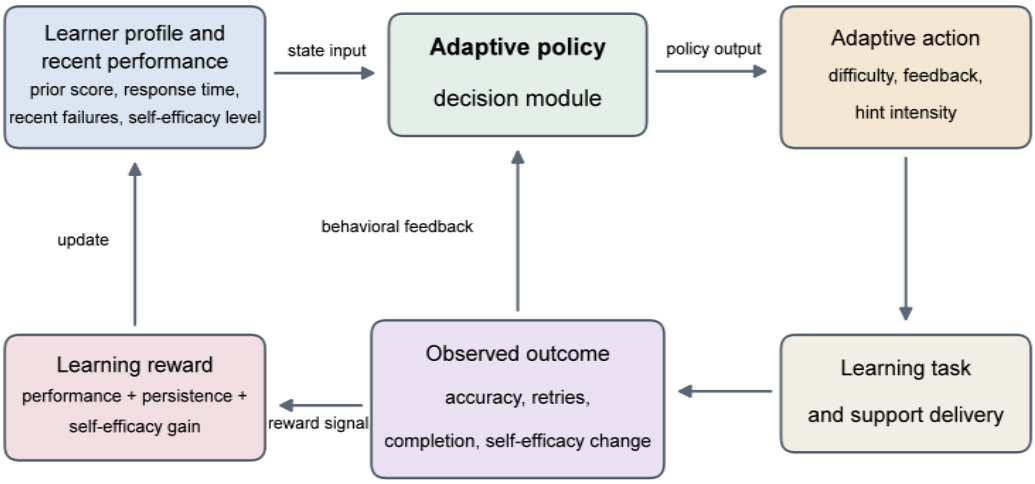


Figure 1: RL-based adaptive learning interaction framework.

2.2 RL Algorithm Design and Reward Mechanism

This research utilizes a restricted DQN to choose discrete teaching actions, hence the major methodological contribution hence lies in how learner information, teaching choices, reward signals, and control constraints are together arranged. As it is displayed in Figure 2, the policy input is constructed from three origins: near-term learning records, behavior tracks, and feeling signals. These include past-cycle correctness[13], behavior on the most recent three jobs, continuous unsuccessful attempts, answer speed, clue asks, try frequency, a self-ability substitute, mental burden, and a short angry mood test. Through the combination of these signals, the model can capture both immediate changes and continuous learning tendencies, which therefore helps to support more stable action distribution among continuous interactive processes.

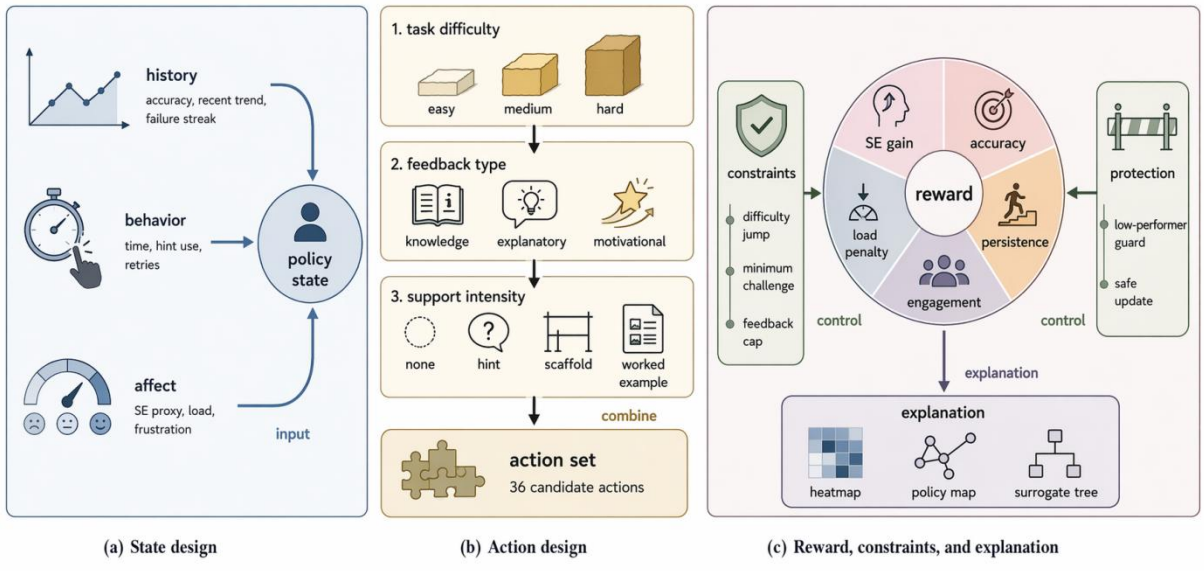


Figure 2: Method schematic of state design, action design, and reward-control interface in the proposed reinforcement learning framework.

The action layer is organized as a three-dimensional combination. In each round, the system jointly selects task difficulty, feedback type, and support intensity. Difficulty has three levels-simple, middle, and difficult; Feedback contains three types, which are knowledge type, explanatory type and motivational type; The scope of support includes no support, faint clue, supporting frame and finished example[14]. When put together, these different aspects constitute 36 candidate operation items. As is displayed in Figure 2, the action space is designed for regulating challenge, explanation, and assistance at one time, thus the policy can give response to learner demands more completely than a system which only gives recommendation of content.

The reward function is expressed as:

$$R_t = 0.35\Delta SE_t + 0.30Acc_t + 0.20Persist_t + 0.10Engage_t - 0.05Load_t \quad (1)$$

where ΔSE_t represents the incremental self-efficacy gain, Acc_t indicates the correctness of the current task, $Persist_t$ reflects the probability of continued learning or completion, $Engage_t$ characterizes the level of engagement, and $Load_t$ penalizes overload and consecutive failures. This formulation integrates psychological improvement, performance enhancement, and sustained engagement into a single optimization objective, helping to curb strategy drift that solely pursues immediate accuracy rates[15-19].

In the training process, the policy is constrained by four control regulations: restricting sudden difficulty changes, keeping a lowest degree of difficulty, stopping too frequent feedback, and starting a protection mechanism when continuous failures go beyond a threshold. At the same time, the model maintains an interpretability handling interface which contains action-frequency heatmaps, policy visualization, feature importance degree, and surrogate trees, hence permitting the analysis of action patterns, policy movement, and the influence of key input elements. In Figure 2, the module on right hand side makes summary of these control components and interpretation components, which will get further examination in the experiments that are later.

2.3 Baselines, Experimental Design, and Statistical Analysis

To investigate the sources of performance gains from the proposed method, the experiment was designed with four control groups, as shown in Table 2.The Rule-based group assigns tasks and provides feedback according to predefined instructional rules, such as "reduce difficulty and provide a hint after two consecutive errors"; the Supervised recommendation group trains a classifier based on historical logs to predict the action most likely to yield a correct answer in the next round; the Bandit group employs a contextual bandit to update action preferences under local rewards; and the Proposed RL group uses the constraint-based reinforcement learning strategy proposed in this paper, simultaneously optimizing self-efficacy, performance, and sustained engagement.In Table 2, all groups share the same course content, interactive interface, and response time settings; the only difference lies in the action generation mechanism. This comparative approach facilitates the distinction of performance differences among fixed-rule[20], supervised prediction, lightweight adaptive, and sequential decision-making strategies. Research on interpretable reinforcement learning emphasizes that strategy comparisons should encompass both decision quality and interpretable outputs; therefore, in addition to main effect tests, this paper also retains action heatmaps, strategy transfer maps, and agent tree analysis interfaces.

Table 2: Baseline Group Settings and Statistical Comparison Objectives

Group	Decision Mechanism	Basis for Action Generation	Primary Comparison Objectives	Explanatory Output
Rule-based	Fixed Rules	Teacher-defined thresholds and trigger conditions	Serves as a minimum baseline for instruction	Rule Log
Supervised recommendation	Supervised Learning	Next-Action Prediction Based on Historical Logs	Evaluate the relative gain of "single-step prediction"	Feature Importance
Bandit/Shallow RL	Lightweight Adaptive	Contextual Reward Updates	Testing whether local adaptation is sufficient	Action Frequency Heatmap
Proposed RL	Constrained RL	Multi-objective Long-term Reward	Verifying the overall gain of sequential policies	Heatmaps, Strategy Maps, and Agent Trees

This experiment has adopted a design of parallel groups. Altogether 228 effective participants were layered according to pre-test achievement and beginning self-efficacy, hence after were randomly distributed into four groups with 57 in each. This interference has continued for total 8 weeks, and the system has recorded task difficulty degree, feedback type, support strength, reaction time, hint utilization, retry number, correct rate, and exit situation in each interactive process. Pre-test data were got together before the course, post-test data at the finish of Week 8, and retention data two weeks after that time. The main resulting items include self-efficacy promotion, post-test correctness, task finishing, quitting rate, and reservation score. Because adaptive feedback can bring about different influences on different results, therefore this research carries out evaluation on intervention effects by using multiple indicators, instead of depending on one single performance measurement.

Statistical analysis proceeded in three layers. The first layer examined mean differences between groups. Pre-test equivalence was checked with one-way ANOVA or independent-samples t-tests, while post-test outcomes were mainly analyzed with ANCOVA, using pre-test scores and prior levels as covariates[21]. The model is written as:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 Pre_i + \beta_3 Prior_i + \varepsilon_i \tag{2}$$

Here, Y_i represents the post-test results for the ith student, corresponding to the self-efficacy post-test, post-test accuracy, or retention score; G_i is the group variable; Pre_i is the corresponding pre-test score; $Prior_i$ represents the prior learning foundation; β_0 is the intercept term; β_1, β_2, and β_3 are the coefficients to be estimated; and ε_i is the random error. For significant results, both Cohen's d and partial eta squared (η_p^2) are reported to reflect the magnitude of the effect.

A second-order mixed-effects model is established for repeated interaction data to analyze whether the effect of the strategy accumulates over time. The model is written as:

$$Y_{it} = \gamma_0 + \gamma_1 Treat_i + \gamma_2 Time_t + \gamma_3 (Treat_i \times Time_t) + u_i + \epsilon_{it} \tag{3}$$

where Y_{it} represents the outcome for the ith student in the tth interaction round, which may be a single-round measure of accuracy, hint use, or a self-efficacy proxy; $Treat_i$ indicates whether the target strategy was adopted; $Time_t$ represents the interaction round; $Treat_i \times Time_t$ is used to test changes in strategy effects over time; u_i is the individual random intercept, used to account for unobserved differences at the student level; ϵ_{it} is the residual term. If the interaction term is significant, it indicates that the advantage of the strategy is not concentrated at a single point in time but unfolds gradually during continuous learning[22-25].

The third level includes the testing of mechanism. This research uses middle variable analysis to test whether the promotion of self-efficacy partially carries out the mediation function in the connection between algorithmic intervention and academic achievement, and thus carries out grouping analyses according to starting self-efficacy degrees, pre-examination scores, and earlier programming experience. Effect magnitudes for the subgroup that has low starting self-efficacy shall be reported separately for the addressing of H3. Results are given by a unified standard form, for example, "self-efficacy promotion: +0.42SD, p<0.01," "post-test accuracy: +6.8%," "dropout rate: −11.3%," and "low-SE subgroup effect size: d=0.61." This report method enables the concurrent displaying of statistical meaningfulness and teaching meanings, and accords with the multi-dimensional result showing which is stressed in latest studies on explainable feedback and adjustable support.

3 Results and Discussion

3.1 Main Effects on Self-Efficacy and Learning Performance

Just like what Figure 3 shows, the whole effect is mainly manifested on three index items: self-efficiency, study achievement, and task finishing. The promotion of self-efficacy in the Proposed RL group was 0.49±0.23, it is higher than what the Rule-based group got (0.23±0.22), the Supervised group got (0.26±0.22), and the Bandit group got (0.41±0.21). One analysis of covariance which takes post-test self-efficacy as the dependent variable, and takes pre-test self-efficacy and previous achievement as covariance items, therefore has discovered a significant main effect of the group: F(3,222)=11.46, p <0.001, ηp^2 =0.13. In terms of learning performance, the RL group's post-test accuracy reached 87.6%, which was 10.3, 7.0, and 5.0 percentage points higher than that of the Rule-based, Supervised, and Bandit groups, respectively; the task completion rate was 94.2%, which was 10.8, 8.0, and 5.0 percentage points higher, respectively. The group effects for these two metrics were F(3,222)=19.15, p<0.001, ηp^2=0.21, and F(3,222)=24.46, p<0.001, ηp^2=0.25. Compared to all control groups, the RL group's self-efficacy gain increased by approximately 0.81SD. These results indicate that self-efficacy is not driven solely by attitude or interest; rather, it is gradually shaped through the continuous process of "use-feedback-experience of success." Recent research on AI self-efficacy also suggests that usage experience, interest, and relevant abilities collectively influence self-efficacy levels.

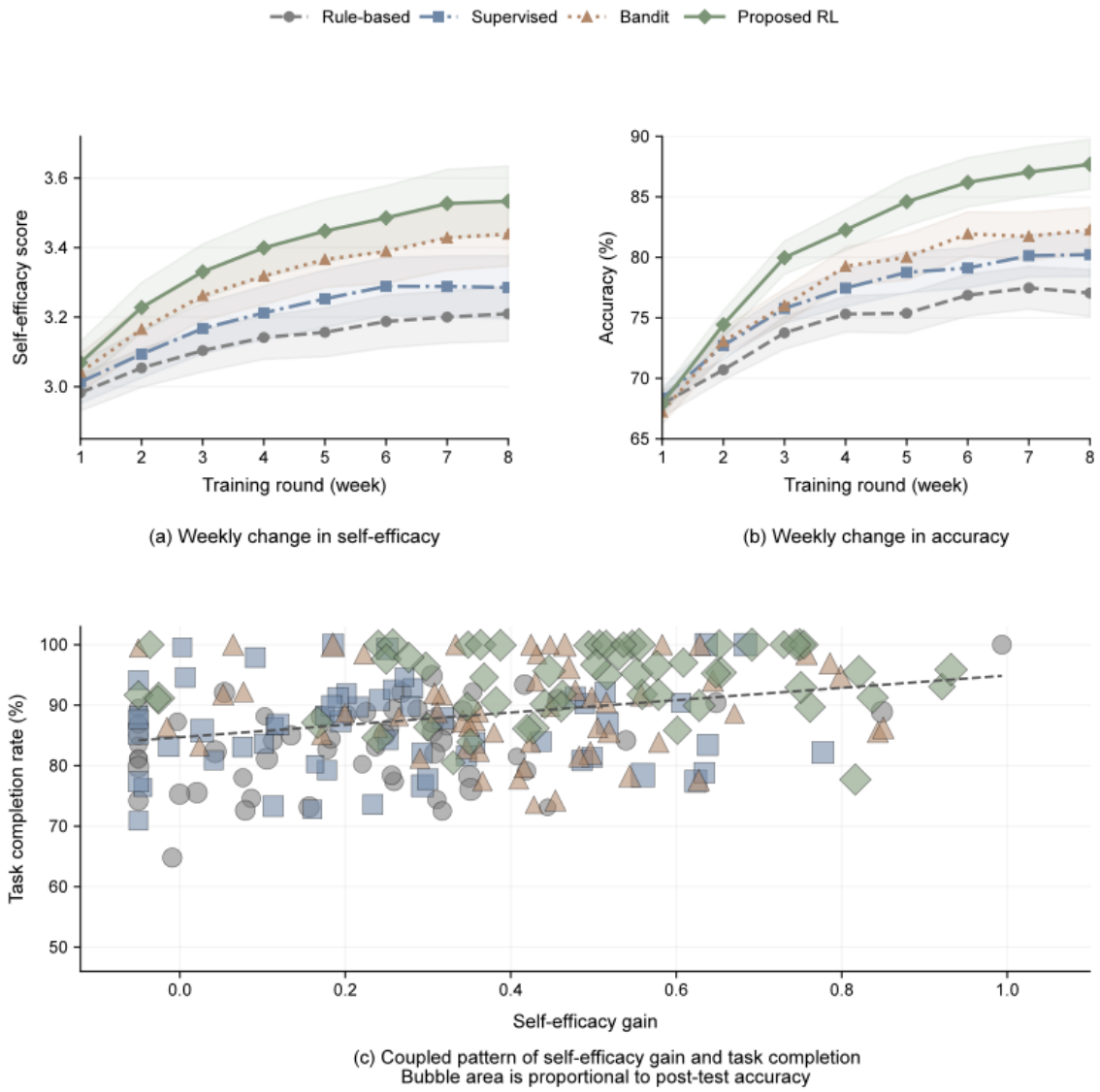


Figure 3: Main effects of reinforcement learning on self-efficacy and learning performance.

Figures 3(a) and 3(b) further illustrate the two time-series curves. The differences among the four groups at the starting point in Week 1 were small, with self-efficacy scores all around 3.0 and accuracy ranging from approximately 67% to 69%; Starting from Week 3, the gap between the RL group and the other three groups gradually widened. By Week 8, the RL group's self-efficacy score had risen to 3.54, while those of the Rule-based, Supervised, and Bandit groups were 3.21, 3.29, and 3.44, respectively; their accuracy at the same time point was 87.6%, 77.2%, 80.6%, and 82.5%, respectively. The mixed-effects model revealed that the strategy-time interaction term was significant in the self-efficacy trajectory (β=0.021, p<0.001) and in the accuracy trajectory (β=0.010, p<0.001). As shown in Figure 3(b), the RL group's advantage was not concentrated in the final test but entered a relatively stable accumulation phase after Week 4, which is consistent with findings that supportive conditions and prior experience continue to influence the formation of self-efficacy.

Figure 3(c) has illustrated the relation among self-efficacy increase, task finishing, and post-test correctness. The data points as a whole have a tendency going up to the right, this indicates that students who get bigger increases in self-efficacy often keep higher rates of task completion

and obtain better results in the post-test. Related analysis shows that the related coefficient between self-efficacy increase and completion rate is r=0.31, p<0.001, and the correlation coefficient between self-efficacy gain and post-test accuracy is r=0.44, p<0.001. After controlling for pre-test self-efficacy, prior academic performance, and group assignment, self-efficacy gain still has an independent predictive effect on persistence,β=0.225, p<0.001. From the perspective of distribution, the samples of the RL group were more concentrated in the upper-right area of Figure 3(c), which shows that the advantages of this strategy went beyond the psychological enhancements on the questionnaire level, and were in accordance with continuous participation and learning results. Related mixed-methods studies have also demonstrated a consistent association between technical self-efficacy and the use of learning strategies, learning satisfaction, and motivational resources, providing empirical support for the findings of this study.

3.2 Policy Behavior, Ablation, and Heterogeneity Analysis

Figure 4 can let us see that the allocation of policy has systematic changes along with the state of learner. In Figure 4(a), students who have low pre-SE and often meet continuous failures are more frequently arranged to medium difficulty tasks with explanatory feedback or scaffolding, the selection probabilities are 0.26 and 0.22, therefore the probability of getting hard difficulty with knowledge feedback is only 0.09. To learners who have stronger earlier knowledge and quicker reaction speed, the distribution moves in the direction of higher challenge, and the probabilities of hard tasks that match with explanatory or knowledge feedback rise to 0.36 and 0.30. This mode tells us that the policy adjusts challenge and feedback depth according to learner situation, not that it uses over and over a fixed action template. Recent summary research also points out that the worth of reinforcement learning in educational fields lies in continuous state-perceiving support and individualized interference, not in one-time recommendation correctness.

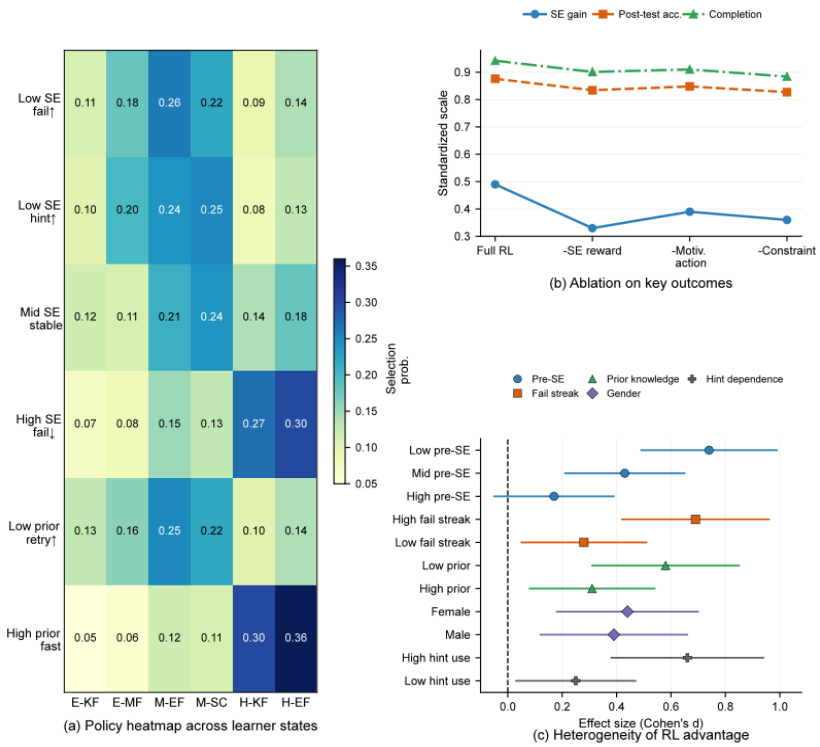


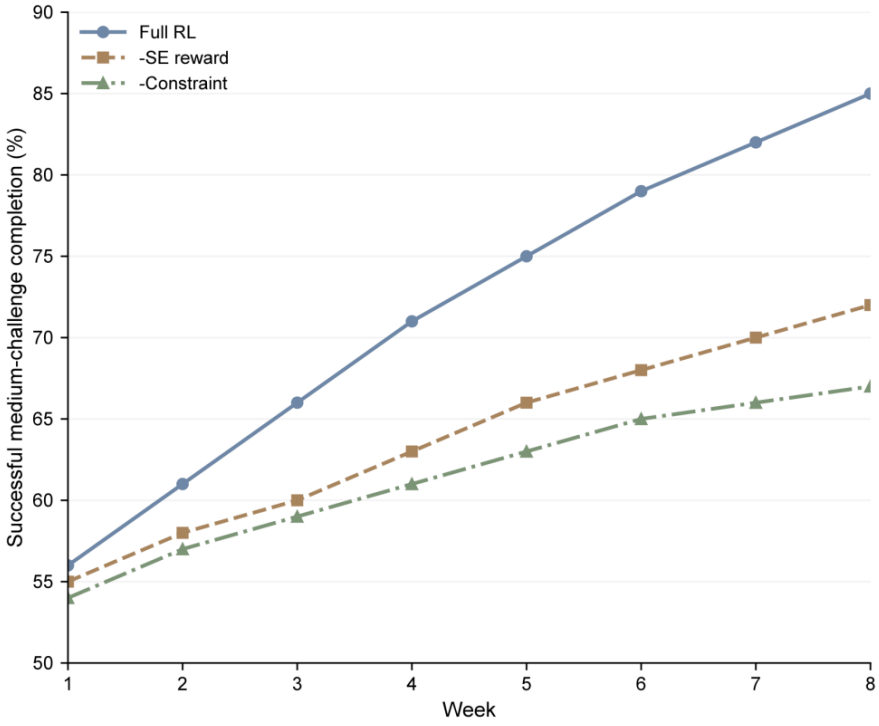
Figure 4: Policy behavior, ablation, and heterogeneity of the proposed RL method.

Figure 4(b) reports the ablation results for the main modules. In the full model, self-efficacy gain reaches 0.49; this value drops to 0.33 after removing the self-efficacy reward, to 0.39 after removing motivational feedback, and to 0.36 after removing the constraint. Post-test accuracy declines from 87.6% to 83.4%, 84.8%, and 82.7%, while completion rate falls from 94.2% to 90.1%, 91.0%, and 88.4%. The largest deterioration appears when the constraint is removed. Strategy logs further show that the share of low-self-efficacy learners assigned to high-difficulty actions increases from 18.4% to 31.7%. Together, these results suggest that reward shaping, motivational support, and control boundaries work jointly to keep the policy stable and instructionally appropriate.

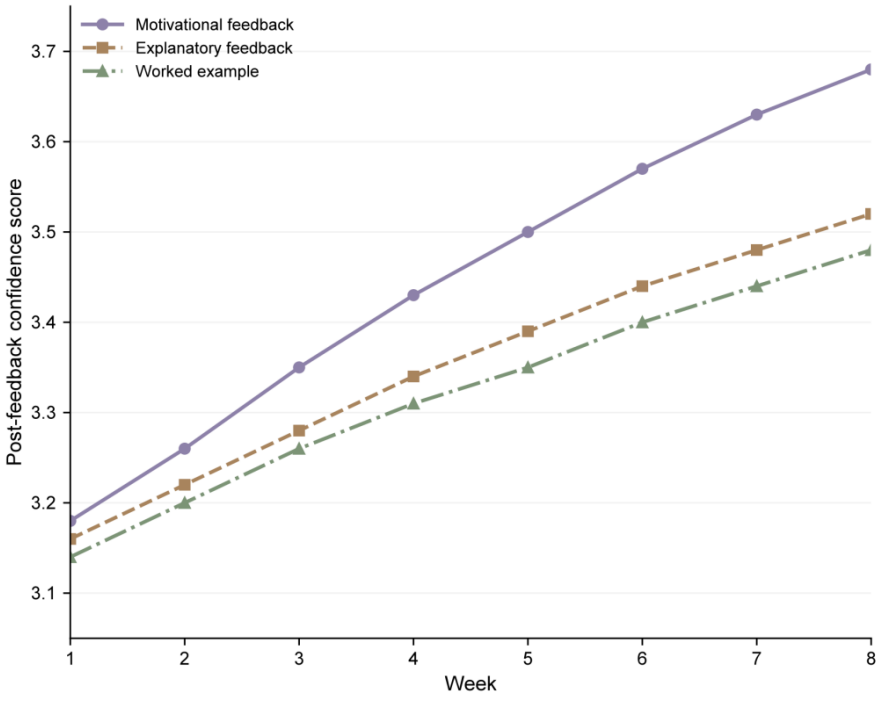
Figure 4(c) illustrates the heterogeneity effects. The advantage of the proposed RL over the combined control group reached an effect size of 0.74 (d) among students with low pre-SE, 0.43 among those with medium pre-SE, and dropped to 0.17 among those with high pre-SE.When grouped by learning process, the effect size for students with a high failure streak was 0.69, significantly higher than the 0.28 for students with a low failure streak; when grouped by support needs, students with high hint usage reached 0.66, while those with low hint usage were at 0.25. Gains were also higher for students with lower prior knowledge: the effect size for low prior knowledge was 0.58, higher than the 0.31 for high prior knowledge;Gender differences were relatively limited, with effects of 0.44 for females and 0.39 for males. This point shows that there is a clear layered distribution of gains, wherein advantages are gathered among groups which are featured by lower beginning self-efficacy, more often continuous failures, and heavier dependence on feedback. Related studies in high-level education indicate that self-efficacy and outside support together affect the building of attitudes towards AI tools and using intentions, this accords with the direction of group differences that is shown in Figure 4(c).

3.3 Mechanism Discussion, Educational Implications, and Limitations

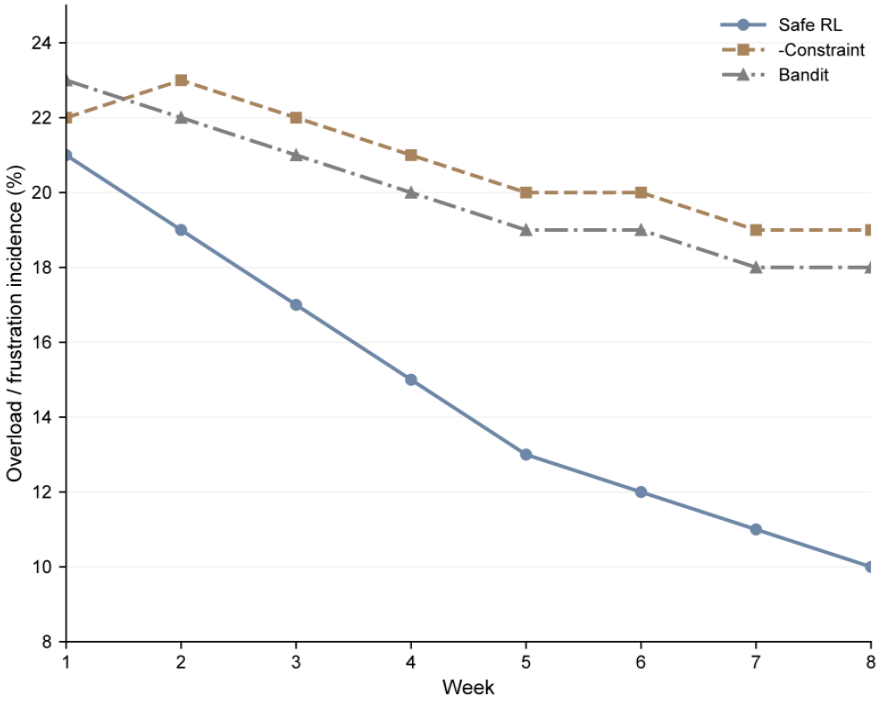
Figure 5 gives three lines of proof for the reason that the strategy we put forward can help sustained growth in self-efficacy. In Figure 5(a), the Full RL group has promoted its success probability on medium difficult tasks from 56% in the first week to 85% in the eighth week, while the edition that has no self-efficacy reward arrived at 72% and the edition that has no constraints arrived at 67%. This type of mode indicates that continuous challenge and reward shaping can help maintain task requirements in a controllable scope, therefore letting learners construct mastery through repeated successful attempts. Figure 5(b) displays an analogous mechanism on the feedback side: confidence values rose from 3.18 to 3.68 in the case of motivational feedback, this is compared with 3.52 in the case of explanatory feedback and 3.48 in the case of worked examples. In this place, motivation feedback mainly strengthens self-confidence by means of language encouragement, while completed examples can give a substitute source of support.



(a): mastery experience under different reward and constraint settings.



(b): Confidence response under motivational, explanatory, and worked-example support.



(c): Affective-state regulation under safe RL, unconstrained RL, and bandit adaptation

Figure 5: Mechanism validation of the proposed reinforcement learning strategy across mastery experience, support-induced confidence response, and affective-state regulation.

Figure 5(c) corresponds to physiological/affective states. The rate of overload or frustration events in the Safe RL group decreased from 21% to 10%, while removing the constraint reduced it to only 19%; the Bandit group remained stable at around 18%. These results indicate that the constraint is not merely an ancillary setting; it directly influences emotional fluctuations and the accumulation of failures during the learning process. At the algorithmic level, multi-objective rewards are more suitable for educational decision-making than a single accuracy target, as they simultaneously capture information regarding performance, persistence, and emotional load. On the education aspect, the rising self-efficacy is normally followed by higher finish rates and more steady development routes of performance. In the application respect, this mechanism has no limit on a single subject, and therefore thus can be shifted to platforms which are for programming, mathematics, language study. Recently done research on very strong adaptive path suggestion and unified path modeling has also taken the synergistic optimization of learning efficiency, personal state, and cognitive burden as a core problem.

This study still has several limitations that require clarification. First, the sample comes from a single programming micro-course scenario, and external validity still needs to be tested across more courses.Second, although self-efficacy was analyzed in conjunction with process proxy variables, the primary measurement still relies on a questionnaire. Third, online RL is still subject to cold-start effects in its early stages. Fourth, long-term transfer effects have not yet been tracked across semesters. Recent research in the field of safe reinforcement learning indicates that constraint or shielding mechanisms can reduce unnecessary high-risk actions; this approach also holds reference value for robust deployment in educational settings.

4 Conclusion

This research carries out development and carries out verification on a reinforcement learning frame for adaptive learning which takes self-efficacy as a core goal. Instead of only paying attention to accuracy or completion, this frame work puts perceived ability, continuous participation and learning effect into the identical decision-making process, hence making the strategy more able to respond to the way that learners actually make advancement in the course of study via difficulty, self-belief and continuous effort. The results coming from controlled experiments, ablation tests and heterogeneity analysis further indicate that the contribution which reinforcement learning makes in education lies not only in recommendation efficiency, but also in its capability that it can regulate support timing, feedback depth as well as challenge level along with the passage of time.

(1)This present article establishes a reinforcement learning framework which is used for the promotion of self-efficacy. It utilizes the study history, behavior tracks, and emotion signals, while it defines behaviors through the common selection of task hardness, feedback category, and support strength. A multi-goal reward connects self-efficacy increments, working results, involvement, and overload punishments inside one identical optimization procedure. By this method, the policy is made to sustain continual and in teaching meaningful individualization, not separate suggestions.

(2)The experiment's result proves that this framework can enhance self-efficacy, learning achievement and persistence at one time. Compared with the control situation, it keeps a more stable superiority in the aspects of self-efficacy promotion, post-test correctness, task finishing and decrease of quitting, it does not just promote one result. Extra researches on change along with time and basic working rules suggest that these benefits accumulate through continuous tasks, hence leading to higher self-confidence, better working outcomes, and stronger readiness to go on with learning.

(3)Hence, this research also puts forward several directions which are able to be utilized for the further development of work. One viewpoint is that we must strengthen explainable reinforcement learning, hence let teachers and learners have a better grasping of the method by which actions are distributed. Another method is to enlarge this frame by making use of safe reinforcement study, through the employment of more powerful constraints and protection rules to maintain more stable real-world utilization. A third item is to examine cross-domain transference in circumstances like programming, mathematics, and language learning, hence the framework can go past a single experimental setting.

Funding

One of the phased achievements of the research project "Practical Research on Cementing the Sense of Community for the Chinese Nation in Vocational College Classes from the Perspective of Mental Health Promotion" (2023ZJBZR14) of Zhang Pei's Famous Head Teacher Studio in Vocational Colleges of Xinjiang Uygur Autonomous Region (Document No. Xin Dang Jiao Chuan [2023] No. 128).

About the Author

Pei Zhang was in Urumqi, Xinjiang, an associate professor with a master's degree in psychology. She serves as the director of the Mental Health Education Teaching and Research Section, program director, and academic leader at the School of Educational Sciences, Xinjiang

Teacher’s College. She is a national second-level psychological counselor, a member of the Health Management and Promotion Professional Committee and the Psychological Counseling Sub-Committee under the National Health Vocational Education and Teaching Steering Committee, a director of the Xinjiang Psychological Society, an autonomous region-level expert for the Ministry of Education - UNICEF Social-Emotional Learning Project, as well as a director of both the Xinjiang Psychological Counselors Association and the Xinjiang Mental Health Association. She is also the host of the Xinjiang Uygur Autonomous Region Vocational College Famous Head Teacher Studio.

Her paper Interpersonal adjustment and depression in college students: The mediating effect of core self-evaluation and moderating effect of gender was published in the Journal of Psychology in Africa. She has published over 20 papers in domestic journals such as Modern Education Science and Xinjiang Social Sciences Forum. Additionally, she has published works including Research on Mental Health Education in Primary and Middle Schools by Xinjiang People's Publishing House and College Students' Mental Health Quality Training by Northeast Normal University Press.

References

[1] Mon, B. F., Wasfi, A., Hayajneh, M., Slim, A., & Abu Ali, N. (2023). Reinforcement learning in education: A literature review. Informatics, 10(3), 74.

[2] Li, X., Xu, H., Zhang, J., & Chang, H.-h. (2023). Deep reinforcement learning for adaptive learning systems. Journal of Educational and Behavioral Statistics, 48(2), 220–243.

[3] Osakwe, I., Chen, G., Fan, Y., Rakovic, M., Li, X., Singh, S., Molenaar, I., Bannert, M., & Gašević, D. (2023). Reinforcement learning for automatic detection of effective strategies for self-regulated learning. Computers and Education: Artificial Intelligence, 5, 100181.

[4] Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. Learning and Instruction, 83, 101620.

[5] Memarian, B., & Doleck, T. (2024). A scoping review of reinforcement learning in education. Computers and Education Open, 6, 100175.

[6] Ruan, S., Nie, A., Steenbergen, W., He, J., Zhang, J. Q., Guo, M., Liu, Y., Nguyen, K. D., Wang, C. Y., Ying, R., Landay, J. A., & Brunskill, E. (2024). Reinforcement learning tutors better supported lower-performing students in a math task. Machine Learning, 113, 3023–3048.

[7] Yun, Y., Dai, H., An, R., Zhang, Y., & Shang, X. (2024). Doubly constrained offline reinforcement learning for learning path recommendation. Knowledge-Based Systems, 284, 111242.

[8] Zhang, D. (2024). Using deep reinforcement learning to optimize the motivational incentive mechanism of online English learners. In Proceedings of the International Conference on Decision Science & Management (pp. 179–183). ACM.

[9] Getenet, S., Cantle, R., Redmond, P., & Albion, P. (2024). Students' digital technology attitude, literacy, and self-efficacy and their effect on online learning engagement. International Journal of Educational Technology in Higher Education, 21, 3.

[10] Shin, D. D. (2024). Curiosity promotes self-regulated learning and achievement in online courses for students with varying self-efficacy levels. Educational Psychology, 44(4), 455–474.

[11] Milani, S., Topin, N., Veloso, M., & Fang, F. (2024). Explainable reinforcement learning: A survey and comparative review. ACM Computing Surveys, 56(7), Article 168, 1–36.

[12] Bekkemoen, Y. (2024). Explainable reinforcement learning (XRL): A systematic literature review and taxonomy. Machine Learning, 113, 355–441.

[13] Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. Computers and Education: Artificial Intelligence, 8, 100349.

[14] Bauer, E., Sailer, M., Niklas, F., Greiff, S., Sarbu-Rothsching, S., Zottmann, J. M., Kiesewetter, J., Stadler, M., Fischer, M. R., Seidel, T., Urhahne, D., Sailer, M., & Fischer, F. (2025). AI-based adaptive feedback in simulations for teacher education: An experimental replication in the field. Journal of Computer Assisted Learning, 41(1), e13123.

[15] Bauer, E., Richters, C., Pickal, A. J., Klippert, M., Sailer, M., & Fischer, F. (2025). Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment in higher education. British Journal of Educational Technology, 56(5), 1735–1757.

[16] Bewersdorff, A., Hornberger, M., Nerdel, C., & Schiff, D. S. (2025). AI advocates and cautious critics: How AI attitudes, AI interest, use of AI, and AI literacy build university students' AI self-efficacy. Computers and Education: Artificial Intelligence, 8, 100340.

[17] Bergdahl, N., & Sjöberg, J. (2025). Attitudes, perceptions, and AI self-efficacy in K–12 education. Computers and Education: Artificial Intelligence, 8, 100358.

[18] Mekheimer, M. (2025). Technological self-efficacy, motivation, and contextual factors in advanced EFL e-learning: A mixed-methods study of strategy use and satisfaction. Humanities and Social Sciences Communications, 12, 677.

[19] Zhao, Z., An, Q., & Liu, J. (2025). Exploring AI tool adoption in higher education: Evidence from a PLS-SEM model integrating multimodal literacy, self-efficacy, and university support. Frontiers in Psychology, 16, 1619391.

[20] Schutte, N. S., & Li, H. (2025). The role of self-efficacy and curiosity in student use of artificial intelligence (AI). International Journal of Educational Technology in Higher Education, 22, 73.

[21] Riedmann, A., Schaper, P., & Lugrin, B. (2025). Reinforcement learning in education: A systematic literature review. International Journal of Artificial Intelligence in Education, 35, 2669–2723.

[22] Ruan, S., & Lu, K. (2025). Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization. Computers and Education: Artificial Intelligence, 9, 100463.

[23] Luo, G., Gu, H., Do ng, X., & Zhou, D. (2025). HA-LPR: A highly adaptive learning path recommendation. Education and Information Technologies, 30(10), 14597–14627.

[24] Zheng, Y., Wang, D., Zhang, J., Li, Y., et al. (2025). A unified framework for personalized learning pathway recommendation in e-learning contexts. Education and Information Technologies, 30(6), 7911–7948.

[25] Gerold, H., & Lucia, S. (2025). Safe reinforcement learning via adaptive robust model predictive shielding. Computers & Chemical Engineering, 206, 109521.