13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Optimizing Collaborative Learning: A Hard-Constraint Reinforcement Learning Approach to Fair Team Composition

Marcell Herbák[a], **Gergely Kovásznai**[a], **Ali Adil Adil**[b]

[a]Eszterházy Károly Catholic University
herbakmarcell@gmail.com
kovasznai.gergely@uni-eszterhazy.hu

[b]University of Debrecen
ali.adil@inf.unideb.hu

Abstract

Collaborative learning strategies are increasingly relied upon in modern educational settings to enhance student engagement and academic performance. However, the efficacy of these strategies is highly dependent on the composition of the student groups. Although well-structured teams can foster peer instruction and social development, creating balanced and effective groups presents a complex combinatorial optimization challenge. Traditional manual assignment methods are not only time-consuming and unscalable, but often fail to account for the student attributes required for optimal group dynamics, leading to suboptimal learning outcomes.

In real-world educational contexts, it is often insufficient to simply cluster students based on general similarity. The allocation process must enforce ”hard constraints” that are non-negotiable for the validity of the grouping. These mandatory requirements typically include maintaining specific gender ratios to ensure diversity and strictly bounding the variance in student performance to prevent radical disparities in knowledge levels within a single team.

The primary motivation for this research stems from the work of Adil et al. [1] regarding the formation of equitable student teams in STEAM education. Their study utilizes a Satisfiability Modulo Theories (SMT) approach to enforce mandatory requirements such as specific gender ratios and minimum aggregate skill levels. However, while their SMT-based solver guarantees constraint satisfaction for small

datasets, it faces scalability challenges and lacks the ability to ”learn” from past assignments. Our research aims to replace this static calculation with a Reinforcement Learning approach, which allows for the dynamic formation of groups that respect these same strict demographic and skill-based constraints but can scale more efficiently to larger student populations.

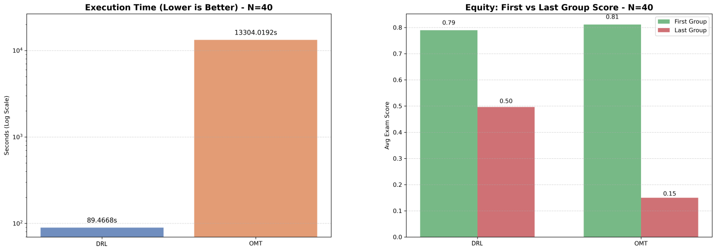
To ensure strict operational limits, Hu et al. introduced the Long and Short-Term Constraints (LSTC) framework for safe autonomous driving [3]. While unexplored in educational contexts, we adapt their ”safe RL” mechanism to student grouping by reinterpreting safety violations as pedagogical constraint violations. Leveraging their short-term feasibility checks, our model guarantees that no student placement breaches non-negotiable rules on gender distribution or knowledge disparity.

Hester et al. introduced Deep Q-learning from Demonstrations (DQfD) [2] to pre-train agents using expert data. We adapt this method – typically used in gaming – by treating historical student groups as ”expert demonstrations.” This initializes the agent with a baseline of valid structures, significantly accelerating the learning of constraints compared to training from scratch.

To train the model, we extracted seven critical features from the Student Performance and Learning Behavior Dataset for Educational Analytics dataset [4] to enable compatibility assessment, constraint enforcement, and skill gap minimization. The Gender feature serves as the primary variable for strictly enforcing diversity ratios. To minimize skill disparities, the model utilizes three quantitative metrics: Exam Score (curriculum mastery), Assignment Completion (effort consistency) and Attendance (reliability). Complementary skill attributes were also included: Extracurricular (social engagement), Discussions (collaboration) and EduTech (digital literacy). This feature selection reduces state space noise while retaining the necessary information for optimizing pedagogical balance and constraint satisfaction.

To address the combinatorial complexity of student group formation, we implemented a Deep Q-Network (DQN) agent that acts as a trainable policy. Unlike static heuristics, the agent makes sequential decisions based on student profiles and global context to optimize a multi-objective reward function. This function enforces strict gender ratios (40-60%) while minimizing skill variance and incentivizing student engagement. By leveraging a windowed, pre-sorted observation space, the agent approximates global optimality and satisfies hard constraints efficiently, avoiding the computational overhead of exhaustive search.

After training our Deep Reinforcement Learning (DRL) agent with 13963 student records, we benchmarked the agent with $N = 40$ students against an exact Optimization Modulo Theories (OMT) solver. The DRL agent achieved a skill variance of 0.1995, nearly identical to the mathematical optimum of 0.1978, while cutting execution time by over 99% (Figure 1a) compared to the OMT solver. The execution time of the DRL model consists of its training and inference (clustering) phase. Qualitatively, the DRL approach demonstrated superior equity: unlike the OMT solver, which optimized the top group at the cost of a ”failure cluster” (Figure 1b), the DRL agent’s sequential decision-making distributed talent more



(a) Execution time.

(b) Equity.

Figure 1. Comparisons of DRL and OMT approaches.

evenly, maintaining a viable final group average of 0.50. This confirms that the DRL approach offers a scalable, equitable alternative to exact methods.

Future work will focus on validating the statistical robustness of the model and transitioning the system to a deployable educational tool. To ensure the reliability of our findings, we plan to expand the experimental framework by conducting a few more independent runs across different random seeds for both the OMT and DRL solvers, alongside a stability test of DRL inference trials to quantify policy consistency. Beyond validation, the project will evolve into a user-facing web application that allows educators to import student datasets and visualize generated team compositions instantly. A critical feature of this deployment will be a ”Continuous Learning” pipeline: as users validate and accept specific group configurations, this data will be fed back into the system to retrain the agent in a background thread, allowing the policy to adapt to real-world nuances over time. Finally, we aim to integrate ”Human-in-the-Loop” constraints, enabling users to define mandatory rules – such as specific student pairings, minimum group score, etc. – which the agent will learn to treat as inviolable hard constraints during the sequential selection process.

References

[1] A. A. Adil, G. Kovásznai, M. Khalid: *Automated fair team formation in STEAM activities using Satisfiability Modulo Theories (SMT)*, Annales Mathematicae et Informaticae 61 (2025), pp. 1–14, doi: `10.33039/ami.2025.10.001`.

[2] T. Hester, T. Hester, O. Pietquin, M. Lancto, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. Agapiou, J. Z. Leibo, A. Gruslys: *Deep Q-learning from Demonstrations*, ACM Transactions on Graphics 32 (2018), pp. 1–12, doi: `10.1609/aaai.v32i1.11757`.

[3] X. Hu, P. Chen, Y. Wen, B. Tang, L. Chen: *Long and Short-Term Constraints Driven Safe Reinforcement Learning for Autonomous Driving*, 2024, arXiv: `2403.18209 [cs.LG]`, url: `https://arxiv.org/abs/2403.18209`.

[4] K. Najem: *Student Performance and Learning Behavior Dataset for Educational Analytics*, version v1.0, Zenodo, 2025, doi: `10.5281/zenodo.16459132`, url: `https://zenodo.org/records/16459132`.