Case Study

A case study has been conducted in a real teaching scenario in the School of Computer Science, South China Normal University. The undergraduate course "Programming" is selected. The course has 68 knowledge points and a set of 240 exercises, each exercise contains 1 to 3 knowledge points. The students can do the exercises online. In the case study, we used 533 students' 384,945 exercise answer records, from 09/2020 to 01/2021, to train a DKT model (AUC of 0.83606). In order to train ExamGAN, following method introduced in Section 4.4, we randomly generated 100 groups from these 533 students (each with 30 students) to generate the training data.

Three real classes of the same course in the new semester from 02/2021 to 04/2021 are involved in the case study. For each class, a real exam script is generated at the end of the semester using ExamGAN. For each class, the knowledge mastery level of students is identified based on exercises answered in the semester following Section 4.1 and fed to the generator. Each exam script consists of questions from the exam question bank which includes the set of exercises and 100 more questions. Each exam script contains 40 questions, and the full score is 100, 2.5 for each question. From test results of the three classes, difficulty, distinguishability, rationality, and validity (denoted as Dif, Dis, Rat and Val respectively) of the generated exam scripts are derived and shown in column 2-4 in Table 4. Clearly, they are highly desirable and consistent for all classes.

TABLE 4. Case study where C-# refers to class-#.

	C-1	C-2	C-3	DKE-based performance estimation			
	ExamGAN			ExamGAN	ExamVAE	GA	RSF
Dif	0.754	0.712	0.723	0.707	0.630	0.759	0.589
Dis	0.359	0.397	0.384	0.369	0.338	0.288	0.465
Rat	0.958	0.962	0.972	0.975	0.866	0.890	0.745
Val	0.934	0.929	0.940	0.926	0.945	0.970	0.912

For each class, we also generate additional exam scripts using ExamGAN and baselines. In the same way as in Section 6.1.2, the trained DKT model is used to predict the results of students in the class if working on these exam scripts. From the predicted results, the four properties of the exam scripts are derived. In column 5-8 in Table 4, average of the properties over three classes are reported. The similarity between column 1-3 and column 4 verifies the experiment results reported for ExamGAN in Section 6.1.2 are reliable, i.e., similar to asking students to really work on the exam scripts. Compared with baselines, the advantage of ExamGAN is demonstrated again in the case study.

To further evaluate the three exam scripts generated using the proposed ExamGAN, we invited three teachers of the course as the human expert. Based on their teaching experiences, they give score 0-5 (higher is better) to each exam script from the perspective that human experts can evaluate, i.e., "Suitability of difficulty level", "Suitability of distinguishability", and "Overall suitability". The averages of their scores are below:

Table 5. Average score of three teachers

	script 1	script 2	script 3
Suitability of difficulty level	4.3	4.3	4.0
Suitability of distinguishability	3.7	3.7	4.0
Overall suitability	4.0	3.7	4.0

The human expert evaluations clearly affirmed the quality of the generated exam scripts.