Student Activity Simulation with Simplified PFA model - Report

Jingquan Li

November 3, 2024

1 - Evaluation

The measurements I chose to evaluate the goodness of the model are accuracy and RMSE. Accuracy = number of correct score predictions / number of false score predictions. RMSE is the root mean square deviation, which is used to evaluate the error between the predicted score and the ground truth score.

The results of the two measurements for each simulation setting are shown in the table below. I also changed the gamma value, either bigger (0.7) or smaller (0.3) to see the influence of gamma value on the results. The corresponding accuracy and RMSE value after changing the gamma value are also included in the following table.

	Accuracy	RMSE
Low learning no-skill (gamma=0.5)	0.6820	0.5639
Low learning no-skill (gamma=0.3)	0.5899	0.6404
Low learning no-skill (gamma=0.7)	0.4147	0.7650
High learning no-skill (gamma=0.5)	0.5853	0.6440
High learning no-skill (gamma=0.3)	0.5899	0.6404
High learning no-skill (gamma=0.7)	0.4700	0.7280
Heterogeneous Learning no-skill (gamma=0.5)	0.5853	0.6440
Heterogeneous Learning no-skill (gamma=0.3)	0.5899	0.6404
Heterogeneous Learning no-skill (gamma=0.7)	0.4147	0.7650
Low learning one-skill (gamma=0.5)	0.6728	0.5720
Low learning one-skill (gamma=0.3)	0.5899	0.6404
Low learning one-skill (gamma=0.7)	0.4147	0.7650
High learning one-skill (gamma=0.5)	0.6037	0.6295
High learning one-skill (gamma=0.3)	0.5899	0.6404
High learning one-skill (gamma=0.7)	0.7880	0.4604

When gamma is set to the default value (i.e. gamma = 0.5), the results of the Low Learning No-Skill PFA model is closest to the gold standard one, making it the best fit simulation model. This is because it achieves the highest accuracy value and the lowest RMSE value, indicating it makes the most correct predictions and the least error value.

When gamma is set to a smaller value (i.e. gamma = 0.3), the results of all the models are the same. When gamma is set to a bigger value (i.e. gamma = 0.7), the results of the High Learning One-Skill PFA model is closest to the gold standard one, making it the best fit simulation model.

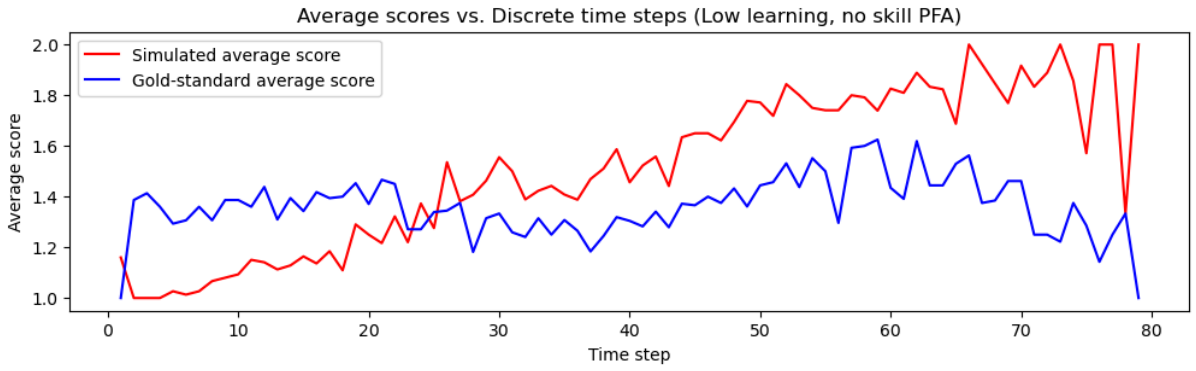
2 - Plots

Figure 1: Plot the average of all simulated student scores on all problems in time as well as the average of all gold-standard student scores in time.

Here I was required to draw box plots, but I tried a lot of times and could not figure out the correct results. I think maybe I didn't understand the box plots clearly, and I hope there would be potential opportunities to discuss this part.

As a substitute option, I plot the line chart of each simulation model. These plots describe the change of average scores of all simulated scores across different time steps. Here are some findings I got from these plots:

- The simulated results using the one-skill model achieve high average scores more quickly at the early time steps. We can particularly see this kind of performance from the last plot (high learning one-skill model). The curve goes up quickly at first and achieves a quite high and stable state early.
- At early time steps, the gold-standard average score curve goes up more quickly than the simulated one, but after several time steps, the simulated may go higher than the gold-standard one. This indicates that at early time steps, the results simulated by the PFA model usually achieve lower scores than the ground truth scores, while at later time steps, the simulated results usually achieve higher scores than the ground truth ones. The simulation effect is best at a crossing point, whose time step is approximately around 25 for the no-skill model, and around 10 for the one-skill model.
- At a high learning parameter setting, the students tend to achieve higher scores when the time steps are still small. This potentially indicates that under a higher learning rate, students may learn quickly from the former experiences, enabling them to perform well on the following problems earlier.



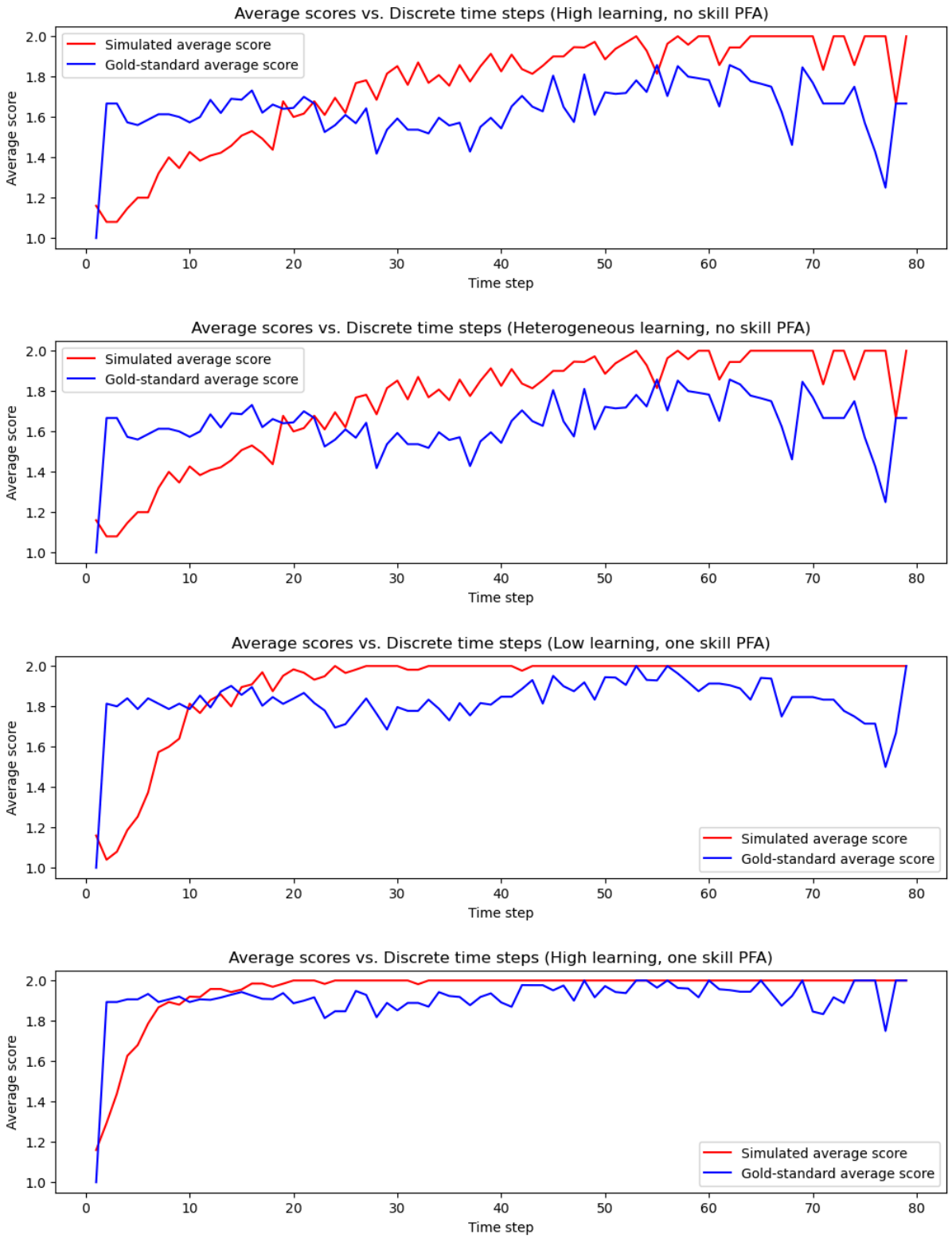
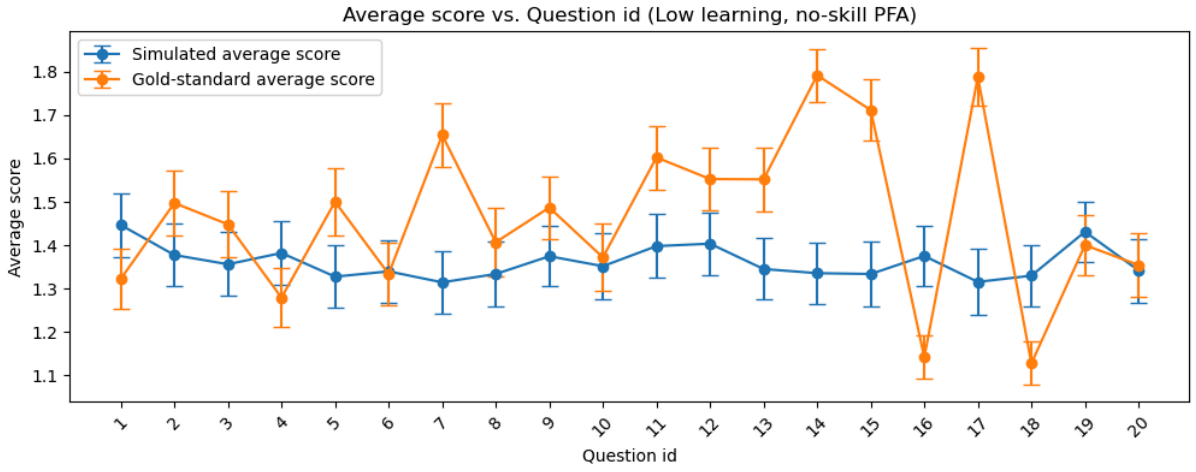


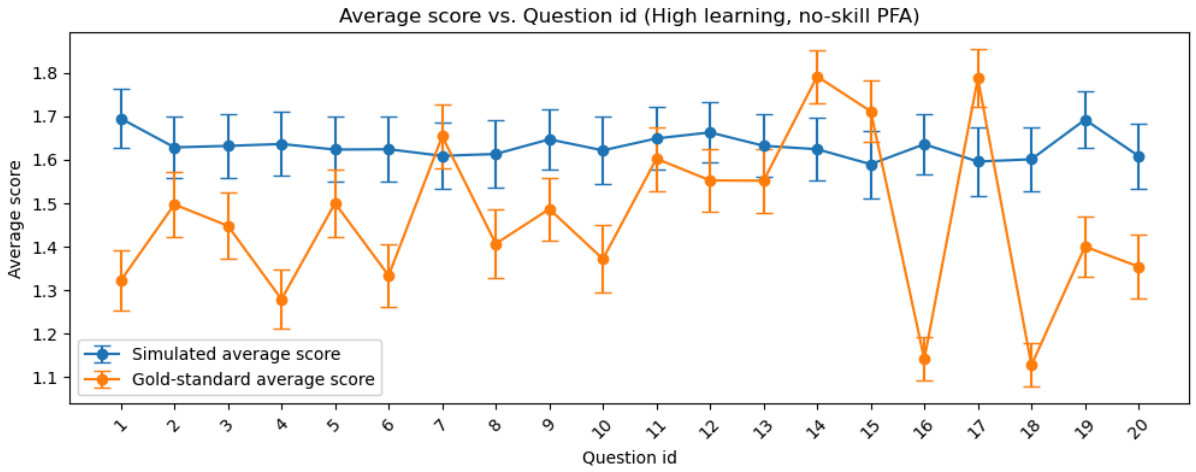
Figure 2: Plot the average student scores over all time-steps for each question for each simulated setting as well as the gold-standard.

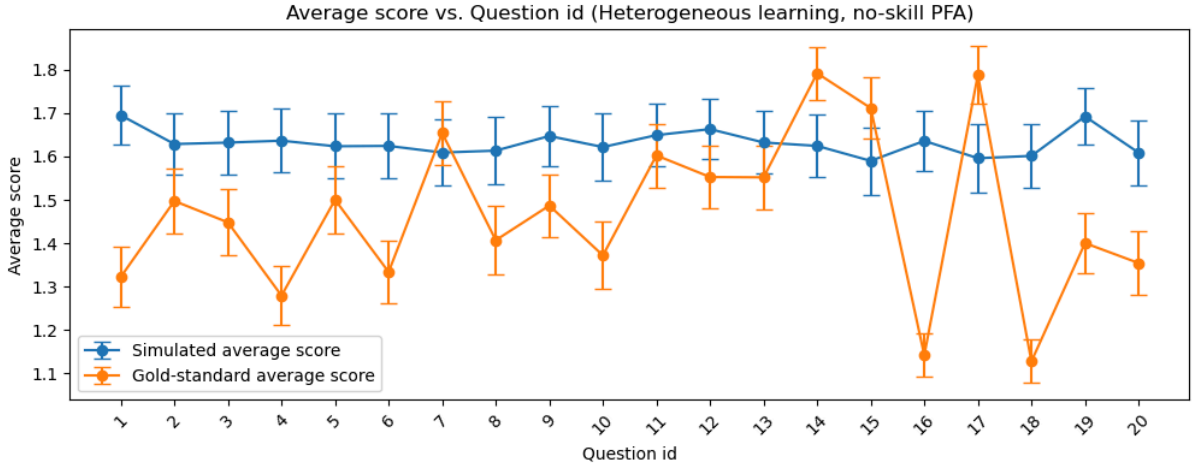
Here are some findings I drew from the following figures:

- Overall, the results simulated by no-skill models achieved relatively lower scores than the ground truth value, while those simulated by one-skill models tend to achieve much higher scores than the actual value.

- The scores curve simulated by these PFA models look much more stable than the actual one. The one-skill model performs a little better than the no-skill model, which may probably be due to its feature of considering different skill sets affiliated to specific problems when doing the simulation.
- However, although the no-skill model perform worse than the one-skill model when simulating the dynamics features of the curve, it has smaller error values than the one-skill models, as its values go around the ground truth curve, while the curve of the one-skill models goes high up beyond the ground truth curve.







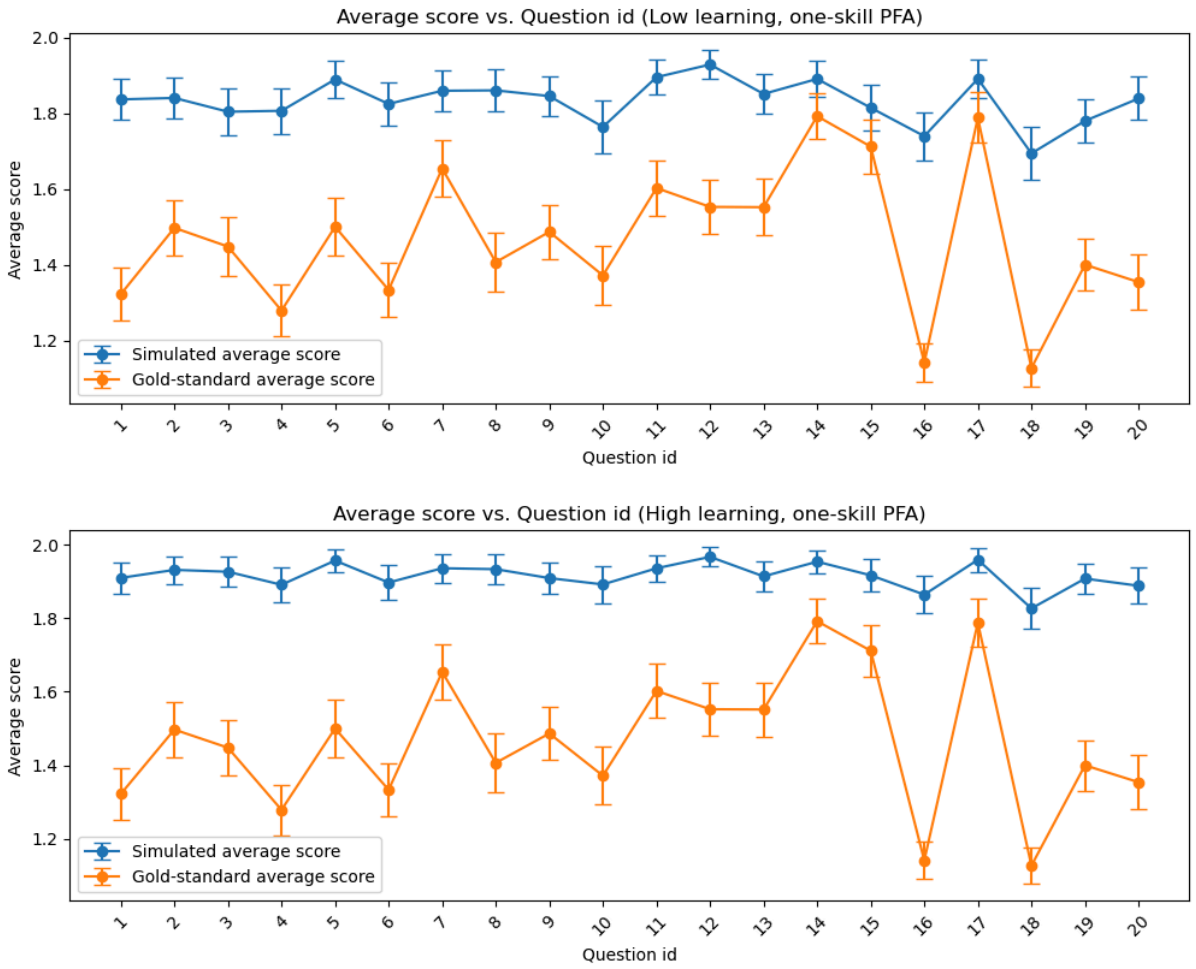
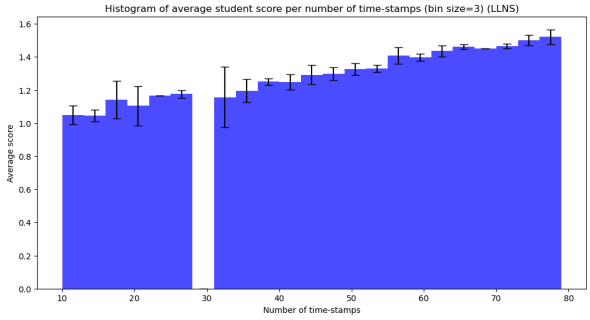
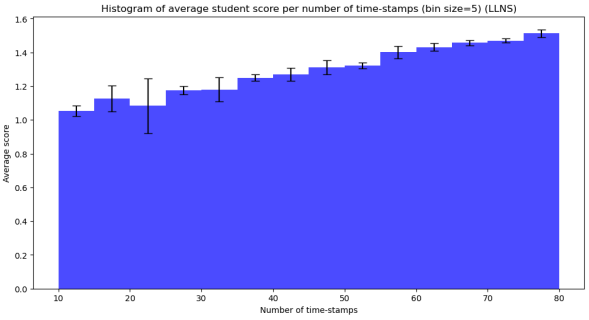


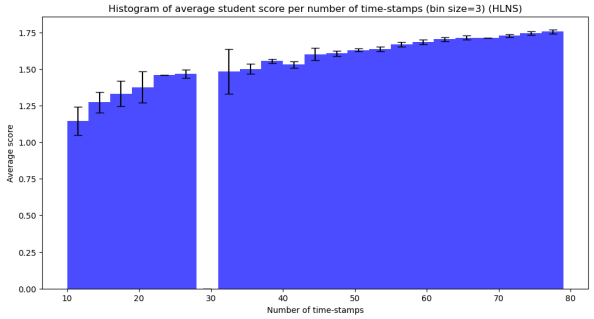
Figure 3: Plot the histogram of average student scores over all problems and all time-stamps per the number of time-stamps for each student.

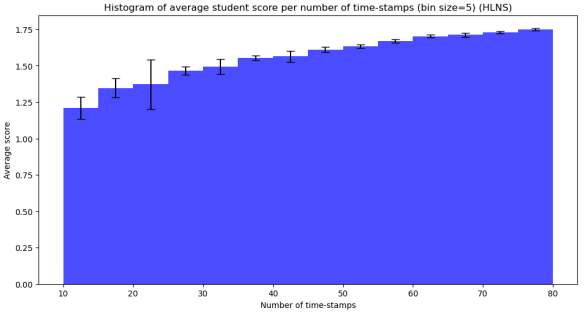
These figures mainly demonstrate how the students' scores change as the time steps increase. The findings I drew from these figures are as follows:

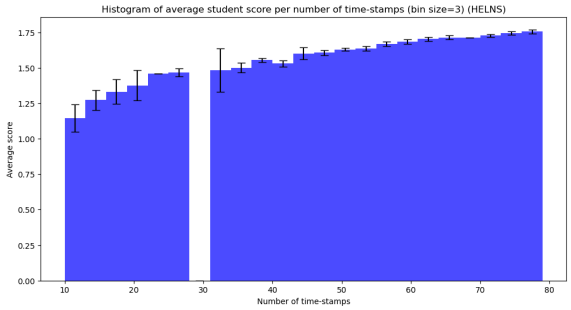
- Overall, the students achieve better scores as the time steps increase. This is consistent with the intuitive, because as students take more exercises, they will accumulate more experiences on solving different problems, and may achieve better results.
- When the bin size is small (i.e. bin_size = 3), the figures display more details and the range between the max value and min value is in total larger. The error bar of the small bin size graphs tends to be smaller than that of big bin size graphs.
- When the time steps grow big, the confidence intervals tend to be small, indicating the stability of the scores at high time steps.

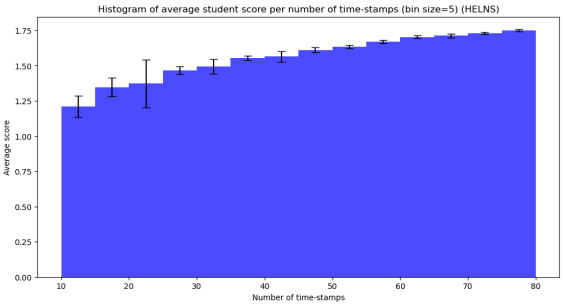


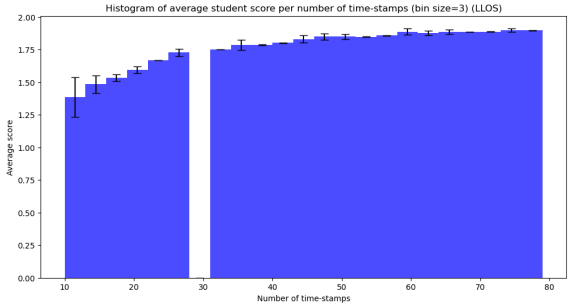


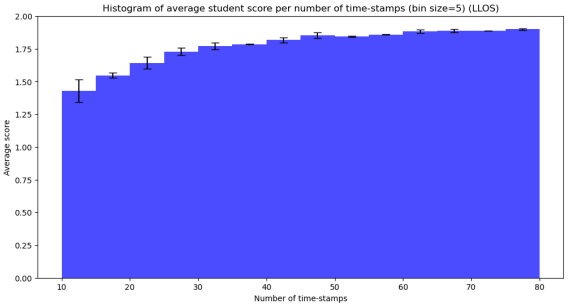


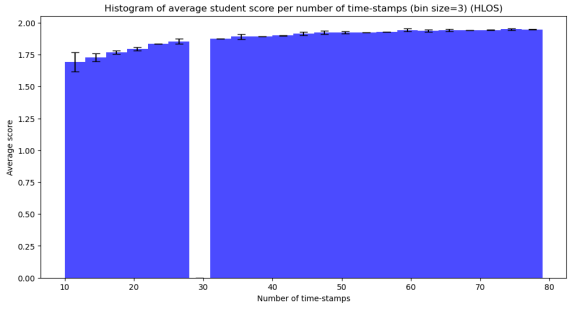


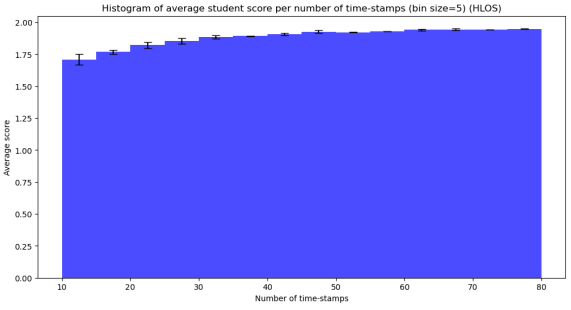












3 - Discussion

Discussion-1: Which simulation setting fits the observed scores best in each plot setting?

In the first plot setting, the High Learning One-Skill model fits the observed scores best, since the distance between the simulated curve and the gold-standard curve is overall smaller than others. In the second plot setting, the Heterogeneous Learning No-Skill model fits the observed scores best, since the simulated curve crosses more times with the gold-standard curve, and it mainly swings around the ground truth value. The third plot setting didn't plot the gold-standard curve, so we can't draw the conclusion of the best fit model from this plot setting.

Discussion-2: What is a possible explanation of the observed behaviors?

One skill model can deal with early stage learning better, because it considers the specific skill of each problem. That's why it can achieve high performance quickly in the first plot setting (figure 1). However, sometimes it may overestimate the learning scores. No skill model is simpler and adapts slowly, so it can keep the score close to the ground truth value and more stable with smaller errors. In figure 3, the confidence intervals become smaller as the time steps increase, indicating that as the exercises increase, the student's scores become more predictable, which is consistent with the intuition that experiences bring better performance.

Discussion-3: How can you further improve the fit?

One way is to combine the no-skill model and one-skill model, so that we can combine their features of quick early-stage learning and stable performance, which may lead to more accurate simulation. Another trying direction may be the use of a changeable learning rate. The model could use a high learning rate at an early stage, and we can decrease its learning rate as the students gain more experience.