Comparing Linear and Adaptive Mathematics Tests: A Pilot Study Using Automated Testing

Øyvind Aas[1] and Wasana Leithe[1]

Kristiania University College
oyvind.aas@kristiania.no

We conduct a lab-based randomized controlled trial to assess the effects of linear versus adaptive mathematics test on learning outcomes. In the linear test, each question has a set difficulty, while the adaptive test modifies its difficulty based on the participant's performance. Our research question is: do adaptive tests generate better learning outcomes than linear ones?

We hypothesize that adaptive tests may be more effective for learning. They give feedback on relevant exercises for the students' level, allow students to progress at their own pace and proficiency, and potentially save time by quickly aligning to the student's level, unlike linear tests which might become tedious.

1 Experimental design

The participants in the experiments are 47 students. The experiment was conducted in the spring of 2023, and the participants received a giftcard of 300NOK.

Questions used were sourced from graded problem sets accumulated over several years of teaching, allowing us to determine the percetange of students who answered each question correctly. A question that 95% of students answer correctly, for example, was assigned a difficulty level of 5 (calculated as $100-95$). These multiple-choice questions are from introductory calculus topics.

Participants were randomly assigned to either a linear or adaptive test. The linear test presented 6 questions with set difficulty level, starting at 30 going to 80 on a 0-100 scale. In contrast, the adaptiv test has 6 questions with changing difficulty levels determined by a Bayesian learning model. Participants began with a rank of 50 on a 0-100 scale. The system presents a question where the likelihood of a correct answer is roughtly 50%. A correct answer would increase the participant's rank while decreasing the question's difficulty level. Subsequent questions were presented similarly.

Both tests provide feedback after each question and a concluding summary. This summary present a visual progress chart, a written evaluation of scores and grades, and a recommendation for which topics to focus on next. After testing, participants were given an identical 4-question exam to measure their learning, and then completed a brief survey.

2 Data and Results

We use a regression model to estimate the causal effect of the test type on the exam score, controlling for experiment group fixed effects. We measure the exam

score for individual i as a function of a constant term and a linear term which is a dummy variable that takes the value 1 if the test is linear and 0 if the test is adaptive. Last, we control for the three different experiment rounds.

Table 1. Regression results of exam results on test type.

	(1)	(2)
VARIABLES	Exam Results	Exam Results
Type of math test: linear	-26.18**	-27.57***
	(10.03)	(9.705)
Experiment group: 2		0.0307
		(10.31)
Experiment group: 3		-28.54*
		(14.84)
Constant	53.26***	58.21***
	(7.745)	(9.067)
Observations	47	47
R-squared	0.132	0.212

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Model (1) in Table 1 shows that participants from the linear test group scored significantly lower by 26.18 percentage points compared to the adaptive test group ($p < 0.05$). This finding is the same after controlling for experiment group fixed effects in Model (2).

After the exam, participants provided feedback via a brief survey on the system. They found it user-friendly and believed it could enhance their academic performance. They didn't think the questions were easy. Feedback, including the graphical overview and the score and grade, was well received. Most notably, guidance on which topics to concentrate on received the highest average score. Furthermore, participants said they would use the system if it was available.

3 Discussion and Conclusion

Our results suggest that adaptive testing can improve mathematics learning. However, it is important to consider these findings in the context of a pilot study. The limited number of participants constrains the study's statistical power. Additionally, using just four exam questions might lead to inaccuracies in assessing learning outcomes. Further research should consider expanding the sample size, refining the learning outcome measurements, and investigating other facets of adaptive testing like motivation, engagement, and self-regulation.