Automated Adaptive Testing vs. Linear Testing in Undergraduate Mathematics*

Øyvind Aas[1] and Wasana Leithe[1]

Kristiania University College
oyvind.aas@kristiania.no

Abstract. We conduct a lab-based randomized controlled trial with 47 undergraduate students in mathematics, comparing an automated adaptive testing system, which adjusts difficulty based on performance, to traditional linear tests. Results shows that students using the adaptive test, for formative assessment, scored 26.2 percentage points higher on a subsequent exam than those in the linear test group ($p < 0.05$). Feedback indicate that participants found the system user-friendly, believed it could improve their performance, and valued the tailored feedback, particularly guidance on focus areas.

Keywords: Automated adaptive test · Undergraduate teaching · Randomized control trial.

1 Introduction

Mathematics plays a central role in engineering, data science, and economics. To excel in these fields, students need to develop strong analytical skills, mathematical proficiency, and a problem-solving mindset. Consequently, introductory mathematics courses are often a foundational component across multiple study programs and tend to have high enrollment numbers. However, students entering these courses come from diverse educational backgrounds, leading to varying levels of mathematical preparedness. These disparities can create knowledge gaps that hinder effective learning and contribute to high attrition rates. A more individualized approach, providing tailored feedback and support, can be essential for improving learning outcomes and reducing dropout rates.

Traditionally, tests are used to evaluate learning outcomes or assign grades, often referred to as summative assessment. However, they can also serve as a powerful tool for learning improvement. The "testing effect" suggests that taking practice tests can improve knowledge retention and understanding, as a formative assessment [12]. Such practice tests, often referred to as low-stakes assessments, can help students identify misconceptions or learning gaps, thereby improving their learning without the stress associated with high-stakes exams.

* This work was supported by the Pedagogical Development Fund at Kristiania University College. The funding source had no involvement in the project. Declaration of interest: none.

Computerized adaptive testing (CAT) has a long history in education, offering strong psychometric properties and personalized assessments [2]. Klinkenberg et al. [9] developed a web-based CAT system for primary school arithmetics using the Elo rating model, finding high correlations with standardized test scores (ranging from .78 to .84). Jansen [8] conducted a six-week field experiment with grades 3-6, using the Klinkenberg system. They had one control group, and three treatment groups: easy, medium and difficult. The easy group showed greater improvement than the control group, while the medium and difficult groups did not.[1] Ling et al. [10] compared CAT and fixed-item tests (FIT) for middle school students, finding no significant performance differences, but noted that immediate feedback improved outcomes across all test types, highlighting its positive impact on learning.

Previous studies have focused primarily on basic arithmetic skills in primary and middle school students [9, 8, 10]. To evaluate adaptive testing in higher education, we conducted a lab-based randomized controlled trial (RCT) comparing linear and adaptive mathematics tests. In the linear test, questions increased in difficulty at a fixed rate, while in the adaptive test, the difficulty level is adjusted based on the participants responses. We call it a linear test because it is a fixed-item test that gets progressively more difficult. Our research question is: Does adaptive testing improve learning outcomes more effectively than linear testing in a digital environment?

We hypothesize that adaptive tests improves learning by providing exercises at the appropriate level, allowing students to progress just beyond their current proficiency. This personalized approach creates a sense of achievement. In contrast, linear tests with fixed difficulty may not match individual learning needs, potentially causing disengagement and frustration.

2 Related literature

Roediger et al. [12] highlight the benefits of testing beyond summative assessment, focusing on its role in enhancing long-term learning. They identify ten benefits, such as improved retention, better knowledge organization, and increased metacognitive awareness. Frequent testing promotes regular study habits, offers feedback, and aids knowledge transfer, underscoring its value as a pedagogical tool in education. We contribute to this literature by causally estimating the effect of formative assessment on learning outcomes.

Klinkenberg et al. [9] develop a computerized adaptive practice (CAP) system for primary school arithmetic using an Elo-based model that accounts for accuracy and response time. Data from 3648 children over ten months show the system provides precise ability measurements and valuable diagnostic insights, supporting targeted interventions and high student engagement.

Conejo et al. [3] introduced the SIETTE system, a web-based assessment platform combining classical test theory, item response theory, and computer-

[1] The sucess rate was set to 90% for the easy group, 85% for the medium and 60% for the difficult.

ized adaptive testing. It supports various question types and features like hints, feedback, and spaced repetition, enhancing both formative and summative assessments. Their results show that SIETTE effectively supports adaptive learning and integrates well with systems like Moodle for comprehensive educational assessment.

Adaptive assessment systems have been applied in programming courses, music psychology and cognitive diagnostics. Yang et al. [13] use a quasi-experiment to integrate CAT with a memory retention algorithm in a 7-week "Introductory Programming Language" course, improving performance and engagement compared to traditional and CAT-only methods. Harrison et al. [7] developed a melodic discrimination test using item response theory, CAT, and automatic item generation, demonstrating its reliability, validity, and efficiency as a model for psychological testing. McGlohen and Chang [11] combine CAT with cognitive diagnosis models to enhance both ability estimation and diagnostic feedback. By optimizing item selection based on ability and attribute mastery, their method improves the precision of both metrics while controlling item exposure, making it valuable for personalized assessments and targeted instructional interventions.

We contribute to the literature by causally estimating whether a CAT system improves learning outcomes over a FIT system in higher education mathematics. Using a lab-based randomized experiment, we compare actual learning outcomes under both testing regimes, unlike some of the previous studies that focus on test consistency or self-reported outcomes without a control group.

Last, a small literature explores teachers' use and perception of adaptive tools to identify student needs. Alfageh et al. [1] found that adaptive diagnostic assessments help elementary teachers tailor instruction, improve student grouping, and enhance lesson planning and communication. Teachers reported these tools as beneficial for improving mathematics education. We propose a novel way to use learning paths together with CAT to improve learning outcomes.

3 Methods

3.1 Participants

We carry out the experiment in the spring of 2023 on campus in regular classrooms, without any connection to actual courses. Participants were recruited from previous student cohorts, spanning three years. All participants voluntarily participate and all the data is anonymous. The students receive a giftcard of 300NOK to participate in the experiment. The experiment took about 30 - 40 min to complete. We had 47 participants across three rounds of experiments.

3.2 The database of questions

We have built a database of questions over several years of teaching. Based on the last four years of student cohort we have graded all problem sets and know the fraction of students who answer each question correct. We assign a difficulty

rank of a question based on the fraction of students who completed the question. For example, an easy question that 95% of students can answer, gets a difficulty rank of $100-95=5$, while a hard question that only 10% of students completed, gets a difficulty rank of $100-10=90$.

Our database of questions are standard calculus exercises for undergraduate mathematics, like "compute the derivative of f", "find the local maximum", or text assignments resulting in integration operations.

3.3 Experimental protocol

Participants are randomly assigned to one of two conditions: a linear test or an adaptive test.

Linear test: In the linear test, participants answer six questions, with difficulty levels starting at 30 and increasing by 10 for each subsequent question, ranging from 30 to 80. After each question, participants are provided with immediate feedback with the correct answer. The difficulty level for each question is fixed and does not depend on the participant's performance.

Adaptive test: The theoretical foundation of our adaptive test is based on item response theory, and provides a robust method for estimating students abilities and tailoring educational experiences [2]. We combine it with the Elo [4] rating system, traditionally used in chess competition . Klinkenberg et al. [9] is an earlier integration of the Elo system with CAT in their web-based platform for adaptive math practice. Our model, like Klinkenberg et al.[9], allows for real-time estimation of students ability and the question difficulty.

The principle of the test is as follows. All participants begin with an initial rank, denoted $r_0 = 50$ on a scale from 0 to 100. The system selects questions from a database based on the participant's current rank, r_t, and a random fluctuation component. Specifically, the difficulty level of the selected question, d_i, is determined as follows:

$$d_i = r_t + 10 \cdot (\phi_1 - \phi_2) \tag{1}$$

where $\phi_1, \phi_2 \sim U(0,1)$ are uniformly distributed random variables. For example, if $\phi_1 = 0.6$ and $\phi_2 = 0.4$, then the difficulty level is calculated as $d_i = 50 + 10 \cdot (0.6 - 0.4) = 52$. We want the question difficulty level to be similar to the participants rank with some noise to allow the participant to face slightly easier or slightly harder questions that their current rank. This randomness means that the participant will converge quicker towards their true ranking level.

Participants answer the question and receive immediate feedback. If the participant's answer is correct, their rank r_t is updated upward, and the question's difficulty level d_i is updated downward. If the answer is incorrect, the participant's rank decreases, and the question's difficulty increases. The updating rules are given by:

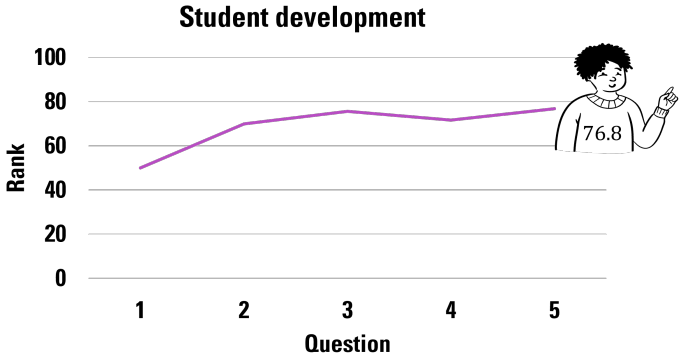
$$r_t = r_{t-1} + 10 \cdot (\text{outcome} - P(\text{participant correct})) \tag{2}$$

$$d_{i,t} = d_{i,t-1} + 10 \cdot (\text{outcome} - (1 - P(\text{participant correct}))) \tag{3}$$

where $P(\text{participant correct})$ is the probability that the participant answers the question correctly, computed by the logistic function:

$$P(\text{participant correct}) = \frac{1}{1 + 10^{(d_i - r)/30}} \tag{4}$$

To illustrate the updating rule consider the following example. Suppose the initial rank is $r_0 = 50$ and the difficulty level of the question is $d_i = 52$, the likelihood that the participant answers correctly is then approximately 46.2%. If the participant answers correctly, their rank is updated to $r_1 = 55.4$, and the question difficulty level is updated to $d_{i,t} = 46.6$. This process continues until the participant has answered six questions. After each question, participants are provided with immediate feedback with the correct answer.



Percentage score = 75%, Grade = B, Feedback = Bra jobbet, sjekk tilbakemelding

Topic	Percentage score	Comment
Analysis	50%	Review material
Combinatorial optimization	100%	Move on

Fig. 1. A graphical illustration of the overall feedback screen after the test. This screen is independent of the treatment of the experiment.

Feedback: Regardless of the test condition, all participants receive the same overall feedback after completing the test. Figure 1 illustrates the feedback screen presented to participants. The feedback includes:

1. A graphical representation of their rank progression throughout the test.
2. A written summary of their overall performance, including their percentage score, grade, and feedback on specific areas of strength and weakness.
3. An overview of the topics covered, their respective scores, and suggestions for further study.

Final exam: After viewing the feedback, participants take a final exam consisting of four questions. These questions are the same for all participants,

regardless of whether they took the linear or adaptive test. The score on this final exam serves as the outcome variable for the study.

Outcome measures: The main outcome variable is the participant's score on the final exam. The primary explanatory variable is the test condition (linear vs. adaptive) to which the participant was assigned. Table 1 presents the descriptive statistics of the exam score.

Table 1. Descriptive statistics of the exam score.

	Observations	Mean	Standard deviation	Min	Max
Exam score in	47	39.89	36.36	0	100
(i) Adaptive test	23	53.26	37.16	0	100
(ii) Linear test	24	27.08	31.20	0	100

4 Results

We estimate the causal effect of the adaptive test versus the linear test on exam scores using the following regression model:

$$\text{Exam}_i = \alpha + \beta_1 Z_i + \beta_2 X_i + \varepsilon_i, \tag{5}$$

where Exam_i represents the exam score for individual i. The variable Z_i is a dummy variable that equals 1 if the participant was assigned to the linear test and 0 if assigned to the adaptive test. The coefficient α is the score on the adaptive test. The coefficient β_1 captures the difference in exam scores between the linear and adaptive tests. The variable X_i includes controls for the three experimental rounds. The coefficient β_2 captures any differences between the different experiment rounds, and ε_i is the error term.

Table 2 presents the regression results. In Model 1, we examine the difference in exam scores between the linear and adaptive tests. The coefficient for the linear test is -26.18, indicating that participants in the linear test scored, on average, 26.18 percentage points lower than those in the adaptive test. This result is statistically significant at the 5% level ($p < 0.05$). The mean score on the adaptive test is the constant term, 53.26%.[2]

In Model 2, we include controls for the experimental rounds. The effect of the linear test remains similar, with a coefficient of -27.57, and the significance level increases to 1% ($p < 0.01$). The experimental round coefficients indicate that there are no statistically significant differences between the rounds compared to the first round. The R-squared value increases from 0.132 in Model 1 to 0.212 in Model 2, suggesting that the experimental round controls explain some additional variance in exam scores. The students in the first and second round

[2] Model 1 is equivalent to a t-test comparing the two groups.

Table 2. Regression results of exam scores on test type.

	(1)	(2)
VARIABLES	Exam Score	Exam Score
Type of Test: Linear	-26.18**	-27.57***
	(10.03)	(9.705)
Experiment Round: 2 (first year students)		0.0307
		(10.31)
Experiment Round: 3 (third year students)		-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
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

were all first-year students. In the third round all participants were in the final year of the 3-year bachelor degree. However, the main effect of the linear versus adaptive test remains robust.

4.1 Survey responses

After completing the exam, participants were asked to rate their agreement with several statements about the system on a scale from 1 to 10, with 10 indicating strong agreement. Figure 2 displays the mean responses with 95% confidence intervals. A score of 5 represents indifference to the statement.

Participants found the system easy to use (mean = 6.1) and believed it could improve their academic performance (mean = 6.4). The questions were appropriately challenging (mean = 4.2). Feedback was highly valued, especially the graphical summary (mean = 7.9), score and grade feedback (mean = 8.1), and personalized study recommendations (mean = 8.6). Most participants would use the system again (mean = 7.3). These responses, together with the experiment, indicate that the CAT system effectively supports both exam performance and user experience, with feedback helping users identify areas for improvement.

5 Discussion

In contrast to Ling et al. [10], our study with higher education students shows a significant performance difference between CAT and linear tests, aligning more with the findings in Jansen et al. [8] on improved performance with adaptive methods. This discrepancy may be due to differences in age and ability levels between middle school and university students, suggesting that adaptive testing may be more effective for higher-ability learners.

We believe that students learn best when working on exercises slightly beyond their current ability, placing them in the "zone of proximal development." If

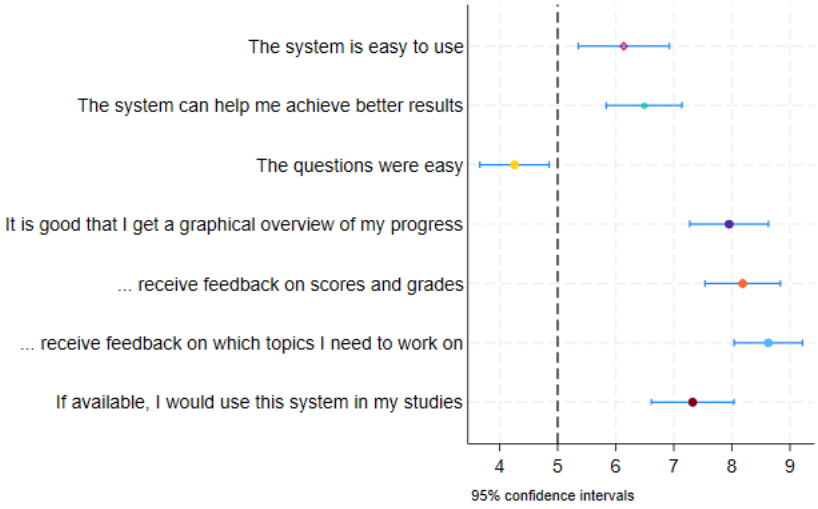


Fig. 2. Survey responses by the participants, on a scale from 1 to 10. The black dotted line marks 5, the point of indifference to the statement.

the adaptive test is initially too easy, it quickly progresses to more challenging material; if too difficult, it becomes easier, ensuring students still feel a sense of achievement. In contrast, linear tests expose students to the full range of questions, which can negatively impact confidence and self-efficacy, ultimately affecting exam performance.

The small sample size of 47 participants, with only 23 in the adaptive test and 24 in the linear test, limits the generalizability of our findings. The statistical power of our tests is relatively low, increasing the likelihood of Type II errors. Additionally, the exam consisted of only four questions, which may not capture the full spectrum of participants' learning outcomes. To address these limitations, future studies should increase the number of participants and include a more comprehensive set of exam questions to enhance statistical power and provide a more precise measurement of learning outcomes. Moreover, given that the testing regime can account for 21% of the variation in the exam scores, it would be interesting to see if there are heterogeneous effects from the treatment. For example, by having a common pre-test before randomly allocating the students into the treatment arms, and then estimating the treatment effect on the exam conditional on the pre-test score.

Green et al. [6] provide technical guidelines for evaluating CATs, focusing on aspects such as dimensionality, measurement error, item parameter estimation, and test validity. They emphasize the importance of a well-calibrated item pool. Our database of questions (item pool) are based on four cohorts of students, so we believe that the difficulty level is accurately measured.

Green et al. [6] also discusses the challenges of equating CAT with traditional paper-based tests and highlights the need for careful consideration of human factors, such as the testing environment and interface design, to ensure the reliability and fairness of adaptive testing systems. Our survey evidence suggest that the participants found our system user-friendly. Last, we suggest using the CAT for formative assessment rather than summative assessment, so the focus is on learning outcomes throughout the course rather than grading at a final exam to ensure fairness.

6 Implications for practice

We propose a novel approach to structuring a university-level mathematics course using adaptive testing and personalized learning pathways. For example, instead of delivering a single lecture to 180 students, our system can potentially allow for what is effectively 180 individualized lectures occurring simultaneously. This is not an online course; rather, we use technology within the classroom to tailor the learning experience to each student's needs.

We envision a gamified learning environment where students progress along individualized learning paths with adaptive difficulty levels. This approach can foster a more engaging and effective learning experience. The following outlines how a typical class session could be structured:

1. Introduction to concepts: The session begins with a brief lecture to the entire class, introducing key concepts and foundational theories. This ensures that all students start with the same basic understanding of the topic.
2. Initial practice exercises: After the brief lecture, students complete a set of practice exercises designed to reinforce the newly introduced concepts. These exercises are relatively straightforward and serves as a warm-up to get a feel of the material.
3. Adaptive testing session: Students then log into the CAT system, which tailors the difficulty of questions based on their performance in real-time. Feedback is provided after each question and at the end of the test, helping students identify their strengths and areas that need improvement.
4. Short videos on key concepts: After learning which areas to improve, the student could watch short videos specifically focused on the key concepts they have identified using the CAT.
5. Instructor's dashboard and group work: The instructor has access to a real-time dashboard displaying each student's progress and performance. This data allows the instructor to:
   (a) Group students by performance level: Form groups of students who are at a similar level to collaborate and support each other.
   (b) Use mixed-ability groups: Create groups with students at varying levels, encouraging peer teaching and collaborative learning.
6. Tailored lecture adjustments: The data collected from the CAT provides valuable insights into the students' comprehension and common areas of

difficulty. This allows the instructor to adjust the subsequent teaching material and lecture focus, ensuring that it addresses the specific needs of the students. By doing so, the instructor can provide targeted support to all students, thereby enhancing the overall learning experience.

We believe there are three key benefits to using a CAT learning system. First, personalized learning paths allow each student to progress at the appropriate level. Ling et al. [10] finds that personalized approach can reduce anxiety and increase engagement.

Second, the system's dashboards provide detailed overview of class performance, enabling instructors to make informed decisions about curriculum adjustments and targeted interventions. Alfageh et al. [1] showed that adaptive diagnostic assessments help teachers tailor instruction, improve student grouping, and enhance communication, benefiting elementary math education.

Third, by incorporating game-based elements and real-time feedback, students are more likely to be engaged and motivated to participate actively in the learning process. Ersozlu [5] found that while online learning increased anxiety during COVID-19, game-based learning and digital interventions reduced math anxiety and boosted engagement in primary students.

However, implementing this system requires further research and development, particularly in refining the adaptive algorithms to determine the optimal difficulty level, how to integrating the building blocks seamlessly into the classroom environment and the optimal course sequence structure and duration. Pilot studies should be conducted to evaluate the effectiveness of this approach and its impact on student performance and motivation.

7 Conclusion

We conducted a lab-based randomized control trial to test the difference between an adaptive test and a linear test. The participants who worked with the adaptive test scored 26.2 percentage points higher than the participants who worked with the linear test. The participants think the adaptive test was easy to use and that the adaptive test can help them get better academic results. The participants did not think the exercises was easy. To ensure robustness of the results, further research should have a larger sample size and more exam questions to improve statistical power and measurement precision of learning outcomes, and a pre-test to measure heterogeneous treatment effects.

Acknowledgements This research project is based on the master thesis Wasana Leithe submitted for the degree MSc in Applied Computer Science. Øyvind Aas was the advisor of the thesis. We thank seminar participants at Kristiania University College, UDIT 2023 and Hans Georg Schaathun for their comments. We thank Phillip Sampaio Smedsrud and Lene Haugsgjerd for excellent research assistance with the experiment.

References

1. Alfageh, D.H., York, C.S., Hodge-Zickerman, A., Xie, Y.: Elementary teachers' use of adaptive diagnostic assessment to improve mathematics teaching and learning: A case study. International Electronic Journal of Mathematics Education **19**(1), 1–15 (2024)
2. Chang, H.H.: Psychometrics behind computerized adaptive testing. Psychometrika **80**, 1–20 (2015)
3. Conejo, R., Guzmán, E., Trella, M.: The siette automatic assessment environment. International Journal of Artificial Intelligence in Education **26**, 270–292 (2016)
4. Elo, A.E.: The rating of chess players, past and present. Arco Pub. (1978)
5. Ersozlu, Z.: The role of technology in reducing mathematics anxiety in primary school students. Contemporary Educational Technology **16**(3, ep517), 1 – 11 (2024)
6. Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., Reckase, M.D.: Technical guidelines for assessing computerized adaptive tests. Journal of Educational measurement **21**(4), 347–360 (1984)
7. Harrison, P.M., Collins, T., Müllensiefen, D.: Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. Scientific Reports **7**(1), 3618:1–18 (2017)
8. Jansen, B.R., Louwerse, J., Straatemeier, M., Van der Ven, S.H., Klinkenberg, S., Van der Maas, H.L.: The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. Learning and individual differences **24**, 190–197 (2013)
9. Klinkenberg, S., Straatemeier, M., van der Maas, H.L.: Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. Computers & Education **57**(2), 1813–1824 (2011)
10. Ling, G., Attali, Y., Finn, B., Stone, E.A.: Is a computerized adaptive test more motivating than a fixed-item test? Applied psychological measurement **41**(7), 495–511 (2017)
11. McGlohen, M., Chang, H.H.: Combining computer adaptive testing technology with cognitively diagnostic assessment. Behavior research methods **40**, 808–821 (2008)
12. Roediger III, H.L., Putnam, A.L., Smith, M.A.: Chapter one - ten benefits of testing and their applications to educational practice. Psychology of learning and motivation **55**, 1–36 (2011)
13. Yang, A.C., Flanagan, B., Ogata, H.: Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning. Computers and Education: Artificial Intelligence **3(100104)**, 1–10 (2022)