Adaptive feedback from an ai tutor in higher mathematics: a markov model of competence acquisition and the role of feedback latency

ALIKHANOV OLIMJON

Nagangan State University, Namangan, Uzbekistan
e-mail: webmasteruzpre@gmail.com

Abstract. *Large language model tutors return feedback within seconds, whereas in a conventional higher mathematics course it arrives a week later. This paper asks how much of the measured advantage of AI tutoring is attributable to latency as such. A three-state Markov model of competence acquisition is developed in which the per-attempt acquisition probability is an explicit function of feedback delay, with a slip-and-guess observation model linking latent competence to observed performance. Over a semester of 45 practice attempts the model predicts a test score of 73.1 % under an AI tutor against 65.0 % under weekly feedback — a gain of 8.1 percentage points, d ≈ 0.54, consistent with meta-analyses of intelligent tutoring systems. The advantage has a half-life of about 41 hours in feedback delay: it is almost fully preserved within an hour, but a third is lost at 24 hours. Verification requires 54 students per group.*

Keywords: *artificial intelligence in education, AI tutor, large language models, adaptive feedback, feedback latency, knowledge tracing, Markov chain, higher mathematics, effect size.*

1. Introduction

Feedback is among the most powerful levers available to an instructor, but in a mass higher mathematics course it is also the scarcest. A lecturer with several hundred students cannot comment on individual solutions more than once a week, and by the time an annotated problem set is returned the cognitive state that produced the error is no longer accessible. Research on formative feedback holds that its value lies in when and how it is delivered, not in its mere presence [8, 9].

Intelligent tutoring systems were designed to relieve this bottleneck, and meta-analyses report effect sizes of roughly 0.4 to 0.7 against conventional instruction [3, 4, 5]. Large language model tutors have changed the economics again, since a competent conversational tutor can now be built for a single course at negligible marginal cost: a recent randomised trial with 194 undergraduates reported that a purpose-built AI tutor produced roughly twice the learning gains of an excellent active-learning class, in less time [6].

What such studies do not isolate is the mechanism. An AI tutor differs from a human instructor in three respects at once: it is available at any hour, it responds within seconds, and it adapts to the individual student. Because these are confounded in practice, it is not known how much of the benefit would survive if the system were equally adaptive but slower — a question with budgetary consequences, since latency and adaptivity are

purchased in very different ways.

The literature on feedback timing suggests the question is not idle: for procedural and applied learning, delayed feedback is markedly less effective than immediate feedback, though not without value [7]. If latency is a principal active ingredient, a tutoring desk answering within the hour should capture most of the benefit, while a sophisticated AI system whose replies are batched overnight should lose much of it.

The aim of the study is to construct a Markov model of competence acquisition in which feedback delay enters as an explicit parameter, to quantify how the expected learning outcome and the effect size depend on that delay, and to determine the sample size required to test the resulting predictions empirically.

2. Materials and methods

2.1. The configurations compared. Two organisations of the same higher mathematics module are compared. Under the conventional arrangement the student submits a weekly problem set and receives annotated feedback at the next class, so the delay is $\tau = 168$ hours. Under the AI-tutor arrangement the student works through the same problems in a dialogue with a course-specific large language model tutor that evaluates each step and responds within seconds, so $\tau \approx 0$. The number of problems, their content and the total study time are held equal; only the latency of feedback differs.

2.2. The competence model. Following the knowledge-tracing tradition [2], competence in a skill is represented by a latent variable with three states: N — not acquired, P — partially acquired, and M — mastered. The unit of time is one practice attempt k. Per attempt the student advances from N to P with probability p, and from P to M with probability $q = cp$, where $c < 1$ reflects the fact that consolidation is slower than first acquisition. Decay acts in the opposite direction: $M \to P$ with probability f and $P \to N$ with probability g. The transition matrix is therefore

$$\Pi = [[1-p,\ p,\ 0],\ [g,\ 1-g-q,\ q],\ [0,\ f,\ 1-f]]. \quad (1)$$

The chain is a birth-and-death process on three states, so its stationary distribution follows from detailed balance:

$$\pi_P/\pi_N = p/g, \quad \pi_M/\pi_P = q/f, \quad (2)$$

which already exposes the structure of the problem: the long-run share of mastered skills is governed by the ratio of acquisition to decay probabilities, and p is the only quantity in that ratio on which the instructional design acts.

2.3. Feedback latency. The acquisition probability is modelled as an exponentially decaying function of delay with a non-zero floor, since delayed feedback retains some value [7]:

$$p(\tau) = p_\infty + (p_0 - p_\infty)\cdot e^{-\tau/\tau_0}, \quad (3)$$

where p_0 is the per-attempt acquisition probability under immediate feedback, p_∞ the limiting probability under indefinitely delayed feedback, and τ_0 the characteristic decay time of feedback value.

2.4. Observation model. Latent competence is not observed directly; it is inferred from test performance. Following the standard formulation [2], the expected score is

$$E[S] = \pi_M(1 - s) + 0.5\cdot\pi_P + \pi_N\cdot\gamma, \quad (4)$$

where s is the slip probability (a mastered skill applied incorrectly) and γ the guess probability. The partially acquired state is scored at one half, reflecting partial credit for an incomplete solution.

2.5. Parameters and criteria. The following values were adopted: $p_0 = 0.28$, $p_\infty = 0.17$, $\tau_0 = 48$ h, $c = 0.6$, $f = 0.08$, $g = 0.06$, $s = 0.10$, $\gamma = 0.15$, and $K = 45$ practice attempts, corresponding to three problems per week over a fifteen-week semester. Comparative effectiveness was assessed by the absolute

difference Δ in expected score, by Cohen's effect size $d = \Delta/\sigma$ with $\sigma = 0.15$, and by the sample size required at $\alpha = 0.05$ with power 0.80 [10]:

$$n = 2(z_{1-\alpha/2} + z_{1-\beta})^2 / d^2. \quad (5)$$

The chain was propagated numerically from the initial state $\pi_0 = (1, 0, 0)$; the limiting values agreed with the analytical stationary distribution obtained from (2) to four decimal places.

3. Results

Equation (3) gives $p = 0.280$ for the AI tutor and $p = 0.173$ for weekly feedback, a ratio of 1.62. The resulting trajectories of expected test score are shown in Table 1 and in Figure 1a.

Table 1. Modelled expected test score as a function of practice attempts

Practice attempts, *k*	0	9	18	27	36	45
Conventional, score %	15.0	49.8	60.5	63.7	64.7	65.0
AI tutor, score %	15.0	63.1	71.5	72.8	73.0	73.1
Difference Δ, pp	0.0	13.3	11.0	9.1	8.3	8.1
Relative gain, %	—	26.7	18.2	14.3	12.8	12.5

Note: computed from (1)–(4) with the parameters of Section 2.5.

At the end of the semester the expected score is 65.0 % under conventional feedback and 73.1 % under the AI tutor. The absolute difference is $\Delta = 8.1$ percentage points, the relative gain 12.5 %, and the effect size $d = 0.54$, a medium effect in Cohen's classification [10]. From (5), confirming an effect of this size requires at least 54 students per group, or 108 in total. The corresponding probabilities of full mastery are $\pi_M = 0.489$ and $\pi_M = 0.633$.

The efficiency gain exceeds the attainment gain. As Figure 1a shows, the AI-tutored student reaches at attempt 11 the score that conventional instruction achieves only after all 45 attempts — roughly three quarters less practice for the same outcome, because raising p shifts both the level and the time constant of the trajectory.

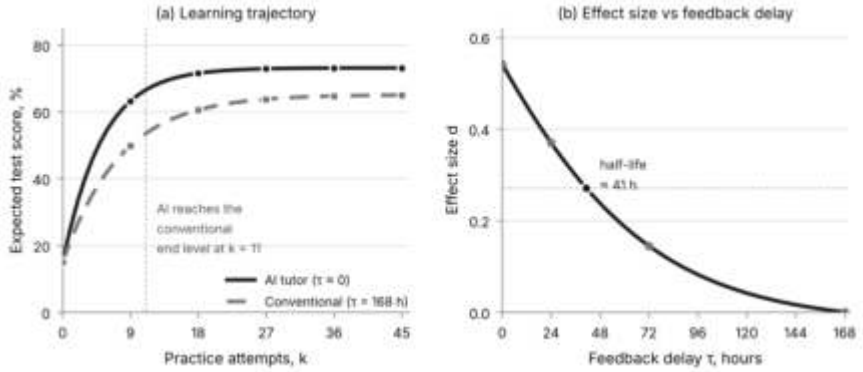


Figure 1. Learning trajectory (a) and effect size versus feedback delay (b)

The central result concerns the shape of the dependence on latency, shown in Figure 1b and quantified in Table 2. The advantage is not a step function of "AI versus no AI": it decays smoothly with delay, with a half-life of approximately 41 hours. Feedback delivered within one hour retains 98 % of the full advantage, whereas a delay of 24 hours — a tutor that replies the next morning — preserves only 68 % of it, and a delay of three days preserves 27 %.

Table 2. Predicted effect as a function of feedback delay

Feedback delay τ	$p(\tau)$	Score, %	Δ, pp	Effect size d	n per group
≈ 1 min (AI tutor)	0.280	73.1	8.1	0.54	54
1 hour	0.278	73.0	8.0	0.53	56
24 hours	0.237	70.5	5.5	0.37	115
72 hours	0.195	67.1	2.2	0.14	753
168 hours (baseline)	0.173	65.0	0.0	0.00	—

Note: Δ, d and n are computed relative to the 168-hour baseline (score 65.0 %).

4. Discussion

The predicted effect size of $d \approx 0.54$ sits inside the range established for intelligent tutoring systems: $g = 0.43$ against conventional instruction [4], a median of about 0.66 [5], and approximately 0.76 for step-based systems against no tutoring [3]. The model is thus calibrated to this body of evidence rather than to the more striking results of recent large language model trials [6].

The substantive contribution is the latency curve. It yields a prediction that is sharp enough to be refuted: an AI tutor whose feedback is deliberately batched to the following day should retain roughly two thirds of its advantage, while a human tutoring desk that answers within the hour should capture almost all of it. If this holds, the operative variable in AI tutoring is response latency rather than the sophistication of the underlying model, and institutions with limited budgets should prioritise synchronous availability over model capability.

The curve also offers a reading of the heterogeneity in the empirical literature: deploying an AI tutor inside a synchronous practice session and offering the same system as an out-of-hours resource are configurations that this model places at different points of a steeply declining curve, although both are reported under the same label.

A further implication concerns measurement. Because the difference narrows as practice accumulates — from 13.3 percentage points at attempt 9 to 8.1 at attempt 45 — the apparent effect of an intervention depends on when the post-test is administered. A study that tests early will report a substantially larger effect than one that tests at the end of the course, with no difference in the underlying mechanism.

Limitations. The exponential form of (3) and the value of τ_0 are assumptions rather than measurements; $\tau_0 = 24$ h raises the predicted effect size to 0.57 and $\tau_0 = 96$ h lowers it to 0.43, so the qualitative conclusion is robust while the precise half-life is not. The model also treats skills as independent, ignores differences in prior preparation, and assumes that AI feedback equals instructor feedback in quality — an assumption that fails where the tutor is confidently wrong, a failure mode with no counterpart in the delayed condition. Finally, only cognitive outcomes are represented, not the motivational effects reported in [6].

5. Conclusions

1. A three-state Markov model of competence acquisition was developed in which feedback delay enters explicitly through the per-attempt acquisition probability, and observed performance is linked to latent competence by a slip-and-guess observation model.

2. Over a semester of 45 practice attempts an AI tutor is predicted to raise the expected test score from 65.0 % to 73.1 %, a gain of 8.1 percentage points with an effect size of $d \approx 0.54$,

consistent with meta-analytic evidence on intelligent tutoring systems.

3. The efficiency gain exceeds the attainment gain: the end-of-semester level of conventional instruction is reached after 11 of 45 practice attempts.

4. The advantage decays with feedback delay with a half-life of about 41 hours; it is almost fully preserved within one hour but a third is lost at 24 hours, which identifies latency as a principal active ingredient of AI tutoring.

5. Because the difference narrows with accumulated practice, the timing of the post-test materially affects the measured effect, which may explain part of the published heterogeneity.

6. Testing these predictions requires at least 54 students per group. The decisive experiment is a three-arm trial in which the same tutor is delivered at deliberately different latencies.

References

1. Decree of the President of the Republic of Uzbekistan No. PF-5847 of 8 October 2019 “On approval of the Concept for the Development of the Higher Education System until 2030”. – Tashkent: Lex.uz, 2019.
2. Corbett A.T., Anderson J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge // User Modeling and User-Adapted Interaction. – 1995. – Vol. 4, No. 4. – P. 253–278.
3. VanLehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems // Educational Psychologist. – 2011. – Vol. 46, No. 4. – P. 197–221.
4. Ma W., Adesope O.O., Nesbit J.C., Liu Q. Intelligent tutoring systems and learning outcomes: A meta-analysis // Journal of Educational Psychology. – 2014. – Vol. 106, No. 4. – P. 901–918.
5. Kulik J.A., Fletcher J.D. Effectiveness of intelligent tutoring systems: A meta-analytic review // Review of Educational Research. – 2016. – Vol. 86, No. 1. – P. 42–78.
6. Kestin G., Miller K., Klales A., Milbourne T., Ponti G. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting // Scientific Reports. – 2025. – Vol. 15. – Art. 97652.
7. Kulik J.A., Kulik C.C. Timing of feedback and verbal learning // Review of Educational Research. – 1988. – Vol. 58, No. 1. – P. 79–97.
8. Shute V.J. Focus on formative feedback // Review of Educational Research. – 2008. – Vol. 78, No. 1. – P. 153–189.
9. Hattie J., Timperley H. The power of feedback // Review of Educational Research. – 2007. – Vol. 77, No. 1. – P. 81–112.
10. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. – Hillsdale: Lawrence Erlbaum Associates, 1988. – 567 p.

Mechanisms for improving the intellectual competence of students with disabilities based on an inclusive approach .....Ошибка! Закладка не определена.

Solikhudjaeva Dildora Abduraimovna.....3