CAN THE SIMULATED STUDENT LEARN? FITTED KNOWLEDGE-TRACING MODELS DISAGREE ON SPACING, RECENCY AND TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Tutoring agents are increasingly scored against a simulated student. In benchmarks that track what the learner knows, the simulated student is a knowledge-tracing model fitted to interaction logs. The score is the change in the predicted probability of a correct response after teaching. It measures learning gain only if the model responds to teaching as a learner does. This paper introduces four contrasts that test this condition on five architectures fitted to two corpora at three seeds each. The contrasts are spacing against a position-matched control, recency, transfer from a related concept and the error penalty. Each contrast appends two sequences to the same learner history. They differ in only one systematic respect: the order of practice, the concept of the items practised first or whether the first response is correct. The effect is the paired difference in predicted gain on held-out items of the target concept, averaged across histories. A contrast passes when its effect has the expected sign at $p < 0.01$ in a two-sided paired t-test. That sign is positive for the three schedule contrasts and negative for the error penalty. A failure is an absence if the effect is not significant and an inversion if it is significant the other way. Responses are written in rather than sampled, because they move the probe score a median of 10 times more than the concept practised does. Every simulator penalises a wrong answer at every seed. Spacing and recency each fail in six of 10 simulators at the first seed, and transfer fails in eight. Across the three seeds, some absences and one small inversion persist, but the largest inversions vanish. Only five of 10 simulators give the same verdicts at every seed, and new target concepts change verdicts at least as often as a refit. DKVMN fails spacing at every seed and target concept. Validating DKT on ASSIST2017 alone finds no failure at the first seed. When the simulator answers, as in a benchmark, a spaced tutor beats a massed tutor on 16 of 30 fitted models, ties with it on 12 and loses on two. The fitted simulator that a benchmark adopts therefore limits which tested phenomena the benchmark can credit. A four-contrast certification costs minutes and reports which of the four a fitted simulator registers.

1 INTRODUCTION

A tutoring agent has to be evaluated against a learner, but recruiting one learner per condition does not scale. The field therefore substitutes a simulator, either a language-model agent that role-plays the learner (Gao et al., 2025; Xu et al., 2025) or a knowledge-tracing model. In benchmarks that track what the learner knows, the knowledge-tracing model is fitted to recorded interaction data, holds the learner state and answers the exercises the agent selects (Lee et al., 2026). The change in the predicted probability of a correct response on a held-out probe is reported as learning gain.

That protocol reads a conditional prediction as a causal response to an intervention. A knowledge-tracing model is fitted to predict the next response in sequences that real students and real tutors produce together. The fitting objective does not require the model to respond as a student does to a sequence that a new tutor constructs. Whether the model does is an empirical question that precedes every other question about such benchmarks. A scorer that does not register spacing gives no basis for crediting a tutor for spacing practice, however carefully the rest of the benchmark is built.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This paper asks the empirical question by construction. Each of four phenomena becomes a contrast between two sequences appended to the same learner history: spacing against a position-matched control, recency, transfer from a related concept and the error penalty. The two sequences of a contrast differ in only one systematic respect: the order of the practice, the concept of the items practised first or whether the first response is correct. Human learners show spacing, recency and transfer (Trumble et al., 2024; Nickl & Bäuml, 2023; Fraundorf et al., 2023), and any model that reads a wrong answer as evidence of less knowledge must show the error penalty (Corbett & Anderson, 1995). Because the responses are written in rather than sampled, a contrast measures how the model responds to the schedule itself rather than a difference in how the simulated learner answers.

The paper makes four contributions. The first is a **model-agnostic construct-validity suite for simulated learners**, four contrasts that isolate the response of a fitted model to order and content from its response to the answers shown to it. The isolation matters because the response channel is larger than the content channel in every simulator, by a median factor of 10 (Section 3.2). The second is **a measurement across architectures, seeds, item draws and target concepts** (Figure 1). Across 10 simulators the error penalty passes in every fit at every item draw. At the first seed, spacing and recency each fail in six of 10, transfer fails in eight, and one simulator passes every contrast. Five absences and one small inversion hold at every seed, whereas the largest inversions vanish. Of these six, five fail again at every seed in two further draws of the items (Appendix H). New target concepts change verdicts at least as often as a refit (Appendix J). Only five of 10 simulators give the same verdicts at every seed, and 115 of the 120 verdicts hold with learners as the unit (Appendix E). The third is **a direct demonstration of the blind spot in single-simulator validation**. In isolation, DKT / ASSIST2017 shows every phenomenon measured here in the human direction, whereas each of the other nine simulators fails at least one contrast. The fourth is **evidence that leaderboard outcomes depend on the fitted simulator**. A spaced and a massed tutor teach from shared item pools and are scored while the simulator answers. The spaced tutor wins on 16 of 30 fitted models, neither wins on 12, and the massed tutor wins on two. With learners as the unit, all 30 comparisons hold. DKVMN cannot separate the two tutors at any seed or draw of items or targets. The controlled spacing verdict takes minutes and anticipates the benchmark outcome in 22 of the 30.

The simulated student that a benchmark scores against may not respond to a schedule at all. What the benchmark can reward is a property of the fitted simulator and has to be measured, not assumed.

2 Related work

Simulated learners and their validity. Simulated students serve teacher training, hypothesis testing and the evaluation of adaptive systems, but a recent survey identifies validity and evaluation as the open problems (Yuan et al., 2026). Recent simulators built on language models (Gao et al., 2025; Xu et al., 2025; Duan et al., 2026; Liu et al., 2025) come with frameworks for judging them (Moreau-Pernet, 2026). The same survey argues that a simulator used to make claims about learning trajectories needs an explicit student state that evolves with instruction (Yuan et al., 2026). More broadly, simulated users for agent evaluation are known to be unreliable proxies (Seshadri et al., 2026), to be more cooperative than people (Chopra et al., 2026) and to leave a measurable gap to reality (Zhou et al., 2026). All of that work asks whether a simulator resembles a person. This paper asks a narrower question that can be answered without a person: whether a fitted model responds to a schedule in the direction human learners do.

Knowledge tracing as the simulator. The models used as simulated students are the standard knowledge-tracing architectures: two-state per-skill models (Corbett & Anderson, 1995), recurrent (Piech et al., 2015), memory-augmented (Zhang et al., 2017) and attentive (Pandey & Karypis, 2019; Ghosh et al., 2020; Liu et al., 2023) successors, and recent variants that read language-model representations (Norris et al., 2025), make difficulty explicit (Huang et al., 2026), model affect (Sun et al., 2025) or model diverse learning states (Wang et al., 2025). Two lines of work are close to this study. Work on responsible knowledge tracing quantifies prediction uncertainty (Cheng et al., 2025), decides when to defer the prediction to a teacher (Mitton et al., 2026) and builds in educational structure (Hooshyar et al., 2026). This paper asks the complementary question of what a fitted model already does when used as a student. Work on mastery learning uses a fitted model to decide when practice has become redundant (Xia et al., 2025), whereas this paper measures whether the model responds to practice at all.

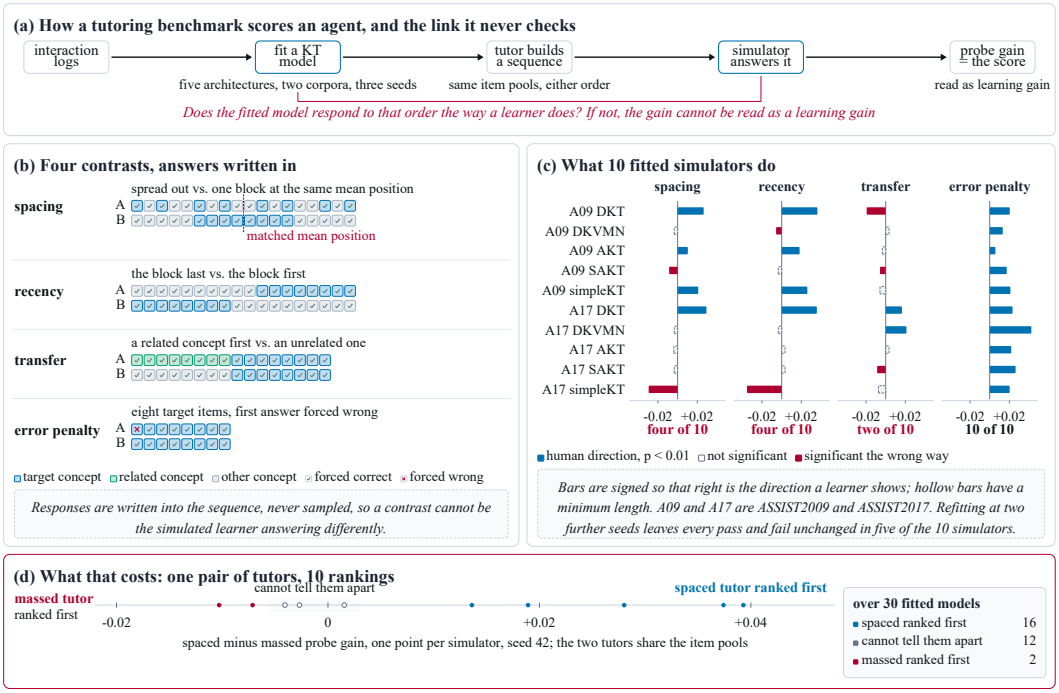

Figure 1: Certifying a simulated student. (a) A tutoring benchmark fits a knowledge-tracing model to interaction logs, has a tutor build a sequence, lets the fitted model answer it and reads the change in probe score as learning gain. The link this paper measures is whether the fitted model responds to the order of that sequence as a learner does. (b) Each contrast writes the responses in and varies one thing systematically: the order, the items practised first or one response. The massed control sits at the mean position of the spaced practice, so recency is matched on average. (c) The recorded effect of each contrast on 10 fitted simulators at seed 42, on one shared scale. A bar to the right marks the direction a learner shows. The error penalty passes in all 10 cells, whereas spacing and recency pass in four each and transfer passes in two. (d) In the benchmark setting the simulator generates its own responses, and different simulators rank the same two tutors differently. The spacing verdict in (c) anticipates the ranking in seven of the 10 simulators.

Evaluation protocols for tutoring agents. A benchmark such as EduClaw-Bench (Lee et al., 2026) couples a tutoring agent to a knowledge-tracing simulator and reports one leaderboard from one fitted model. Doubts about such a leaderboard usually concern noise in its estimates or bias in the scorer, and both doubts take the responses of the simulator as given. This paper asks the prior question of whether the simulator can register the behaviours being ranked at all. Fitted models are usually compared by next-response AUC, which leaves this question open.

3 Protocol

3.1 Setting and notation

A fitted knowledge-tracing model s maps a history x of item–response pairs to the predicted probability $p_s(q \mid x)$ of a correct response to any item q. Each learner record in the test split is cut into segments of at most 200 interactions, and the history x_u is the first $t_0 = 50$ interactions of segment u. The concept of an item is its first concept identifier, and a probe Q_c of eight held-out items measures the state of a target concept c. The probe score after a sequence y is

$$P_s(y) = \frac{1}{|Q_c|} \sum_{q \in Q_c} p_s(q \mid x_u \oplus y), \tag{1}$$

with each probe item appended on its own, so no probe measurement conditions another. The gain of y is $P_s(y) - P_s(\varnothing)$. Because a contrast subtracts one gain from another, the shared baseline probe

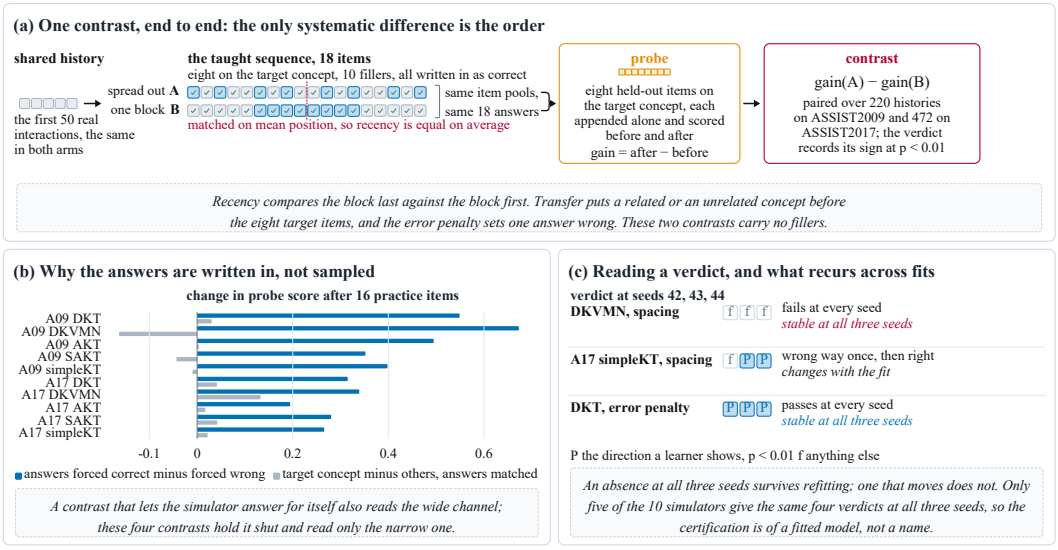

Figure 2: How one contrast is built and why the responses are written in. (a) After a shared history of 50 real interactions, a spaced and a massed sequence each draw 18 items from the same pools by the same procedure. The massed block sits at the mean position of the spaced practice, and every answer is written in as correct. Eight held-out items yield the probe score before and after each arrangement, and the contrast is the paired difference in gain. (b) The two channels a probe gain can come from, on the 10 simulators. Forcing the answers correct rather than wrong moves the gain a median of 10 times more than attaching the same answers to other concepts. (c) At each of three seeds, a verdict records the sign of the contrast at $p < 0.01$. An absence at all three seeds recurs across the tested fits, and one that changes across seeds is specific to a fit.

score $P_s(\varnothing)$ cancels in the paired difference. A tutoring benchmark reports exactly the gain of y, where the agent chooses y and s itself generates the responses inside y.

3.2 Holding the responses fixed

In the benchmark setting the simulator generates the learner responses. A sequence of practice is therefore also a transcript of answers, and the two are inseparable after the fact but separable by construction. Every contrast below writes the responses into y rather than sampling them, so that the only systematic difference between its two sequences is the order of practice, the items practised first or the single response that the contrast varies.

Section 4.4 reports what fixing the responses removes. With 16 practice items and every response forced correct, the probe score rises by 0.10 to 0.34 across the 10 simulators. With every response forced wrong, the same items move it by -0.56 to $+0.04$. The answers alone thus open a gap of 0.19 to 0.67. With responses matched and the effect averaged over both forcings, practising the target concept rather than other concepts is worth -0.16 to $+0.13$. The response channel exceeds the content channel by a median factor of 10. A contrast that lets the simulator sample its answers measures the response channel, whatever else it measures.

3.3 Four contrasts

Each contrast is built from $k = 8$ practice items on the target concept, f filler items from other concepts and the probe. The two sequences vary one factor systematically and otherwise draw their items at random from the same pools. Each gain is averaged over two samples. Responses are written in. Every response is correct except the first response of the error-penalty contrast, so the expected difference between the two sequences reflects only what the contrast varies. Figure 2 shows the spacing contrast. Figure 1(b) shows all four, and only spacing and recency carry fillers. The suite reads nothing but predicted probabilities on the probe, so it runs on any fitted knowledge-

tracing model with concept-labelled items. Correctly answered practice raises the probe score in all 10 simulators (Appendix C), so the only question is whether its arrangement does.

Spacing. The massed arrangement places the eight practice items contiguously, whereas the spaced arrangement distributes them evenly through the fillers. Because spaced practice also lies nearer the probe, the massed control is placed at the mean position of the spaced practice. Spacing is read as spaced minus that position-matched control. If spacing is read against a massed block at the start instead, the effect is confounded with recency. Appendix B compares the two readings.

Recency. The contrast is practice last minus practice first, with the block of eight practice items placed after or before the f fillers.

Transfer. The contrast is eight items of a related concept minus eight items of an unrelated concept, placed before the target practice. With no prerequisite graph available, the related concept is the one that co-occurs most often with the target concept in the learner histories, and the unrelated concept is the one that co-occurs least. A batch with no such pair is skipped.

Error penalty. The contrast is eight target items with the first response forced wrong, minus eight target items with every response forced correct.

The three schedule effects are above zero in human studies, and the error penalty is below zero in any model that reads a wrong answer as evidence of less knowledge. A simulator passes a contrast when the paired effect over histories has the expected sign at an uncorrected $p < 0.01$. It fails by absence when the effect is not significant and by inversion when the effect is significant the other way.

3.4 Scale

Five architectures (DKT, DKVMN, AKT, SAKT and simpleKT) are fitted to ASSIST2009 and ASSIST2017 at seed 42 and refitted at seeds 43 and 44 for Section 4.2. The third corpus used elsewhere in this literature is NIPS34, which assigns every item the same first concept identifier. Because the protocol reads that identifier as the concept of an item, NIPS34 cannot support a concept-based contrast and is excluded rather than reported as empty. Every contrast is paired over the same histories, 220 from 167 learners on ASSIST2009 and 472 from 162 learners on ASSIST2017. With learners rather than histories as the unit, 115 of the 120 verdicts are unchanged (Appendix E). Spacing and recency use $f = 10$ fillers. All four contrasts are fixed in writing before the sweep.

4 Results

4.1 The error penalty is robust; the schedule effects are not

Figure 3 and Table 1 give the 10 cells. The error penalty passes in 10 of 10 simulators, with effects from -0.042 to -0.006. Spacing and recency each pass in four of 10 simulators, and transfer passes in two of 10. Only DKT on ASSIST2017 passes all four contrasts.

Most failures are absences. The spacing of DKVMN is absent on both corpora, the recency of SAKT is absent on both, and all three schedule effects of AKT are absent on ASSIST2017. Across the three seeds, each individual 99% interval of an absence excludes the pooled median effect of the fits that pass the same contrast (Appendix E). Seven effects are significant the wrong way. Three are large: simpleKT on ASSIST2017 has spacing at -0.029 and recency at -0.035, and DKT on ASSIST2009 has transfer at -0.019. The other four are small but significant at these sample sizes. DKVMN on ASSIST2009 has recency at -0.006, SAKT on ASSIST2009 has spacing at -0.009 and transfer at -0.006, and SAKT on ASSIST2017 has transfer at -0.009. Section 4.2 refits every simulator at two further seeds to see which absences and inversions recur across fits.

4.2 Which verdicts recur across fits

Every simulator is refitted at two further seeds and the four contrasts are rerun. Table 5 marks every contrast at each seed as a pass or a fail. Five of 10 simulators give the same verdicts at all three

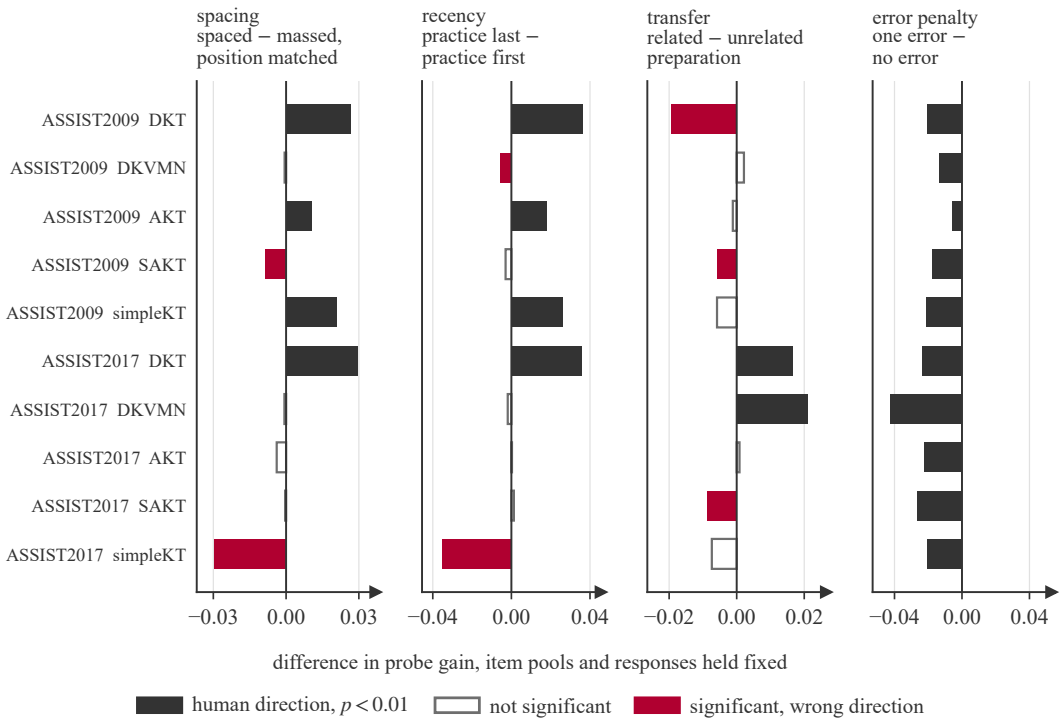

Figure 3: Four learning phenomena, 10 fitted simulators. Each bar is the difference in probe gain between two sequences whose only systematic difference is the order, the items practised first or one response. Effects are at seed 42. Filled bars have the human direction at $p < 0.01$, hollow bars are not significant, and accented bars are significant the wrong way at this seed. Table 6 shows that only the recency of DKVMN on ASSIST2009 is an inversion at all three seeds. The error penalty passes everywhere. Of the three schedule effects, spacing and recency pass in four of 10 cells, and transfer passes in two of 10. Spacing and recency each have two inversions, and transfer has three. The other failures are absences.

Table 1: The four contrasts on 10 simulators at seed 42, with the responses written in. A starred effect has the human direction at $p < 0.01$. The Fail column lists contrasts that are absent (*a*) or significant the wrong way at this seed (*i*). Unmarked failures are absences. Table 6 shows that only the recency of DKVMN on ASSIST2009 is inverted at all three seeds.

Corpus	Model	Spacing	Recency	Transfer	Error	Fail
ASSIST2009	DKT	$+0.027^*$	$+0.036^*$	-0.019	-0.020^*	transfer (i)
ASSIST2009	DKVMN	-0.001	-0.006	$+0.002$	-0.013^*	all three; recency (i)
ASSIST2009	AKT	$+0.010^*$	$+0.018^*$	-0.001	-0.006^*	transfer (a)
ASSIST2009	SAKT	-0.009	-0.003	-0.006	-0.017^*	all three; two (i)
ASSIST2009	simpleKT	$+0.021^*$	$+0.026^*$	-0.006	-0.021^*	transfer (a)
ASSIST2017	DKT	$+0.029^*$	$+0.036^*$	$+0.017^*$	-0.023^*	none
ASSIST2017	DKVMN	-0.001	-0.002	$+0.021^*$	-0.042^*	spacing, recency
ASSIST2017	AKT	-0.004	$+0.000$	$+0.001$	-0.022^*	all three
ASSIST2017	SAKT	-0.000	$+0.001$	-0.009	-0.026^*	all three; transfer (i)
ASSIST2017	simpleKT	-0.029	-0.035	-0.007	-0.020^*	all three; two (i)

seeds. The error penalty passes in 30 of 30 fitted models. Spacing and recency pass in 14 each, transfer passes in six, and the three schedule contrasts are inverted in 19 of the 90 measurements. A Bonferroni correction over the 120 measurements keeps 27 of the 30 error-penalty passes and 11 of the 19 inversions (Appendix E). DKVMN never registers spacing on either corpus at any seed, nor recency on ASSIST2017. On ASSIST2009 its recency is inverted at every seed, and each inversion is small but significant. SAKT on ASSIST2009 registers none of the three in the human direction at

Table 2: A spaced tutor against a massed tutor on the same item pools at seed 42, in the benchmark setting where the simulator generates learner responses. Difference is spaced minus massed probe gain before rounding. Winner is decided at $p < 0.01$. Over all three seeds the tally is spaced 16, neither 12, massed two. Bold marks the two massed wins, both at seed 42.

Corpus	Model	n	Massed	Spaced	Difference	Winner
ASSIST2009	DKT	220	-0.022	$+0.016$	$+0.037$	spaced
ASSIST2009	DKVMN	220	-0.006	-0.004	$+0.002$	neither
ASSIST2009	AKT	220	$+0.002$	$+0.030$	$+0.028$	spaced
ASSIST2009	SAKT	220	$+0.017$	$+0.013$	-0.004	neither
ASSIST2009	simpleKT	220	$+0.010$	$+0.029$	$+0.019$	spaced
ASSIST2017	DKT	472	-0.005	$+0.034$	$+0.039$	spaced
ASSIST2017	DKVMN	472	$+0.004$	$+0.001$	-0.003	neither
ASSIST2017	AKT	472	$+0.021$	$+0.034$	$+0.014$	spaced
ASSIST2017	SAKT	472	$+0.035$	$+0.028$	-0.007	**massed**
ASSIST2017	simpleKT	472	$+0.009$	-0.001	-0.010	**massed**

any seed. Those failures recur in all three fits. By contrast, the two largest inversions at the first seed are the spacing and recency of simpleKT on ASSIST2017, and both take the human direction at the two further seeds. Of the 12 simulator–contrast pairs inverted at some seed, five are inverted at two seeds and six are inverted at one. Only the recency of DKVMN on ASSIST2009 is inverted at all three. Every simulator on ASSIST2017 changes at least one verdict with the seed, whereas none of the five on ASSIST2009 does. On ASSIST2017, whether a fitted model registers a schedule effect depends on the fit.

Stable failures are not confined to one architecture. Over both corpora and three seeds, DKT, simpleKT, AKT, SAKT and DKVMN pass 12, 10, eight, three and one of their 18 schedule verdicts. Of the 14 simulator–contrast pairs that fail at every seed, the three attention models have eight, DKVMN has five and DKT has one. Of the five absences that hold at every seed, DKVMN has three. The other two are the recency of SAKT on ASSIST2009 and the spacing of AKT on ASSIST2017. Two further draws of the items change 12 of the 120 verdicts, five of them in transfer (Appendix H). Of the five stable absences and the inverted recency of DKVMN on ASSIST2009, all but the recency of SAKT fail at every seed in both draws. AKT shows no recency on ASSIST2017, although its architecture includes an explicit exponential decay intended to model forgetting (Ghosh et al., 2020). A built-in mechanism does not guarantee that the fitted model uses it.

The corpus matters too. At the first seed, DKT passes every contrast on ASSIST2017 and inverts transfer on ASSIST2009. On ASSIST2009, simpleKT and AKT reproduce spacing and recency. On ASSIST2017 at that seed, simpleKT inverts both and AKT shows neither. Over the three seeds, transfer passes in six fits on ASSIST2017 and in none on ASSIST2009. A simulator is a model fitted to a corpus, and its behaviour is a property of the pair. Fit quality does not reveal any of these verdicts. Refitting moves test AUC by at most 0.006 in any simulator (Appendix A) but changes at least one verdict in five of 10. Only one of the 30 fitted models passes all four contrasts.

Each batch of a run takes as its target the concept with the most distinct items, which gives four target concepts on ASSIST2009 and two on ASSIST2017 (Appendix I). Two further runs draw the target of each batch at random and bring the concepts tested to 29 (Appendix J). New target concepts change a pass or fail verdict at least as often as a refit, in 72 of 360 comparisons against 20 of 120. With drawn targets, recency and transfer pass more often and fewer effects invert. No target concept inverts the error penalty, and DKVMN fails spacing for every target concept at every seed.

4.3 Which tutor wins depends on the simulator

The contrasts above write the responses in, whereas a benchmark lets the simulator generate them. To see what the failures cost in the benchmark setting, this paper scores two tutors against every simulator, as on a leaderboard. A massed tutor teaches eight target items before 10 fillers, and a spaced tutor spreads them evenly among the fillers. Both draw their items from the same pools. Human evidence on spacing and on recency puts the spaced tutor ahead (Trumble et al., 2024; Nickl & Bäuml, 2023). Table 2 reports which simulators agree with it.

At seed 42, five of 10 simulators place the spaced tutor ahead at $p < 0.01$, three cannot separate the two tutors, and two (SAKT and simpleKT on ASSIST2017) place the massed tutor ahead. Refitting at two further seeds gives 30 fitted models. The spaced tutor wins on 16, neither tutor wins on 12, and the massed tutor wins on two. Both massed wins occur at the first seed only. At the other two seeds those simulators either put the spaced tutor ahead or cannot separate the two tutors, so the massed wins are properties of one fit. In contrast, the six DKVMN ties are stable. DKVMN cannot separate the two tutors on either corpus at any of the three seeds, nor in two further draws of the items (Appendix H). It is the one architecture whose failure of spacing and recency in Table 5 recurs in every fit. The 12 fitted models that cannot separate the two tutors include DKVMN at every seed, and no benchmark built on one of them credits the spaced tutor. Eight of these 12 show a spacing absence in Table 6. In 22 of the 30 fits, a spacing pass, absence or inversion anticipates a spaced win, a tie or a massed win in turn. At each of the two further draws of the items, the spaced tutor wins on 14 fitted models, and the massed tutor wins on two. The rule then holds in 26 and 24 fits. With the target concept drawn at random, the spaced tutor wins on 18 and 16 fitted models, and the massed tutor wins on one and two. DKVMN still cannot separate the two tutors under either target draw (Appendix J).

4.4 The response channel, measured

The contrasts hold responses fixed because sampling them measures something else. With 16 practice items, forcing every response correct rather than wrong moves the probe score by 0.19 to 0.67 depending on the simulator. With responses fixed, changing only whether the items belong to the target concept moves the probe score by -0.16 to $+0.13$, averaged over the two forcings. Across the 10 simulators the response channel exceeds the content channel in every cell, by a median factor of 10. The ratio runs from 2.6 on DKVMN / ASSIST2017 to 124.1 on AKT / ASSIST2009, whose content channel is $+0.004$. Every schedule effect in Table 1 is smaller than 0.07. Hence the response channel is large enough to dominate the verdict of a contrast that samples its responses.

A validity study of DKT / ASSIST2017 alone misses the failures reported here. On that simulator at seed 42, all six phenomena measured here take the human direction, namely the four contrasts, the dose–response of Appendix C and the content channel of Appendix F. Yet DKT / ASSIST2017 is the only cell in Table 1 that passes every contrast, and it loses spacing at the second seed. A single-simulator validity study reports the validity of one fitted model.

5 What follows for benchmark practice

A benchmark that scores against one fitted simulator is scoring against one row of Table 1. Six of 10 rows give no certified basis for crediting a tutor for spacing its practice, and eight give none for ordering related concepts. Some rows penalise the tutor for either. None of these failures is visible in the benchmark outputs, since a fitted model produces a probe score whether or not it registers the schedule.

The remedy is a certification rather than a fix. The four contrasts take between 16 and 202 seconds on a fitted model, with a median of 25 seconds. They identify which tutor behaviours the simulator can register. Before adopting a simulator, a benchmark should run the contrasts over several target concepts and draws of the items. It should report the four verdicts, with their effects and p-values, beside the leaderboard. A simulator that fails spacing should not be used to compare schedulers that differ in how they space practice. Because refitting changes a verdict in five of 10 simulators, the certificate covers the fitted model that a benchmark ships, not the architecture.

6 Limitations

The contrasts test direction against human learners but not magnitude. They vary the order of practice rather than elapsed time, since the fitted models read no timestamps. A simulator that reproduces spacing at $+0.02$ is not thereby shown to reproduce it at the magnitude human learners show. Because no prerequisite graph is available for either corpus, the related concept is read from co-occurrence, which is a proxy. Two corpora, five architectures and 29 target concepts do not exhaust the space of corpora, architectures or concepts. The grid of the main text rests on six of these con-

cepts, one of which covers 415 of the 472 ASSIST2017 histories. The third common corpus is unusable under the standard preprocessing. Three seeds per simulator distinguish recurrence across the tested fits from a result specific to one fit, but do not estimate how often a given architecture registers a given phenomenon. On ASSIST2017, every simulator changes at least one verdict with the seed. The right number of seeds, item draws and target concepts for a certification is an open question. Every effect is measured on a fitted model, and nothing here says what a person learns from the same schedule. A simulator that registers all four contrasts can still rank tutors wrongly for reasons outside them. The contrasts are a floor rather than a full validity battery.

7 Conclusion

At the concepts with the most items, a knowledge-tracing model used as a simulated student penalises a wrong answer in all 30 fits tested, but 11 of the 30 pass no schedule contrast. Across 10 simulators the phenomena a tutor controls through scheduling are not reproduced in most and are inverted in some. Which phenomena a simulator shows differs between architectures, between fits of one architecture and between target concepts, and the fit matters most on ASSIST2017. The fitted simulator that a benchmark adopts therefore limits which of these phenomena the benchmark can credit. Learner-level tests leave 115 of the 120 verdicts unchanged. No target concept inverts the error penalty, and DKVMN fails spacing at every seed, item draw and target concept. Five of the six failures that hold at every seed also fail at every seed in two further draws of the items. Identifying them takes minutes.

AI use statement

Generative AI tools provide methodological feedback, and language-model reviews of the manuscript inform revisions to the claims. The tools also implement the code for the experiments and analyses, help interpret the results and create the figures, find related work, and draft and edit the manuscript. Generative AI is not used to translate, develop the conceptual framework, propose or refine the hypotheses, design the contrasts, or clean, reformat, generate, alter or select any data. The paper states no theorem and analyses no qualitative data. All AI-assisted work is reviewed by the authors, every cited work is checked against its published record, and the main results are recomputed from the recorded outputs by a verification script. The authors take responsibility for the text, claims and artefacts of this work.

Ethics statement

The experiments use public knowledge-tracing corpora released for research and carrying no personally identifying information. No new data are collected from human subjects, and no human subject interacts with any system in this study.

Reproducibility statement

Appendices A and B give the fitting recipe, the seeds and the construction of every contrast, and Appendix E gives the statistical procedure. The contrasts are fixed in writing before the sweep, and scripts recompute every table and every plotted value from the recorded outputs.

References

Weihua Cheng, Hanwen Du, Chunxiao Li, Ersheng Ni, Liangdi Tan, Tianqi Xu, and Yongxin Ni. Uncertainty-aware knowledge tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27905–27913, 2025. doi: 10.1609/aaai.v39i27.35007. arXiv:2501.05415.

Harshita Chopra, Kshitish Ghate, Aylin Caliskan, Tadayoshi Kohno, Chirag Shah, and Natasha Jaques. Beyond cooperative simulators: Generating realistic user personas for robust evaluation of LLM agents. *arXiv preprint arXiv:2605.12894*, 2026.

Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995. doi: 10.1007/BF01099821.

Zhangqi Duan, Shuyan Huang, Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. Who am I? History-aware profiles for student simulation in tutoring dialogues. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. arXiv:2605.30051.

Scott H. Fraundorf, Zachary A. Caddick, Timothy J. Nokes-Malach, and Benjamin M. Rottman. Cognitive perspectives on maintaining physicians' medical expertise: IV. Best practices and open questions in using testing to enhance learning and retention. *Cognitive Research: Principles and Implications*, 8(1):53, 2023. doi: 10.1186/s41235-023-00508-8.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. Agent4Edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23923–23932, 2025. doi: 10.1609/aaai.v39i22.34565. arXiv:2501.10332.

Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2330–2339, 2020. doi: 10.1145/3394486.3403282. arXiv:2007.12324.

Danial Hooshyar, Gustav Šír, Yeongwook Yang, Tommi Kärkkäinen, Raija Hämäläinen, Ekaterina Krivich, Mutlu Cukurova, Dragan Gašević, and Roger Azevedo. Neural-symbolic knowledge tracing: Injecting educational knowledge into deep learning for responsible learner modelling. *arXiv preprint arXiv:2604.08263*, 2026.

Shuyan Huang, Alexander Scarlatos, Jaewook Lee, and Andrew Lan. Interpretable difficulty-aware knowledge tracing in tutor-student dialogues. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 612–623, 2026. doi: 10.18653/v1/2026.bea-1.43. arXiv:2605.01097.

Unggi Lee, Sookbun Lee, Yeil Jeong, Eunjoo Lee, Minchul Shin, and Hoilym Kwon. EduClaw-Bench: A long-horizon benchmark for pedagogical LLM agents with simulated learners. *arXiv preprint arXiv:2608.03206*, 2026.

Zhengyuan Liu, Stella Xin Yin, Bryan Chen Zhengyu Tan, Roy Ka-Wei Lee, Guimei Liu, Dion Hoe-Lian Goh, Wenya Wang, and Nancy F. Chen. The imperfect learner: Incorporating developmental trajectories in memory-based student simulation. *arXiv preprint arXiv:2511.05903*, 2025.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simpleKT: A simple but tough-to-beat baseline for knowledge tracing. In *International Conference on Learning Representations*, 2023. arXiv:2302.06881.

Joshua Mitton, Prarthana Bhattacharyya, Ralph Abboud, and Simon Woodhead. Knowing when to defer: Selective prediction for responsible knowledge tracing. In *Proceedings of the Impactful and Responsible AI Systems for Education Workshop*, volume 339 of *Proceedings of Machine Learning Research*, pp. 22–42, 2026. arXiv:2509.21514.

Baptiste Moreau-Pernet. EvalConvoLearn: An open-source framework for evaluating grounded learner simulations in tutoring conversations. In *Proceedings of the Impactful and Responsible AI Systems for Education Workshop*, volume 339 of *Proceedings of Machine Learning Research*, pp. 180–186, 2026. arXiv:2608.07497.

Anna T. Nickl and Karl-Heinz T. Bäuml. Retrieval practice reduces relative forgetting over time. *Memory*, 31(10):1412–1424, 2023. doi: 10.1080/09658211.2023.2270735.

Max Norris, Kobi Gal, and Sahan Bulathwela. Next token knowledge tracing: Exploiting pretrained LLM representations to decode student behaviour. *arXiv preprint arXiv:2511.02599*, 2025.

Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining*, pp. 384–389, 2019. arXiv:1907.06837.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, volume 28, pp. 505–513, 2015. arXiv:1506.05908.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. Lost in simulation: LLM-simulated users are unreliable proxies for human users in agentic evaluations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47423–47439, 2026. doi: 10.18653/v1/2026.acl-long.2192. arXiv:2601.17087.

Xinjie Sun, Kai Zhang, Qi Liu, Shuanghong Shen, Fei Wang, Yuxiang Guo, and Enhong Chen. DASKT: A dynamic affect simulation method for knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1714–1727, 2025. doi: 10.1109/TKDE.2025.3526584. arXiv:2502.10396.

Emma Trumble, Jason Lodge, Allison Mandrusiak, and Roma Forbes. Systematic review of distributed practice and retrieval practice in health professions education. *Advances in Health Sciences Education*, 29(2):689–714, 2024. doi: 10.1007/s10459-023-10274-3.

Shanshan Wang, Xueying Zhang, Xun Yang, Xingyi Zhang, and Keyang Wang. Learning states enhanced knowledge tracing: Simulating the diversity in real-world learning process. *Expert Systems with Applications*, 274:126838, 2025. doi: 10.1016/j.eswa.2025.126838. arXiv:2412.19550.

Meng Xia, Robin Schmucker, Conrad Borchers, and Vincent Aleven. Optimizing mastery learning by fast-forwarding over-practice steps. In *Two Decades of TEL. From Lessons Learnt to Challenges Ahead (EC-TEL 2025)*, Lecture Notes in Computer Science, pp. 549–563, 2025. doi: 10.1007/978-3-032-03870-8_37. arXiv:2506.17577.

Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2025. doi: 10.1145/3706598.3713773. arXiv:2502.02780.

Zhihao Yuan, Yunze Xiao, Ming Li, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. Towards valid student simulation with large language models. *arXiv preprint arXiv:2601.05473*, 2026.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 765–774, 2017. doi: 10.1145/3038912.3052580. arXiv:1611.08108.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. Mind the Sim2Real gap in user simulation for agentic tasks. In *Conference on Language Modeling*, 2026. arXiv:2603.11245.

A CORPORA, SIMULATORS AND HOW WELL THEY FIT

ASSIST2009 carries 15,911 items over 123 concepts and ASSIST2017 carries 3,162 items over 102 concepts. Each interaction carries up to four concept identifiers. Each learner record is cut into consecutive segments of at most 200 interactions, and a segment shorter than three interactions is dropped. Learners are split $70/10/20$ into training, validation and test at split seed 42, so the three groups hold the same learners for every architecture at every model seed. The test split holds 50,587 interactions on ASSIST2009 and 183,653 on ASSIST2017. Every model is trained with Adam at learning rate 10^{-3}, batch size 64, dropout 0.2 and embedding width 64, for at most 100 epochs with early stopping after 10 epochs without improvement in validation AUC. Each architecture keeps its reference sizes: the recurrent state of DKT is 200 wide, DKVMN holds 50 memory slots, and AKT, SAKT and simpleKT use four attention heads, simpleKT with two layers and AKT with a Rasch regulariser of 10^{-5}. Nothing is tuned per architecture, because the sweep examines the fitted simulator that a benchmark builder gets from the standard recipe.

Table 3 lists every fitted simulator. Test AUC runs from 0.727 to 0.783, and expected calibration error stays below 0.06 everywhere. Yet 16 of the 30 fits do not pass the spacing contrast. A next-response AUC well above chance does not ensure that a fitted simulator responds to spacing. Refitting barely changes test AUC. Across the three seeds, test AUC varies by at most 0.006 in any simulator and by 0.003 at the median, yet five of 10 simulators change at least one verdict (Appendix D).

Table 3: Every fitted simulator and how well it fits, with the three seeds side by side. Validation AUC selects the checkpoint, and test AUC and accuracy are on the test split. Test AUC runs from 0.727 to 0.783, and expected calibration error stays below 0.06 throughout. The table omits expected calibration error for width. Fits of similar quality give different verdicts, so next-response AUC does not settle whether a simulator reproduces a learning phenomenon.

		Valid AUC			Test AUC			Test ACC		
Corpus	Model	42	43	44	42	43	44	42	43	44
ASSIST2009	DKT	0.750	0.746	0.752	0.755	0.754	0.759	0.741	0.737	0.739
ASSIST2009	DKVMN	0.740	0.744	0.746	0.757	0.751	0.755	0.735	0.733	0.736
ASSIST2009	AKT	0.739	0.739	0.740	0.750	0.748	0.751	0.734	0.733	0.734
ASSIST2009	SAKT	0.729	0.729	0.733	0.743	0.739	0.744	0.728	0.727	0.728
ASSIST2009	simpleKT	0.736	0.736	0.741	0.749	0.744	0.749	0.731	0.728	0.732
ASSIST2017	DKT	0.782	0.781	0.781	0.783	0.782	0.782	0.723	0.722	0.722
ASSIST2017	DKVMN	0.770	0.769	0.768	0.769	0.768	0.768	0.714	0.713	0.712
ASSIST2017	AKT	0.766	0.766	0.767	0.765	0.767	0.768	0.709	0.712	0.712
ASSIST2017	SAKT	0.732	0.732	0.733	0.728	0.727	0.727	0.694	0.694	0.694
ASSIST2017	simpleKT	0.738	0.739	0.737	0.734	0.734	0.733	0.700	0.700	0.700

B How each contrast is constructed

Every contrast is a pair of sequences appended to the same history and scored with the same probe. The sample depends on the corpus alone, so every architecture and seed is tested on the same histories. The segments of the test split are read in their stored order, in batches of 64. A history is the first $t_0 = 50$ interactions of a segment, taken verbatim. The sample is every segment in the first eight batches that holds at least t_0 interactions. The sample size is $n = 220$ histories from 167 learners on ASSIST2009 and $n = 472$ histories from 162 learners on ASSIST2017.

The target concept, the probe and the item pools are fixed once per batch, before the two arrangements are drawn. The target concept is the concept with the most distinct items in the histories of the batch, and a batch needs at least 12 such items. The probe is eight of these items drawn at random, and the rest form the practice pool, so no probe item is practised. The fillers come from the items of other concepts in the same histories. Each of the first eight batches meets these conditions on both corpora. Each sequence holds $k = 8$ practice items in one of two arrangements, and the spacing and recency sequences also hold $f = 10$ fillers. Each arrangement draws its practice items and fillers from these pools at random and with replacement, and its gain is averaged over two samples. Every appended item and every probe item enters the model with all of its concept identifiers, as in fitting. Responses are written in rather than sampled. Unless a contrast specifies otherwise, every response in both arrangements is set correct. Because the two arrangements are otherwise drawn from the same pools in the same way, the expected difference between them is a function of the one factor the contrast varies. One random seed per run fixes the probe and every sample. The reported runs use seed 31 for the spacing and recency contrasts, seed 37 for the other two, seed 41 for the tutor comparison and seed 11 for the dose–response check. A run with another seed is a fresh draw of the items, and Appendix H repeats every contrast and the tutor comparison at two further draws on the same histories.

Spacing. Arrangement A places the eight practice items at evenly spaced positions. Arrangement B places its eight practice items in one contiguous block centred on the mean position of the practice in A. The two means are equal at 10 fillers, the setting of every reported contrast, and half a position apart at five. The contrast is $\mathrm{gain}(A) - \mathrm{gain}(B)$. Against a massed block placed first rather than at the matched position, the spacing effect on DKT / ASSIST2017 is $+0.030$ at five fillers and $+0.034$ at 10. Against the position-matched control, it is $+0.035$ and $+0.029$. However, on a simulator with recency but no spacing, the unmatched control reports recency as a spacing effect. The position-matched control is therefore used throughout.

Recency. A places the block of practice at the end and B places it at the start, with fillers elsewhere. The contrast is A minus B.

Transfer. A precedes the eight target items with eight items on a related concept, and B precedes them with eight items on an unrelated one. Because no prerequisite graph is published for either corpus, the related concept is the one that co-occurs most often with the target concept in the histories of the batch, and the unrelated concept co-occurs least. Only concepts with at least four distinct items in the batch and at least one co-occurrence with the target are candidates. Both arrangements end with eight target items from the same pool in the same positions, so the preparation is the only systematic difference between them.

Error penalty. A and B are built in the same way from the same pool, but in A the first response is set incorrect and in B every response is correct. The contrast is A minus B, and its human direction is negative.

The probe is eight target-concept items held out from practice. Each item is appended separately to the sequence and scored, so that scoring one cannot condition the state of the model for the next. Gain is the probe score after the sequence minus the probe score before it.

C THE DOSE–RESPONSE CHECK

Whether the arrangement of practice matters is a question only if practice itself matters. Table 4 gives probe gain against the number of target-concept practice items, with every response forced correct. Gain increases monotonically with k in all 10 cells. This increase is the premise a tutoring benchmark rests on. Fitted knowledge-tracing models do respond to correctly answered practice. The main text then asks whether the arrangement of practice affects them as it affects a person.

Table 4: Probe gain against the number of practice items on the target concept, with every response forced correct. Every simulator improves monotonically with practice, the premise a tutoring benchmark needs. The main text tests not whether practice helps but whether the arrangement of practice matters.

Corpus	Model	$k = 0$	$k = 4$	$k = 16$	$k = 32$
ASSIST2009	DKT	+0.000	+0.091	+0.158	+0.183
ASSIST2009	DKVMN	+0.000	+0.081	+0.109	+0.116
ASSIST2009	AKT	+0.000	+0.109	+0.135	+0.149
ASSIST2009	SAKT	+0.000	+0.042	+0.101	+0.128
ASSIST2009	simpleKT	+0.000	+0.083	+0.105	+0.132
ASSIST2017	DKT	+0.000	+0.180	+0.286	+0.319
ASSIST2017	DKVMN	+0.000	+0.160	+0.342	+0.388
ASSIST2017	AKT	+0.000	+0.119	+0.238	+0.286
ASSIST2017	SAKT	+0.000	+0.082	+0.219	+0.299
ASSIST2017	simpleKT	+0.000	+0.098	+0.245	+0.294

D EVERY CONTRAST AT EVERY SEED

Table 5 gives the verdict of every contrast on every simulator at each of the three seeds, and the main text argues from these verdicts. Table 6 gives the effect and the paired p-value behind each verdict. Of these 120 measurements, the main text reports the 40 at seed 42. The two tables separate a failure that recurs in all three fits from one specific to a single fit.

A row of Table 6 shows which verdicts hold at all three seeds. The error penalty is present at every seed in every cell. DKVMN shows no spacing at any seed on either corpus and no recency on ASSIST2017. The recency of DKVMN on ASSIST2009 carries a dagger at all three seeds. SAKT on ASSIST2009 shows none of the three schedule effects in the human direction at any seed. The inversions are less stable than the absences. The two largest inversions at seed 42 are the spacing and recency of simpleKT on ASSIST2017, and both take the human direction at the two further seeds. The inverted transfer of DKT on ASSIST2009 survives one refit and becomes an absence at the other. Of the 12 simulator–contrast pairs with a dagger at some seed, five have it at two seeds and six have it at one. Only the recency of DKVMN on ASSIST2009 has it at all three. Of the 21 pairs with an absence at some seed, five have it at all three.

Table 5: Verdict of each contrast at seeds 42, 43 and 44, in that order. `P` denotes the human direction at $p < 0.01$, and `f` denotes anything else. The last column indicates whether all three seeds agree on every contrast. The seeds agree in five of 10 simulators.

Corpus	Model	Spacing	Recency	Transfer	Error	Stable
ASSIST2009	DKT	PPP	PPP	fff	PPP	yes
ASSIST2009	DKVMN	fff	fff	fff	PPP	yes
ASSIST2009	AKT	PPP	PPP	fff	PPP	yes
ASSIST2009	SAKT	fff	fff	fff	PPP	yes
ASSIST2009	simpleKT	PPP	PPP	fff	PPP	yes
ASSIST2017	DKT	PfP	PPf	PPf	PPP	no
ASSIST2017	DKVMN	fff	fff	Pff	PPP	no
ASSIST2017	AKT	fff	fff	fPP	PPP	no
ASSIST2017	SAKT	fPf	fPf	ffP	PPP	no
ASSIST2017	simpleKT	fPP	fPP	fff	PPP	no

A single run cannot show which verdicts change, or where. Five of the 10 simulators change at least one verdict with the seed, and all five are on ASSIST2017. DKT on ASSIST2017 is the one cell that passes all four contrasts at seed 42. It loses spacing at the second seed and loses both recency and transfer at the third. SAKT on the same corpus passes spacing and recency only at the second seed and passes transfer only at the third. simpleKT turns its two largest inversions into passes at both further seeds, AKT gains transfer at the second seed and keeps it, and DKVMN loses transfer after the first. A single certification run therefore describes one fit. On ASSIST2017 every simulator changes a verdict with the seed, so no single run there speaks for other fits of the same architecture.

Of the 90 schedule measurements, 19 are significant the wrong way: four in spacing, five in recency and 10 in transfer. In all, 53 of the 90 reach $p < 0.01$ in one direction or the other. Of the 56 failures, 37 are absences. The 19 inversions sit in eight cells. The recency of DKVMN on ASSIST2009 is inverted at all three seeds, and its spacing is inverted at the second. The transfer of AKT on ASSIST2009 and the transfer of simpleKT on ASSIST2017 are inverted at the last two seeds, and the transfer of DKT on ASSIST2009 is inverted at the first two. The spacing and transfer of SAKT on ASSIST2009 are inverted at the first and last seeds. The transfer of simpleKT on ASSIST2009 and the recency of AKT on ASSIST2017 are inverted at the last seed, and the transfer of SAKT on ASSIST2017 is inverted at the first. At seed 42, the spacing and recency of simpleKT on ASSIST2017 are the two largest inversions in the study.

Seeds change verdicts despite modest effect changes. Over the three seeds, the effect of a cell–contrast pair changes by a median of 0.009. It changes by less than 0.01 in 25 of the 40 pairs and by less than 0.02 in 30. The largest change anywhere is 0.058, and it takes the recency of simpleKT on ASSIST2017 from an inversion to a pass. A verdict flips because a small effect crosses zero or the threshold, not because the fitted model behaves differently by a wide margin. This paper therefore tests direction rather than magnitude. The error penalty is the steadiest of the four contrasts. Its effect moves by at most 0.010 across the three seeds, and its verdict never changes.

E Statistical procedure

For the same history, each contrast compares two arrangements whose only systematic difference is what the contrast varies. The test is a two-sided paired t-test at $\alpha = 0.01$ on the per-history difference in probe gain. A verdict of *pass* requires both the direction human learners show and $p < \alpha$. Anything else is a fail. A fail means that the contrast does not certify the direction at this threshold, not that the effect is zero. A fail that is significant in the opposite direction is reported separately as an inversion, because it differs in kind from an absence.

Over the three seeds the grid holds 37 absences. Each individual two-sided 99% interval, the counterpart of the test, excludes the pooled median effect of the fits that pass the same contrast. The medians are 0.018 for spacing, 0.022 for recency and 0.019 for transfer. With a Bonferroni correction over all 120 measurements, 36 of the 37 intervals still exclude it. Of the 37 intervals, 23 also exclude the pooled smallest passing effect. With the median taken within each corpus instead, each

Table 6: Every contrast on every simulator at every seed, with the effect on probe gain and its paired p-value. A star marks an effect in the human direction at $p < 0.01$, a dagger marks an effect in the opposite direction at $p < 0.01$, and an unmarked effect is not significant. The main text reports the seed-42 columns. The other two seeds separate a failure that recurs in all three fits from one specific to a single fit.

			Seed 42		Seed 43		Seed 44	
Corpus	Model	Contrast	Effect	p	Effect	p	Effect	p
ASSIST2009	DKT	Spacing	$+0.027^{*}$	$< 10^{-3}$	$+0.035^{*}$	$< 10^{-3}$	$+0.021^{*}$	$< 10^{-3}$
		Recency	$+0.036^{*}$	$< 10^{-3}$	$+0.061^{*}$	$< 10^{-3}$	$+0.042^{*}$	$< 10^{-3}$
		Transfer	$-0.019^{\dagger}$	$< 10^{-3}$	$-0.022^{\dagger}$	$< 10^{-3}$	$+0.005$	0.177
		Error	-0.020^{*}	$< 10^{-3}$	-0.017^{*}	$< 10^{-3}$	-0.011^{*}	$< 10^{-3}$
ASSIST2009	DKVMN	Spacing	-0.001	0.602	$-0.003^{\dagger}$	0.007	$+0.001$	0.450
		Recency	$-0.006^{\dagger}$	$< 10^{-3}$	$-0.005^{\dagger}$	$< 10^{-3}$	$-0.005^{\dagger}$	$< 10^{-3}$
		Transfer	$+0.002$	0.165	$+0.003$	0.012	-0.003	0.192
		Error	-0.013^{*}	$< 10^{-3}$	-0.013^{*}	$< 10^{-3}$	-0.016^{*}	$< 10^{-3}$
ASSIST2009	AKT	Spacing	$+0.010^{*}$	$< 10^{-3}$	$+0.009^{*}$	$< 10^{-3}$	$+0.005^{*}$	0.002
		Recency	$+0.018^{*}$	$< 10^{-3}$	$+0.008^{*}$	0.002	$+0.007^{*}$	$< 10^{-3}$
		Transfer	-0.001	0.588	$-0.014^{\dagger}$	$< 10^{-3}$	$-0.016^{\dagger}$	$< 10^{-3}$
		Error	-0.006^{*}	0.009	-0.013^{*}	$< 10^{-3}$	-0.008^{*}	$< 10^{-3}$
ASSIST2009	SAKT	Spacing	$-0.009^{\dagger}$	$< 10^{-3}$	$+0.001$	0.285	$-0.004^{\dagger}$	0.001
		Recency	-0.003	0.049	-0.001	0.581	-0.003	0.012
		Transfer	$-0.006^{\dagger}$	$< 10^{-3}$	-0.003	0.109	$-0.006^{\dagger}$	$< 10^{-3}$
		Error	-0.017^{*}	$< 10^{-3}$	-0.015^{*}	$< 10^{-3}$	-0.011^{*}	$< 10^{-3}$
ASSIST2009	simpleKT	Spacing	$+0.021^{*}$	$< 10^{-3}$	$+0.017^{*}$	$< 10^{-3}$	$+0.027^{*}$	$< 10^{-3}$
		Recency	$+0.026^{*}$	$< 10^{-3}$	$+0.022^{*}$	$< 10^{-3}$	$+0.024^{*}$	$< 10^{-3}$
		Transfer	-0.006	0.050	-0.002	0.508	$-0.007^{\dagger}$	0.004
		Error	-0.021^{*}	$< 10^{-3}$	-0.019^{*}	$< 10^{-3}$	-0.012^{*}	$< 10^{-3}$
ASSIST2017	DKT	Spacing	$+0.029^{*}$	$< 10^{-3}$	$+0.002$	0.470	$+0.009^{*}$	$< 10^{-3}$
		Recency	$+0.036^{*}$	$< 10^{-3}$	$+0.011^{*}$	$< 10^{-3}$	$+0.001$	0.595
		Transfer	$+0.017^{*}$	$< 10^{-3}$	$+0.024^{*}$	$< 10^{-3}$	$+0.001$	0.632
		Error	-0.023^{*}	$< 10^{-3}$	-0.023^{*}	$< 10^{-3}$	-0.021^{*}	$< 10^{-3}$
ASSIST2017	DKVMN	Spacing	-0.001	0.780	-0.003	0.132	-0.001	0.799
		Recency	-0.002	0.446	$+0.006$	0.012	$+0.001$	0.710
		Transfer	$+0.021^{*}$	$< 10^{-3}$	$+0.006$	0.038	-0.003	0.257
		Error	-0.042^{*}	$< 10^{-3}$	-0.052^{*}	$< 10^{-3}$	-0.052^{*}	$< 10^{-3}$
ASSIST2017	AKT	Spacing	-0.004	0.075	-0.004	0.116	-0.000	0.975
		Recency	$+0.000$	0.892	$+0.001$	0.609	$-0.012^{\dagger}$	$< 10^{-3}$
		Transfer	$+0.001$	0.735	$+0.033^{*}$	$< 10^{-3}$	$+0.014^{*}$	$< 10^{-3}$
		Error	-0.022^{*}	$< 10^{-3}$	-0.019^{*}	$< 10^{-3}$	-0.015^{*}	$< 10^{-3}$
ASSIST2017	SAKT	Spacing	-0.000	0.805	$+0.008^{*}$	$< 10^{-3}$	-0.001	0.268
		Recency	$+0.001$	0.429	$+0.021^{*}$	$< 10^{-3}$	-0.003	0.013
		Transfer	$-0.009^{\dagger}$	$< 10^{-3}$	-0.000	0.723	$+0.005^{*}$	$< 10^{-3}$
		Error	-0.026^{*}	$< 10^{-3}$	-0.020^{*}	$< 10^{-3}$	-0.020^{*}	$< 10^{-3}$
ASSIST2017	simpleKT	Spacing	$-0.029^{\dagger}$	$< 10^{-3}$	$+0.019^{*}$	$< 10^{-3}$	$+0.015^{*}$	$< 10^{-3}$
		Recency	$-0.035^{\dagger}$	$< 10^{-3}$	$+0.022^{*}$	$< 10^{-3}$	$+0.014^{*}$	$< 10^{-3}$
		Transfer	-0.007	0.016	$-0.009^{\dagger}$	0.002	$-0.010^{\dagger}$	0.001
		Error	-0.020^{*}	$< 10^{-3}$	-0.020^{*}	$< 10^{-3}$	-0.021^{*}	$< 10^{-3}$

of the 29 intervals with a passing effect of the same contrast on its corpus excludes that median. The other eight are transfer absences on ASSIST2009, where transfer passes in no fit.

The counts are uncorrected for multiple comparisons. This is the conservative choice for the main claim that contrasts fail, because a correction can only turn passes and inversions into absences and so never turns a failure into a pass. The error penalty passes at $p < 0.01$ in all 30 fitted models and at $p < 10^{-3}$ in 29 of them, and a Bonferroni correction over all 120 measurements still leaves it significant in 27 of the 30. Of the 34 schedule passes, 26 survive the correction and eight become absences. The same correction keeps 11 of the 19 inversions and turns the other eight into absences. These eight include the recency inversion of DKVMN on ASSIST2009 at the last seed. The correction changes only the kind of that failure, from an inversion to an absence, and the failure holds at every seed under either threshold. The seed replication screens the positive results for stability. A verdict that appears at one seed and not at the others is reported as unstable, whatever its p-value.

The sample holds 220 histories from 167 learners on ASSIST2009 and 472 histories from 162 learners on ASSIST2017. It is identical across architectures and seeds because it depends on the corpus alone. A learner with a long record contributes several histories, up to four on ASSIST2009 and 11 on ASSIST2017. Averaging the histories of each learner first and repeating every test with learners as the unit leaves 115 of the 120 verdicts unchanged. Four changes are at seed 43. On ASSIST2017 the recency pass of DKT and the transfer inversion of simpleKT become absences. On ASSIST2009 the spacing inversion of DKVMN becomes an absence, and its transfer absence becomes a pass. The fifth is the error penalty of AKT on ASSIST2009 at seed 42, which becomes an absence at $p = 0.0102$. The same analysis leaves all 30 tutor comparisons unchanged, and the learner-level spacing verdict anticipates the benchmark outcome in 23 of the 30 fits. The recency of DKVMN on ASSIST2009 stays inverted at all three seeds, so the one inversion that holds at every seed also holds with learners as the unit. Under the drawn target concepts of Appendix J, the same analysis leaves 222 of the 240 verdicts and 56 of the 60 tutor comparisons unchanged. Of the 18 changes, 13 turn a transfer verdict into an absence. The spacing of DKVMN keeps every verdict, and the error penalty changes once. It becomes an absence for AKT on ASSIST2009 at the last seed under the second drawn targets. Two spaced wins become ties, one tie becomes a massed win, and another tie becomes a spaced win.

F THE RESPONSE CHANNEL IN FULL

For each simulator, Table 7 gives the response and content channels, which are the two magnitudes behind every effect in Table 6. The two forced runs share one random seed, so the response channel compares the same 16 practice items with every answer forced correct against every answer forced wrong. It moves the probe score by 0.19 to 0.67. The content channel compares practice on the target concept against practice on other concepts under the same forcing and is averaged over the two forcings. It moves the probe score by -0.16 to $+0.13$. The response channel is larger in magnitude in every one of the 10 cells.

The four contrasts therefore write the responses in. A contrast that lets the simulator answer for itself mixes the two channels in unknown proportion. The larger channel is the response channel, and it says nothing about what the sequence teaches. It says only that the learner is seen to answer correctly. Two tutors that elicit correct answers at different rates can therefore separate on any benchmark in which the simulator answers for itself, whether or not either tutor teaches anything. The consequence experiment in Section 4.3 deliberately lets the simulator answer for itself, because a leaderboard does. The controlled grid isolates the effect of the order and content of practice.

G THE TUTOR COMPARISON AT THREE SEEDS

Table 8 gives all 30 outcomes of the tutor comparison. The spaced tutor wins on 16, neither tutor wins on 12, and the massed tutor wins on two. Both massed wins occur at seed 42. At the other two seeds, the same simulators either place the spaced tutor ahead or cannot separate the two tutors. The massed wins thus do not recur across fits. The DKVMN ties are harder to dismiss. DKVMN cannot separate the two tutors at any seed on either corpus. It is also the architecture in which spacing and

Table 7: The two channels a probe gain can come from, after 16 practice items. The response channel is the gain from the same items with every answer forced correct minus every answer forced wrong. The content channel is the gain from the target concept minus other concepts under the same forcing, averaged over the two forcings. The response channel is larger in magnitude in every cell, by a median factor of 10. Hence a contrast in which the simulator answers for itself mixes the two channels, and the larger of them is the response channel.

Corpus	Model	Response channel	Content channel	Ratio
ASSIST2009	DKT	$+0.549$	$+0.030$	18.0
ASSIST2009	DKVMN	$+0.673$	-0.163	4.1
ASSIST2009	AKT	$+0.495$	$+0.004$	124.1
ASSIST2009	SAKT	$+0.352$	-0.043	8.1
ASSIST2009	simpleKT	$+0.398$	-0.010	41.2
ASSIST2017	DKT	$+0.315$	$+0.042$	7.6
ASSIST2017	DKVMN	$+0.339$	$+0.133$	2.6
ASSIST2017	AKT	$+0.195$	$+0.017$	11.4
ASSIST2017	SAKT	$+0.280$	$+0.042$	6.7
ASSIST2017	simpleKT	$+0.266$	$+0.022$	12.1
Median		$+0.346$	$+0.020$	9.7

recency fail at every seed in Table 5. A benchmark built on a DKVMN simulator therefore gives the spaced tutor no credit over the massed tutor at any of the three seeds.

Table 8: A spaced tutor against a massed tutor in the benchmark setting, at all three seeds. Each tutor draws its own eight target items and 10 fillers from the same pools by the same rule, and the two tutors differ systematically only in how they arrange them. The simulator answers for the learner. Gain is the spaced minus massed probe gain averaged over four samples, and the winner is decided at $p < 0.01$ on the paired difference. Both massed wins occur at the first seed. DKVMN cannot separate the tutors at any seed on either corpus.

		Seed 42		Seed 43		Seed 44	
Corpus	Model	Gain	Winner	Gain	Winner	Gain	Winner
ASSIST2009	DKT	$+0.037$	spaced	$+0.036$	spaced	$+0.023$	spaced
ASSIST2009	DKVMN	$+0.002$	neither	$+0.000$	neither	$+0.000$	neither
ASSIST2009	AKT	$+0.028$	spaced	$+0.015$	spaced	$+0.010$	neither
ASSIST2009	SAKT	-0.004	neither	$+0.009$	spaced	$+0.000$	neither
ASSIST2009	simpleKT	$+0.019$	spaced	$+0.027$	spaced	$+0.032$	spaced
ASSIST2017	DKT	$+0.039$	spaced	$+0.033$	spaced	$+0.044$	spaced
ASSIST2017	DKVMN	-0.003	neither	$+0.004$	neither	-0.006	neither
ASSIST2017	AKT	$+0.014$	spaced	$+0.002$	neither	$+0.005$	neither
ASSIST2017	SAKT	-0.007	massed	$+0.005$	spaced	-0.003	neither
ASSIST2017	simpleKT	-0.010	massed	$+0.031$	spaced	$+0.022$	spaced

H THE VERDICTS AT THREE DRAWS OF THE ITEMS

The main text reports one draw of the items (Appendix B). Two further draws repeat the four contrasts and the tutor comparison on all 30 fitted models, with the same histories and target concepts and fresh items. The second draw uses seeds 131, 137 and 141 in place of 31, 37 and 41, and the third uses 231, 237 and 241. In every fit, the gain a history receives from massed practice or from errorless practice correlates across draws at $r \geq 0.44$, and at $|r| \leq 0.15$ once the histories are shuffled. Table 9 gives the tallies of the main text at each draw.

The error penalty passes in all 30 fitted models at every draw. At the first seed, spacing passes in four, four and three of the 10 simulators at the three draws. Recency passes in four at each draw, and transfer passes in two, three and two. DKT on ASSIST2017 is the only simulator that passes all four contrasts at the first seed, and it does so at every draw. The three large inversions at the first seed

Table 9: The tallies of the main text at three draws of the items. Each draw repeats every contrast and the tutor comparison on all 30 fitted models, on the same histories and target concepts with fresh practice items, fillers and probe items. The first draw is the one the main text reports, and the last row tests the per-history effect averaged over the three draws. Anticipated counts the fits in which the spacing verdict anticipates the winner.

	Passes (of 30)				Inversions	Winner (of 30)			Anticipated
Draw	Spacing	Recency	Transfer	Error	(of 90)	Spaced	Neither	Massed	(of 30)
First (main text)	14	14	6	30	19	16	12	2	22
Second	14	13	8	30	20	14	14	2	26
Third	14	13	5	30	16	14	14	2	24
Averaged	16	14	6	30	20	17	9	4	26

recur at every draw, and so do three of the four small ones. Five absences and one inversion hold at every seed in the main text. All but the recency of SAKT on ASSIST2009 fail at every seed in both further draws, and that recency passes at one seed of the second draw. With absences and inversions counted alike, 12 of the 30 pairs of a simulator and a schedule contrast fail at every seed in all three draws. The recency of DKVMN on ASSIST2009 never passes, although it is an absence rather than an inversion at one seed in the second draw and at one in the third. At every draw at least half of the simulators change a verdict with the seed, five, seven and seven of the 10 in turn.

Single verdicts near the threshold do change with the draw. Of the 120 verdicts, 108 pass or fail alike at all three draws, and 94 also agree on whether a failure is an absence or an inversion. The 12 that change are small effects, none beyond 0.021 at any draw, and five of them are transfer verdicts. Across the three draws, the median range of an effect within a fit is 0.004, about half its median range of 0.009 across seeds. Averaging each history over the three draws before the test leaves 116 of the 120 verdicts as the first draw reports them.

The tutor comparison depends on the simulator at every draw. The massed tutor wins on two of the 30 fitted models at each draw, and the spaced tutor wins on 16, 14 and 14 in turn. DKVMN separates the two tutors at no seed in any draw. Of the 30 comparisons, 22 are the same at all three draws, and averaging over the draws leaves 27 as the first draw reports them. The spacing verdict anticipates the winner in 22, 26 and 24 of the 30 fits.

I THE VERDICTS BY TARGET CONCEPT

Each batch of a run fixes one target concept, the concept with the most distinct items in the histories of the batch (Appendix B). Over the eight batches this rule selects four concepts on ASSIST2009, tested on 87, 57, 46 and 30 histories. On ASSIST2017 it selects two, tested on 415 and 57 histories. Table 10 repeats each contrast of the first draw within each concept of each fit, with the same paired test and the same uncorrected threshold.

The error penalty keeps the human direction in all 90 cells and passes in 71 of them, although the smallest concept holds 30 histories. No cell inverts it. Spacing never passes for one concept and inverts for another in the same fit, and the spacing of DKVMN is absent in all 18 of its cells. Transfer is the most mixed. It passes in 17 cells and is inverted in 28, and in 10 fits it passes for one target concept and is inverted for another. Recency does so in four fits. A Bonferroni correction over the 360 cells leaves the error penalty passing in 46 cells and never inverted. It leaves transfer passing in eight cells and inverted in 17, with two split fits. The inverted recency of DKVMN on ASSIST2009 is significant the other way in its largest concept at every seed. At the first seed, SAKT and simpleKT on ASSIST2009 both pass transfer for one target concept and invert it for another. A pooled transfer verdict can therefore average effects of opposite sign.

J THE VERDICTS AT DRAWN TARGET CONCEPTS

The main text takes as the target of each batch the concept with the most distinct items (Appendix B). This rule tests four concepts on ASSIST2009 and two on ASSIST2017. Two further runs draw the

Table 10: Every contrast within each target concept of each fit, at the first draw. With four target concepts on ASSIST2009 and two on ASSIST2017, the 15 fits of each corpus give 60 and 30 cells per contrast. A cell passes when its effect has the human direction at $p < 0.01$, is inverted when the effect is significant the other way and is absent otherwise. Human sign counts the cells whose effect has the human direction, significant or not. Split counts the fits in which the contrast passes for one concept and is inverted for another.

Corpus	Contrast	Cells	Pass	Absence	Inversion	Human sign	Split
ASSIST2009	Spacing	60	20	38	2	40	0
ASSIST2009	Recency	60	23	30	7	40	2
ASSIST2009	Transfer	60	8	32	20	19	7
ASSIST2009	Error	60	44	16	0	60	0
ASSIST2017	Spacing	30	6	21	3	15	0
ASSIST2017	Recency	30	9	17	4	14	2
ASSIST2017	Transfer	30	9	13	8	17	3
ASSIST2017	Error	30	27	3	0	30	0

target of each batch uniformly at random from the concepts every contrast can use. These concepts have at least 12 distinct items and at least two candidates for the related and the unrelated concept. The candidates must not all co-occur with the target equally often. Both runs keep the histories and item seeds of the main text and record the target of every batch. The three rules test 14 concepts on ASSIST2009 and 15 on ASSIST2017. Table 11 gives the tallies of the main text under each rule.

Table 11: The tallies of the main text when the target concept of each batch is drawn at random rather than taken as the concept with the most items. Each target draw repeats every contrast and the tutor comparison on all 30 fitted models, on the same histories and with the item seeds of the main text, whose rule is the first row. Anticipated counts the fits in which the spacing verdict anticipates the winner.

	Passes (of 30)				Inversions	Winner (of 30)			Anticipated
Target rule	Spacing	Recency	Transfer	Error	(of 90)	Spaced	Neither	Massed	(of 30)
Most items	14	14	6	30	19	16	12	2	22
First draw	16	15	10	29	13	18	11	1	21
Second draw	15	19	20	30	6	16	12	2	25

The target concept moves verdicts at least as much as the fit does. The 30 fits give 120 verdicts, and each is compared across the three pairs of rules. A verdict changes between pass and fail in 72 of these 360 comparisons, and in 24 of 360 across the three pairs of item draws. Across the three pairs of seeds of the main text, the 40 verdicts of the 10 simulators change in 20 of 120 comparisons. The comparisons share fits and histories, so the rates are descriptive. Five absences and one inversion hold at every seed in the main text. All six still fail at every seed under the first drawn targets. Under the second, the recency of SAKT on ASSIST2009 and the spacing of DKVMN on ASSIST2017 still fail at every seed.

Other results hold under every rule. The error penalty passes in 89 of its 90 verdicts across the three rules. Its one absence, AKT on ASSIST2009 at the last seed under the first drawn targets, keeps the human direction, and none of the 510 concept cells inverts it. DKVMN fails spacing at every seed on both corpora under every rule, and none of its 102 concept cells passes spacing. DKT on ASSIST2017 passes all four contrasts at the first seed under every rule. The tutor tallies of Table 11 change little, although the outcome of 11 of the 30 fits differs between rules. DKVMN separates the two tutors under no rule. A verdict thus belongs to a fitted model on the concepts it is tested on, and a certification should report the concepts it covers.

K WHY THE THIRD CORPUS IS EXCLUDED

NIPS34 is the third corpus this literature commonly uses, but it is not in the grid. Its 948 items carry 86 concept identifiers between them, up to six on one item, and the first identifier is the same

one on every item. The protocol and the diagnostic pass over the fitted model both take the first identifier as the concept of an item. The pass therefore finds one concept in every batch it reads. Spacing, recency and transfer set a target concept against other concepts, and the error penalty uses the batches of transfer. On NIPS34 the other-concept set is empty, so every batch is skipped and the sample size is zero for all four contrasts. The zero sample size reflects the corpus as prepared rather than a fault in the contrasts. Hence none of the four contrasts as constructed here can be run on the corpus under the standard preprocessing. The grid therefore contains two corpora, and any study that reports concept transfer on NIPS34 should state which identifier on an item serves as the concept.

L Reproduction

Given the fitted checkpoints, the study consists of four sweeps over the simulators and a set of analysis scripts. Three of the sweeps are repeated at two further draws of the items and at two random draws of the target concepts. The dose sweep runs three times per simulator, with every response forced correct, with every response forced wrong and with the responses sampled. The two channels in Appendix F are read from the first two runs, and the sampled run is recorded but not reported. On the training host, the four contrasts on one fitted model take 16 to 202 seconds (median 25 seconds over 29 timed runs, with DKVMN the slowest). The times include process start and checkpoint loading. The analysis scripts recompute every table and every plotted value from the recorded outputs, and the two conceptual figures are vector drawings with no raster. The learner-level tests of Appendix E also read the corpus files to map each history to its learner. A verification script recomputes every cell of Tables 1 and 2 from the recorded outputs and runs 199 further checks against the text. These checks cover the inversion and absence counts and both consequence tallies. It fails if any of them disagrees with the recorded outputs.