Inferring Cognitive Strategy Transitions in Multiplication Fact Learning with Graph-Based Skill Dependency Modeling

Ark Deliev
Faculty of Science, University of Amsterdam
ark.deliev@student.uva.nl

Abstract

Cognitive theories propose that multiplication learning progresses from slow calculation-like responding toward faster responding often attributed to retrieval and automaticity. We test whether a data-driven graph-based skill dependency model (GraafTel), fit with k latent components (k = 3–6), recovers an item representation that includes a robust RT-linked "slow" component across learning stages. Using 315,690 practice trials from 540 children ages 6–10 in an adaptive fact-learning system, we relate item-level component requirements to response time and to a system-estimated forgetting parameter α that is partly RT-mediated in Level 3. Across k, one component shows a strong positive association with RT across levels and encounter positions, whereas the remaining components show weaker or more stage-dependent time associations. The RT-linked component also shows positive raw correlations with α, although this relation is reduced after controlling for accuracy. These results suggest that graph-based latent structure can serve as a computational marker of strategy-sensitive dissociations in large-scale learning traces.

Keywords: skill acquisition; multiplication; response time; forgetting; computational modeling

Introduction

How children move from slow arithmetic procedures to fast fact retrieval is a core question in cognitive science. Most evidence comes from laboratory tasks or self-reports. Educational platforms now provide large longitudinal traces, but those traces do not directly label strategy use.

Classic theories describe multiplication learning as a shift from calculation-like procedures, such as repeated addition or decomposition, to direct retrieval and, with practice, greater automaticity (Siegler, 1988; Lemaire & Siegler, 1995; Anderson, 1982; Tenison, Fincham, & Anderson, 2016). Overlapping-waves accounts further predict that multiple strategies can coexist and that different facts can transition at different rates (Siegler, 1988; Lemaire & Siegler, 1995). The expected pattern is item-level heterogeneity rather than a single global stage shift.

Naturalistic traces are valuable but difficult to interpret because item difficulty, practice history, and context effects are entangled, and strategy remains latent. We therefore ask whether a graph-based latent skill model, fit to naturalistic multiplication traces, can recover item-level structure that aligns with strategy-sensitive behavioral signatures such as response time.

GraafTel is a graph-based latent skill dependency model that represents each item with a small vector of latent component requirements and each learner with a matching mastery vector. At a high level, the model predicts success from how strongly an item depends on each latent component and how much the learner has mastered those components.

We test whether graph-learned item components in naturalistic multiplication traces show stable versus stage-specific links to RT, whether this pattern is stable across $k = 3$–6 and under difficulty controls, and how the RT-linked component relates to the system-estimated forgetting parameter α as a secondary RT-dependent signal. This framing also yields a falsifiable validation path for future strategy-grounded studies.

We test three hypotheses. **H1**: Across $k = 3$–6, at least one latent component shows a stable positive association with RT across levels and encounter positions. **H2**: The remaining components show weaker or more stage-dependent RT associations. **H3**: The RT-linked component also shows a positive raw association with the system-estimated forgetting parameter α. Because α in Level 3 is partly RT-mediated, this serves only as a compatibility check and not as independent evidence of strategy.

Related Work

Strategy mixtures and problem-specific selection

Overlapping-waves accounts emphasize that learners use multiple strategies concurrently and adaptively, with strategy selection depending on problem characteristics and experience (Siegler, 1988; Lemaire & Siegler, 1995). This implies that strategy-sensitive signatures should be visible at the item level: some facts may remain computation-like and slower, whereas others become retrieval-dominant.

Qualitative shifts in skill acquisition

Classic accounts of skill acquisition describe qualitative changes in processing as declarative knowledge becomes compiled and performance becomes faster and more automatic (Anderson, 1982). Consistent with this view, work on arithmetic learning has reported phase-like transitions in neural and behavioral signatures (Tenison et al., 2016). Our approach does not directly observe strategies or neural phases, but instead tests whether latent structure discovered from practice traces captures stage-dependent dissociations consistent with these theories.

Learning traces and cognitive modeling

Adaptive learning systems often use knowledge tracing and memory models to estimate latent knowledge states and forgetting from performance and timing data (Abdelrahman, Wang, & Nunes, 2023; van Rijn, van Maanen, & van

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Woudenberg, 2009; Sense, Behrens, Meijer, & van Rijn, 2016). These approaches primarily track latent parameters over time. In contrast, graph-based skill dependency modeling aims to discover interpretable latent structure over items and their dependencies. It can therefore complement parameter-tracking approaches by generating hypotheses about which kinds of items recruit distinct processes.

Methods

Dataset and learning environment

We analyze an anonymized dataset collected in a naturalistic deployment of an adaptive multiplication learning system across 11 primary schools (Iancu et al., 2024). Participants were 540 children ages 6–10 who completed 315,690 trials over 17,575 sessions spanning 197 days. The deployed content focused on single-digit multiplication with operands 1–10, covering up to 100 directed facts.

The system implements three levels that broadly correspond to a progression toward fluency (Iancu et al., 2024). In Level 1, items are presented sequentially and a single correct response marks an item as completed for that level. In Level 2, the system schedules practice adaptively using a cognitive model of memory (van Rijn et al., 2009) and uses accuracy to update item parameters. In Level 3, the adaptive scheduling remains in place but adds time pressure through an 8-second response deadline, and parameter updates incorporate both accuracy and RT.

Because Level 3 updates incorporate RT, any later analyses involving the system-estimated forgetting parameter α are treated as secondary and potentially RT-coupled. We therefore interpret α associations as compatibility checks rather than independent causal evidence.

System memory model and interpretation of α

The platform's scheduler is based on a memory activation model that predicts the strength of a fact from its practice history (van Rijn et al., 2009). Let t_j denote past practice times for fact f. Activation at time t is computed as

$$A(f,t) = \ln \sum_j (t - t_j)^{-d}, \tag{1}$$

and a fact is treated as sufficiently mastered if its predicted activation exceeds a threshold at a short look-ahead, for example 24h (Iancu et al., 2024; van Rijn et al., 2009).

The system additionally maintains a fact-specific speed-of-forgetting parameter α that is system-estimated and model-derived. This parameter modulates how rapidly activation decays between repetitions. Mechanistically, α is initialized around 0.3 and is adjusted in ± 0.01 steps in Level 2 based on accuracy. In Level 3, the updates additionally incorporate RT, making α partly RT-mediated (Iancu et al., 2024). We therefore treat α as the system's internal estimate of forgetting rather than a direct behavioral measure, and note that Level 3 α associations are not independent of RT.

Encounter position is defined per student and level as the first attempt at each fact, a middle attempt halfway between the first and last attempt, adjusted when necessary to differ from the first, and the last or most recent attempt.

Preprocessing. We divide the trial stream into nine level$\times$encounter subsets, that is, Levels 1–3 $\times$ first/middle/last, and compute item-level aggregates within each subset. Accuracy and exposure counts use all attempts in the subset. RT aggregates use only non-missing RTs, which are available in Levels 2–3.

Response time is recorded by the platform in milliseconds for Levels 2–3. We compute mean item RTs using all available non-missing RTs, without additional manual trimming beyond the platform's logging and the Level 3 time constraint.

We analyze the released, pre-cleaned trial dataset (Iancu et al., 2024). Analyses use all trials with recorded correctness and treat missing RTs as missing, since RT is not recorded in Level 1.

Descriptives. RTs are strongly right-skewed, with overall mean ≈ 5.16s, median ≈ 2.51s, and 99th percentile ≈ 36.17s, which motivates log-mean and trimmed-summary robustness checks. Level 3 RTs are also right-censored by the platform's 8-second response deadline. The system-estimated forgetting parameter α shows modest dispersion, with overall mean ≈ 0.29 and SD ≈ 0.05, and is lower on average in Level 3 than in Level 2.

RT in this classroom deployment may reflect several non-strategic factors, including interface and input-device variation, individual typing speed, and local deadline pressure. Broader classroom context may also contribute. Our controls, namely accuracy, exposure, and operand-size proxies, address part of item difficulty and practice structure, but they cannot identify or remove all environment-level and person-level timing confounds.

GraafTel latent skill model

In simple terms, the model represents each multiplication fact in terms of a few hidden requirements and each learner in terms of how much they have mastered those requirements. More formally, GraafTel represents each item with a skill-requirement vector x and each student with a mastery vector s, with elements bounded to $[0,1]$. Given k components, the expected success probability is

$$ES = \prod_{i=1}^{k} (1 - x_i + x_i s_i). \tag{2}$$

Equation (2) has a simple interpretation: each factor is a soft prerequisite gate. If an item does not rely on component i ($x_i \approx 0$), the factor is near 1 and does not constrain success. If an item relies strongly on component i ($x_i \approx 1$), the factor becomes approximately s_i, so success depends on mastery of that component. The product form therefore implements a conjunctive, or "AND-like," combination across components:

items can depend on multiple latent requirements simultaneously, and weak mastery on any strongly required component reduces expected success.

This inductive bias supports cognitive accessibility. If subsets of facts systematically differ in the processes they recruit, for example because some remain more time-demanding and error-prone even with practice, then those subsets should exhibit distinct accuracy trajectories and dependencies in the practice graph. GraafTel can capture such structure as separate latent components because it must explain item-by-student performance patterns using shared, reusable dimensions. We treat the learned item weights as computational markers and evaluate them against strategy-sensitive proxies, namely RT and the system's α estimate.

Model parameters are optimized to reduce prediction error using gradient-based updates with Adam-style adaptive steps (Kingma & Ba, 2015). We fit configurations with $k \in \{3,4,5,6\}$ components, estimating separate models for each level, 1–3, and encounter position, first/middle/last, and training each model for 4000 epochs.

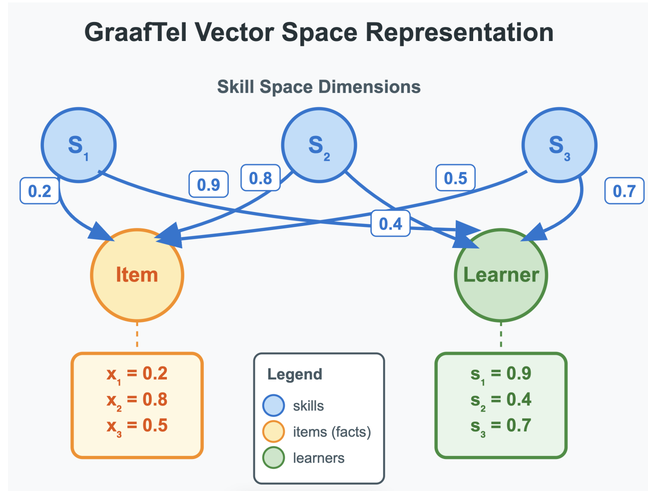

Figure 1: Illustration of the vector representation: item component weights (x) and student mastery (s) jointly determine expected success probability.

Training details and stability checks. Item and student vectors are constrained to $[0,1]$ throughout optimization. In the released model artifacts, training uses 4000 epochs with an Adam-style update rule and step size $\alpha \approx 0.001$. Student updates are computed on subsets of students per iteration for scalability.

To characterize the k trade-off, we summarize final training error and inter-component overlap, $|r|$ averaged over component pairs, for $k \in \{3,4,5,6\}$. Final error is nearly unchanged across k at 0.160–0.161, whereas overlap varies modestly, with mean $|r| = 0.18$–0.28. We report main figures and tables for $k = 3$ for compactness, but all key conclusions are stable across $k = 3$–6 (Table 4).

Because k is a modeling choice and component indices are unlabeled, and can permute across runs or epoch length, we interpret components by their empirical associations rather than by index. For each k, we label as the *RT-linked ("slow") component* the one most positively associated with log-mean RT in a single reference condition, namely Level 3 middle encounter. This labeling step is not itself confirmatory. The main test is whether that same component continues to predict RT across other levels, encounter positions, and model sizes. The main evidential claims therefore rest on generalization across levels, encounters, and k, together with robustness checks and incremental prediction under controls. A falsifying pattern would be that this labeled component fails to generalize outside the reference condition or loses incremental value once controls are included. External validation of strategy interpretation still requires independent measures such as think-aloud reports.

Linking components to strategy-sensitive measures

To test cognitive interpretations, we relate item component weights, that is, the x vectors, to two measures available in the learning system.

Response time (RT). For each level, 2–3, and encounter position, first/middle/last within level, we compute mean item RTs and correlate them with item component weights, applying Bonferroni correction across components for the $k = 3$ main figures.

As robustness checks, we additionally evaluate alternative RT summaries, namely median, log-mean, and top-1% trimmed means, and fit difficulty-controlled item-level regressions including mean accuracy and an operand-magnitude proxy.

All correlations are computed at the item, or multiplication-fact, level.

Forgetting (decay rate α). The adaptive system estimates an item-level speed-of-forgetting parameter α that is updated from performance (Iancu et al., 2024; van Rijn et al., 2009). In Level 2, α updates are based on accuracy, whereas in Level 3 updates incorporate both accuracy and RT. We therefore interpret α correlations as reflecting the system's estimated forgetting rather than a direct behavioral measure. We correlate Level 1 item component weights for $k = 3$ with Level 3 item α values across encounter positions.

To test whether these associations reflect component-specific effects beyond performance, we also report difficulty-controlled regressions of α on component weights with accuracy and operand magnitude controls.

Baseline-model positioning

Our analyses are positioned against two baseline families. First, simpler item-difficulty baselines treat RT as a function of fixed item effects or operand-magnitude proxies. Our controls, namely accuracy, exposure count, and operand magnitude, approximate this family without adding new fitted baselines. Second, knowledge-tracing and performance-tracing models estimate latent learning trajectories over time

Condition	Accuracy	Median RT (s)	Mean RT (s)
L1 first	0.851	–	–
L1 middle	0.824	–	–
L1 last	0.922	–	–
L2 first	0.888	3.10	8.82
L2 middle	0.909	2.77	7.73
L2 last	0.930	2.91	7.63
L3 first	0.817	2.54	3.21
L3 middle	0.847	2.31	2.94
L3 last	0.858	2.35	3.13

Table 1: Baseline descriptives by level and encounter position, reported as item-level means. RTs are available only in Levels 2–3. Level 3 RTs are right-censored by an 8-second deadline.

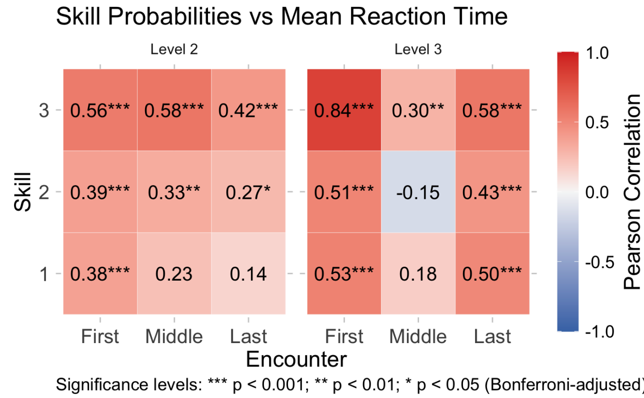

Figure 2: Pearson correlations between item component weights ($k = 3$) and mean response time across Level 2 and Level 3 encounter positions. Asterisks indicate Bonferroni-adjusted significance.

(Abdelrahman et al., 2023; van Rijn et al., 2009; Sense et al., 2016), whereas GraafTel emphasizes item-component structure. We therefore do not claim numeric superiority here. A concrete next step is a held-out benchmark comparing GraafTel-based features against (i) fixed-item or operand RT baselines and (ii) a standard tracing model with item identity.

Results

Baseline learning dynamics in performance

Before linking skills to strategy-sensitive measures, we summarize what learning looks like in the raw traces. Accuracy tends to increase from first to last encounter within each level, whereas RTs in Levels 2–3 decrease modestly across encounters (Table 1). Mean RTs are substantially larger than medians in Level 2, reflecting a heavy right tail. This motivates the use of log-mean and trimmed-mean RT summaries in the robustness analyses. These descriptives provide context for interpreting the hypothesis tests.

Stage-dependent RT profiles of latent components

Testing H1 and H2, Figure 2 shows correlations between item component weights, reported for the $k = 3$ solution, and mean RT across levels and encounter positions. One component, Component 3, the RT-linked "slow" component by our definition, exhibits consistently strong positive correlations with RT. This pattern suggests that items with high loading on this component remain time-demanding across stages. In contrast, the remaining components show weaker and more stage-dependent associations, including attenuated correlations in later encounters. This pattern is consistent with cognitive accounts in which reliance on calculation-like processing decreases and retrieval or automaticity becomes more prominent with practice (Siegler, 1988; Tenison et al., 2016), but the present results do not directly identify strategies.

Controlling for difficulty and practice. The RT-linked component remains the largest positive predictor of log-mean RT after controlling for mean accuracy and an operand-magnitude proxy across all levels and encounter positions (Table 3). To test whether component–RT associations reduce to standard item-level proxies, we fit item-level regressions predicting log-mean RT while controlling for mean accuracy and the sum of operands. The key finding persists after these controls, and adding components improves prediction beyond the baseline controls (Table 5). Thus, the main signature is a stable association between one latent dimension and RT, even after accounting for accuracy, exposure, and operand size.

For the coefficient table, we report a compact specification with mean accuracy and operand magnitude. The incremental-value analysis uses the fuller baseline of mean accuracy, exposure count, and operand magnitude.

This pattern is robust to alternative RT summaries: Component 3 remains positive for median RT, log-mean RT, and top-1% trimmed mean (Table 2). Adding components to baseline controls, namely accuracy, exposure, and operand magnitude, improves prediction across all conditions ($\Delta R^2 = .05$–$.17$; Table 5).

As a stage-sensitivity test, we fit difficulty-controlled item-level models with Component×Stage interactions, where Stage is first/middle/last encounter position within level. Interaction terms for Component 3 are small across levels, with standardized β for middle–first and last–first contrasts ranging from +0.04 to +0.10, which is consistent with a relatively stable RT association across stages.

Cross-k stability of the RT-linked component

To show that the RT signature is not an artifact of fixing $k = 3$, we apply the RT-linked component identification rule described in Methods for each $k \in \{3, 4, 5, 6\}$. Across k, the identified component shows a strong positive association with log-mean RT across levels and encounter positions (Table 4). Mean incremental predictive value beyond baseline difficulty and practice proxies is similar across k, with $\Delta R^2 \approx .11$–

Condition	Mean RT r	Median RT r	Log-mean RT r	Trim top 1% r
L2 first	+0.33 [+0.18,+0.59]	+0.68 [+0.60,+0.75]	+0.74 [+0.66,+0.81]	+0.70 [+0.61,+0.77]
L2 middle	+0.32 [+0.20,+0.65]	+0.71 [+0.64,+0.78]	+0.78 [+0.72,+0.84]	+0.72 [+0.65,+0.79]
L2 last	+0.45 [+0.35,+0.63]	+0.71 [+0.65,+0.78]	+0.75 [+0.68,+0.81]	+0.65 [+0.57,+0.73]
L3 first	+0.83 [+0.79,+0.87]	+0.79 [+0.74,+0.83]	+0.83 [+0.79,+0.87]	+0.84 [+0.80,+0.87]
L3 middle	+0.83 [+0.78,+0.87]	+0.81 [+0.75,+0.86]	+0.81 [+0.76,+0.87]	+0.83 [+0.78,+0.88]
L3 last	+0.49 [+0.13,+0.87]	+0.80 [+0.75,+0.85]	+0.81 [+0.76,+0.86]	+0.84 [+0.80,+0.88]

Table 2: Robustness of the Component 3–RT association, that is, the RT-linked "slow" component in the $k = 3$ solution, across RT summaries: mean, median, log-mean, and top-1% trimmed mean. Each cell reports Pearson r with bootstrap 95% confidence interval, computed at the item level within each level and encounter position.

Condition	β_{C1}	β_{C2}	β_{C3}	β_{acc}	β_{sum}
L2 first	+0.19	+0.11	+0.53	-0.28	-0.05
L2 middle	+0.17	+0.04	+0.61	-0.29	+0.11
L2 last	+0.17	+0.16	+0.50	-0.21	+0.28
L3 first	+0.09	+0.10	+0.41	-0.53	-0.07
L3 middle	+0.05	-0.04	+0.51	-0.53	+0.22
L3 last	+0.07	-0.02	+0.66	-0.28	+0.14

Table 3: Difficulty-controlled item-level regressions predicting log-mean RT from item component weights ($k = 3$) and controls, reported as standardized coefficients. Controls include mean accuracy and an operand-magnitude proxy given by the sum of operands.

.12. In difficulty-controlled models, the RT-linked component is also the largest positive component coefficient in all six level×encounter conditions for each k.

An RT-linked component correlates with system-estimated forgetting

Testing H3, Figure 3 links early component weights from Level 1, shown for $k = 3$, to a later system-estimated forgetting parameter, Level 3 α. The RT-linked component, Component 3 in the $k = 3$ solution, shows moderate positive raw correlations with α across encounters. Items with higher loading on this component therefore tend to receive higher system-estimated forgetting values later. The remaining components show smaller and less consistent effects.

Because Level 3 updates to α incorporate RT in addition to accuracy, the Component–α association is potentially partly driven by shared dependence on RT and should not be interpreted as an independent forgetting signature. We therefore emphasize the difficulty-controlled regressions and incremental-value tests below as the primary evidence about whether components explain α beyond standard performance proxies.

Raw correlations show that the RT-linked component is positively associated with later decay rate α, with $r = .56/.61/.57$ for first, middle, and last encounters and all Bonferroni-adjusted $p < .001$. However, in difficulty-controlled regressions including mean accuracy and operand

k	Slow comp.	L2 first	L2 mid	L2 last	L3 first	L3 mid	L3 last	Mean ΔR^2
3	x_3	0.74	0.78	0.75	0.83	0.81	0.81	0.11
4	x_3	0.74	0.78	0.74	0.82	0.80	0.81	0.11
5	x_4	0.75	0.79	0.75	0.82	0.80	0.80	0.12
6	x_1	0.78	0.77	0.73	0.81	0.81	0.80	0.12

Table 4: Cross-k stability of the RT-linked, or "slow," component. For each k, the slow component is defined as the component with the largest positive correlation with log-mean RT in Level 3 middle encounter. Entries report Pearson r between that component's item weights and log-mean RT across level×encounter conditions. Mean ΔR^2 is averaged across the six conditions, comparing baseline models using accuracy, exposure count, and operand magnitude to models that additionally include all k components.

Condition	R^2 (baseline)	R^2 (skills+baseline)	ΔR^2
L2 first	0.562	0.693	0.131
L2 middle	0.605	0.753	0.148
L2 last	0.561	0.732	0.171
L3 first	0.833	0.881	0.048
L3 middle	0.777	0.827	0.050
L3 last	0.640	0.736	0.096

Table 5: Incremental predictive value of item component weights ($k = 3$) for log-mean RT beyond baseline difficulty and practice proxies. The baseline includes mean accuracy, exposure count, and operand magnitude, using sum and max. Entries report R^2 from item-level OLS within each level and encounter position.

magnitude, component coefficients are small (Table 6). This is consistent with much of the α association being mediated by performance rather than reflecting a component-specific forgetting signature.

Consistent with this interpretation, adding Level 1 components to baseline models of Level 3 α improves fit only marginally (Table 7; $\Delta R^2 = 0.00$–0.03).

Interpretability and operand-size checks

To make the latent dimensions more concrete, Table 8 lists example facts whose item–component weight is most *specific* to each component, using Level 2 averages ranked by $x_k - \max_{j \neq k} x_j$. Items can depend on multiple components simultaneously, and the table highlights facts where one dimension is dominant.

Qualitatively, the $k = 3$ exemplars suggest recognizable arithmetic structure. One component, Component 2, concentrates mass on facts in the $\times 10$ row or column, consistent with a "ten-based" dimension. Another, Component 3, the RT-linked component, includes multiple $\times 2$ and related small-operand facts. A third, Component 1, includes a mix of anchor and non-anchor facts, consistent with a broader dimension that is not captured by a single operand rule. Across k, comparable structure is typically present but may split or merge across components.

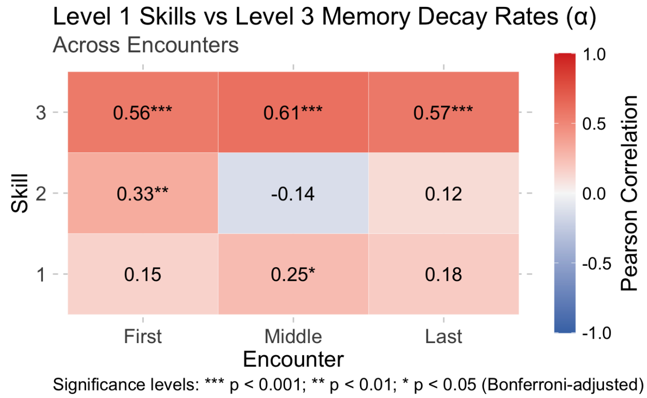

Figure 3: Pearson correlations between Level 1 item component weights ($k = 3$) and Level 3 item decay rate α across encounter positions. Asterisks indicate Bonferroni-adjusted significance.

Encounter (L3)	β_{S1}	β_{S2}	β_{S3}	β_{acc}	β_{sum}
first	+0.16	+0.01	+0.09	-0.76	+0.14
middle	+0.10	+0.03	-0.08	-0.90	+0.18
last	-0.09	+0.08	-0.05	-1.00	+0.23

Table 6: Difficulty-controlled item-level regressions predicting Level 3 decay rate α from Level 1 component weights ($k = 3$) and controls, reported as standardized coefficients.

As a minimal check against the alternative explanation that these weights simply encode operand magnitude, we correlated mean item component weights for $k = 3$ with operand-size proxies. The RT-linked component, Component 3, had weak correlations with max operand ($r = -.05$), sum of operands ($r = .13$), and product ($r = .12$), which provides little evidence that it is a simple operand-size axis. Component 2 showed a larger association with sum and product ($r \approx .35$), which is consistent with partially capturing difficulty or magnitude.

Discussion

The main finding is that GraafTel yields latent components whose relationships to RT and to a system-estimated forgetting parameter differ across learning stages. Across $k = 3$–6, an RT-linked component remains tightly coupled to slower responses, which is consistent with sustained reliance on calculation-like processing. The remaining components show weaker and more stage-dependent associations, which is consistent with increasing reliance on retrieval and emerging automaticity as practice proceeds (Siegler, 1988; Anderson, 1982; Tenison et al., 2016). The α results are secondary: they are compatible with this picture but are not independent evidence because Level 3 α updates partly depend on RT.

These dissociations matter for cognitive theory because they suggest that strategy transitions are fact-specific rather than uniform across the multiplication table. Consistent with overlapping-strategies accounts, not all facts transition at the same rate, and latent structure can highlight which facts appear to remain process-demanding even late in training (Siegler, 1988; Lemaire & Siegler, 1995).

For cognitive science more broadly, this approach offers a way to test theories of strategy mixture and transition in large-scale naturalistic data even when explicit strategy reports are unavailable. It therefore complements laboratory studies by adding ecological scale while retaining item-level resolution. For educational measurement, these markers could help distinguish globally slow learners from selectively process-demanding facts.

Encounter (L3)	R^2 (baseline)	R^2 (full)	ΔR^2
first	0.616	0.644	0.027
middle	0.778	0.780	0.002
last	0.789	0.811	0.022

Table 7: Incremental predictive value of Level 1 item component weights ($k = 3$) for Level 3 α beyond baseline proxies, namely mean accuracy, exposure count, and operand magnitude. Entries report R^2 from item-level OLS within each encounter position.

Component 1	Component 2	Component 3
1×7	10×10	1×8
9×1	6×10	8×2
3×1	2×10	5×2
10×7	5×10	2×8
10×1	7×10	1×6
2×4	8×10	4×2

Table 8: Example facts where one component is the dominant item dependency, using Level 2 averages across encounter positions and ranking by $x_k - \max_{j \neq k} x_j$. Components are unlabeled latent dimensions, and the exemplars provide a hypothesis for interpretation.

Limitations and future work

Latent components are inferred from performance and do not directly observe strategies. We therefore treat the RT-linked dimension as a *computational marker* consistent with calculation-like processing, pending external validation. RT in classroom deployments may reflect device speed, input speed, and time-pressure effects, including the Level 3 deadline. The system-estimated forgetting parameter α is model-derived and, in Level 3, partly RT-mediated, so Component–α associations do not provide independent evidence of forgetting dynamics.

Acknowledgments

Code and analysis scripts: `https://github.com/theodrosmhaile/times_table_tracing`

References

Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, *55*. doi: 10.1145/3569576

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological review*, *89*(4), 369.

Iancu, S. D., Braam, M., Mccabe, N., Wilschut, T., Rijn, H. V., & van der Velde, M. (2024). *An adaptive learning system for stepwise automatisation of multiplication facts in primary education* (Vol. 46).

Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization.* Retrieved from `https://arxiv.org/abs/1412.6980`

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: contributions to children's learning of multiplication. *Journal of experimental psychology: General*, *124*(1), 83.

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*. doi: 10.1111/tops.12183

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275. doi: 10.1037/0096-3445.117.3.258

Tenison, C., Fincham, J. M., & Anderson, J. R. (2016). Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive psychology*, *87*, 1–28.

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of iccm 2009 - 9th international conference on cognitive modeling.*