Fermi: AI-Assisted Practice for Measurable STEM Mastery at Scale

Evidence from student attempts (Aug–Dec 2025)

Executive summary

Students improve in STEM when they complete sustained, targeted practice—especially when they receive immediate, stepwise support that keeps them engaged through difficulty. Fermi is an AI-assisted practice environment built to scale this experience: it measures concept-level understanding, assigns practice questions, and provides scaffolding without turning into an answer-giving tool.

This whitepaper summarizes evidence from an Aug–Dec 2025 practice period combining:

1. Attempt records with concept-level grades (0–10) and coarse difficulty bands (1–3) (**79 students; 7,220 attempts; 15,047 concept-grade rows**),
2. Assistance logs capturing how many hints a student received per attempt, and
3. Qualitative roundtables with the same students and their parents, plus anonymized attempt conversation traces.

Key quantitative findings (observational):

- Sustained practice correlates with large within-student mastery gains: among students with ≥10 attempts (n=57), mean attempt mastery rose from **4.91 to 7.52** (+2.60); among students with ≥100 attempts (n=40), mastery rose from **4.97 to 7.79** (+2.82).
- The largest gains occur after initial struggle: for student × concept sequences with first grade ≤2 (n=347), the next exposure improved by **+3.50** on average, and first-to-last improved by **+4.68**.
- Students become more independent: among students with ≥10 attempts, hints/attempt declined from **1.71 to 1.35** (−21%), and **66.7%** show mastery up and hints down.
- Gains persist when holding difficulty band constant (difficulty=3: mean mastery gain **+3.02**, with hints/attempt declining on average).

Engagement distribution and drop-off:

- Engagement is deep for many students (median **100 attempts**; **50.6%** of students reached ≥100 attempts; **10.1%** reached ≥200).
- Early drop-off exists: **24.1%** of students completed only 1–5 attempts. Early leavers show lower mastery and more low-mastery attempts in their first five attempts, and they also use more hints—suggesting the main scaling constraint is the first-week experience (onboarding, friction, confidence, and return-after-failure), not absence of support.

Key qualitative findings (mechanisms + constraints):

- Students describe AI tutoring as a judgment-free space for repeated questions.
- Students and parents prefer a coach (stepwise guidance) over an answer machine.
- Trust hinges on accuracy; confident wrong answers harm adoption.
- Flow friction (writing, question ingestion) can negate benefits.
- Motivation is shaped by goals and deadlines, but students want autonomy.

1. The problem

Education systems face a structural mismatch: students need far more individualized, error-responsive practice than large classrooms can provide. This is especially true in STEM where progress depends on stepwise reasoning and frequent feedback. At the same time, general-purpose AI tools are widely used by students. Without guardrails, these tools can become answer machines that reduce learning and weaken academic integrity. Public systems need tools that increase practice volume without increasing teacher workload, reinforce learning rather than shortcutting it, provide measurable evidence of mastery growth, and operate safely under student privacy constraints.

2. Fermi's approach

Fermi is an AI-assisted learning platform built around a closed loop: measure current understanding (concept-level mastery), assign practice problems aligned to student needs, support with stepwise hints and guidance when the student struggles, update the learner model based on observed errors and assistance needed, and repeat to build mastery and confidence through volume.

Two principles define the system: practice is the primary learning mechanism (sustained problem solving rather than passive consumption), and the tutor is a coach, not an answer machine (preserving student thinking and enabling repeated questioning in a psychologically safe environment).

3. Evidence base and methods

3.1 Quantitative datasets

Attempt mastery (concept grades): Each attempt is tagged with one or more concepts, each with a grade on a 0–10 scale. We compute attempt-level mastery as the mean grade over concepts tagged in that attempt.

Assistance usage (hints per attempt): Each attempt has a hintCount, the number of hint/supervisor records associated with that attempt.

Cohort definition: We analyze a cohort of **79** high-school students preparing for the Joint Entrance Examination (JEE).

3.2 Core analyses

1) Within-student mastery growth: early vs late window per student (first 20% vs last 20%, minimum 5 attempts per window).
2) Struggle → improvement: for student×concept sequences with first grade ≤2, measure first→second and first→last changes.
3) Support usage: distribution of hints/attempt and relationship to mastery and difficulty.
4) Independence: early vs late hints/attempt per student, and mastery vs hint change quadrant.
5) Difficulty-controlled learning: repeat (1) and (4) within each difficulty band to test whether gains can be explained by easier problem selection.

3.3 Engagement distribution and survivorship bias

Several analyses in this paper focus on students with ≥10 attempts to enable longitudinal comparisons. This can introduce survivorship bias if early drop-off students have systematically different experiences. We therefore report engagement distribution across the full cohort (n=79). **24.1%** of students completed only **1–5 attempts**, while **72.2%** completed **≥10 attempts**, **50.6%** completed **≥100 attempts**, and **10.1%** completed **≥200 attempts** (median **100** attempts; max **307**). Among students who completed >5 attempts (n=60), median usage is **132** attempts.

Early leavers show a harder first-week experience: mean mastery over their first five attempts is **3.01/10** versus **4.55/10** for students who continue, and very low mastery (≤2/10) occurs more often in early leavers' first five attempts (**40.9%** vs **24.0%**). Early leavers also use more hints in their first five attempts (**2.92** vs **2.07** hints/attempt), indicating that drop-off is not explained by a lack of assistance alone. A policy-grade evaluation should therefore treat first-week retention and return-after-failure as first-class outcomes alongside mastery growth.

4. Results: practice leads to measurable mastery gains

4.1 Within-student mastery rises with sustained practice

Among students with **≥10 attempts** (n=**57**), mean attempt mastery increased from **4.91** to **7.52** (**+2.60**), and **96.5%** improved. Among students with **≥100 attempts** (n=**40**), mastery increased from **4.97** to **7.79** (**+2.82**), and **100%** improved.

Within-student mastery change (late − early), students with ≥10 attempts
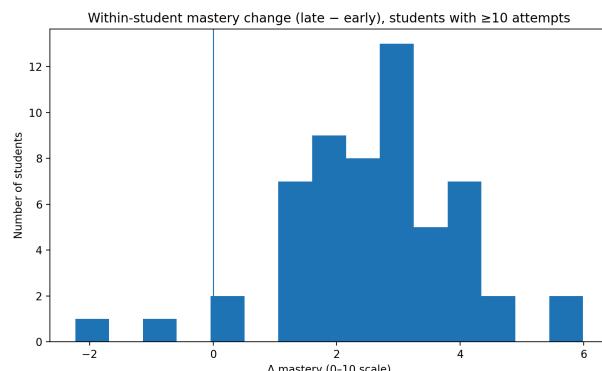
Figure 1. Within-student mastery change (late − early), students with ≥10 attempts.

4.2 Improvement occurs after struggle

For student × concept sequences where the first grade is **≤2** (n=**347**), the next exposure improved by **+3.50** on average, and **79.3%** improved immediately. First-to-last improvement averaged **+4.68**, and **87.0%** improved over time.
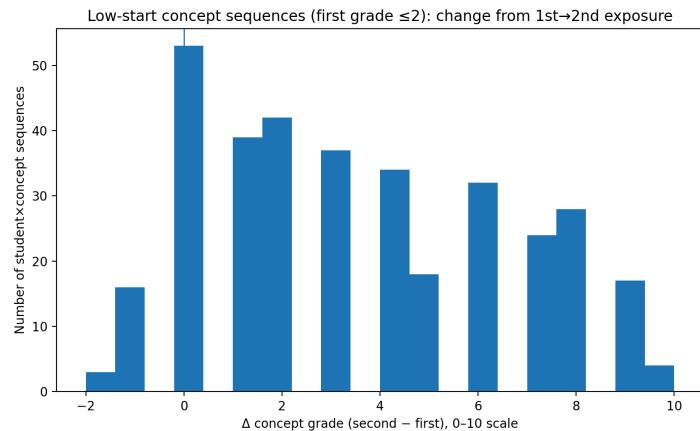

Figure 2. Low-start concept sequences (first grade ≤2): change from 1st to 2nd exposure.

5. Results: AI support enables learning without dependence

5.1 Students use support, and it scales with difficulty

Across all attempts, mean hints/attempt is **1.52** (median **1**). **39.2%** of attempts have **0** hints; **53.9%** have **1–4** hints; **6.9%** have **≥5** hints. Support usage increases with difficulty: mean hints/attempt is **0.86** at difficulty 1, **1.29** at difficulty 2, and **1.74** at difficulty 3 (zero-hint rates: **52.3%**, **43.9%**, **34.4%**, respectively).
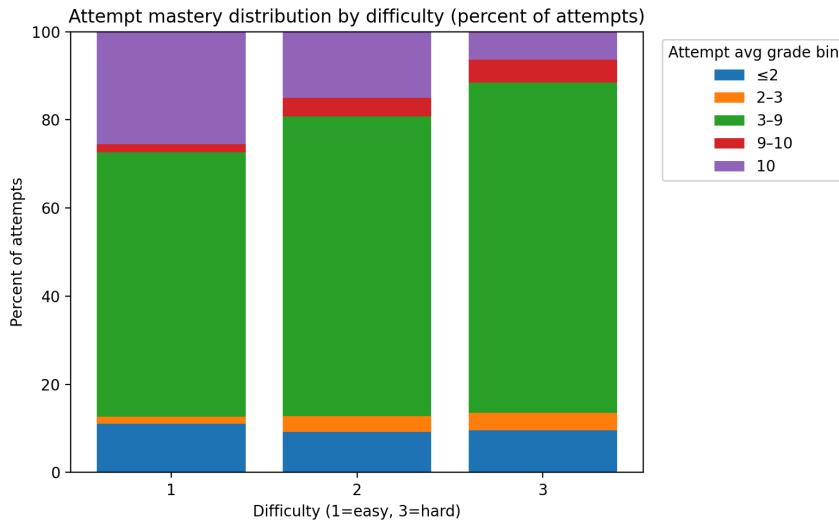
Figure 3. Attempt mastery distribution by difficulty (percent of attempts).

5.2 Over time, students need less help while mastery rises

Among students with **≥10 attempts** (n=**57**), hints/attempt declined from **1.71** to **1.35** (−21.0%), and **66.7%** show mastery up and hints down (68.4% reduced hints overall). Among students with **≥100 attempts** (n=**40**), hints/attempt declined from **1.70** to **1.18** (−30.6%), and **75%** show mastery up and hints down.
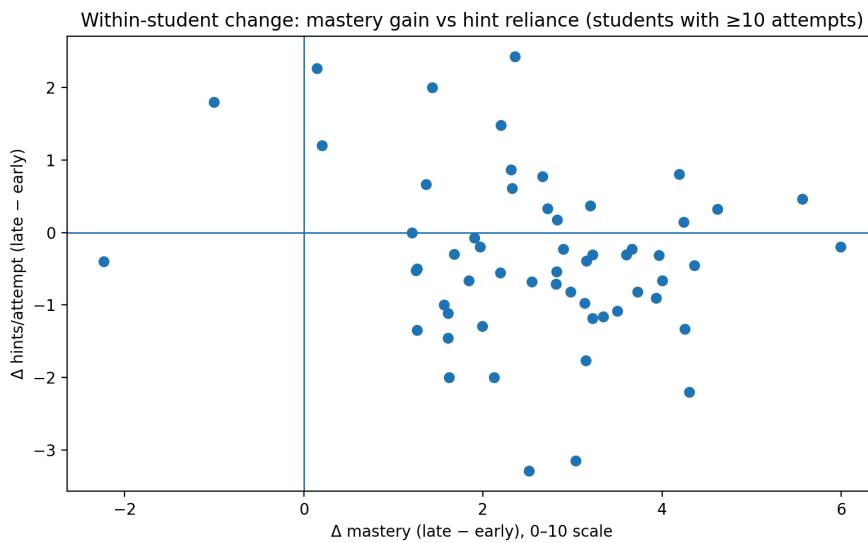

Figure 4. Within-student change: mastery gain vs hint reliance (students with ≥10 attempts).

Hint usage is not a proxy for weakness alone. Moderate hint usage often supports productive struggle and successful completion
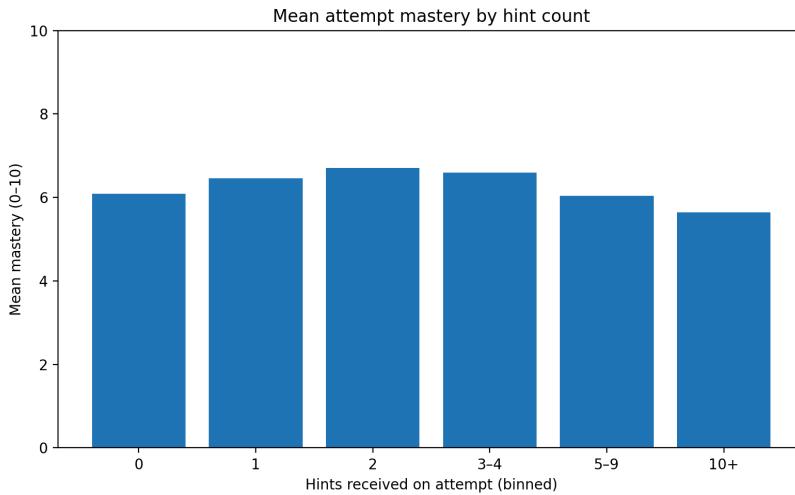
Figure 5. Mean attempt mastery (0–10) by hints received (binned). Moderate hint usage (1–4) coincides with higher average mastery than 0 hints or very high hint counts, consistent with scaffolding during productive struggle.

5.3 Support is concentrated in productive struggle attempts

Hint usage is highest in mid-band mastery attempts (3–9): mean **1.77** hints/attempt, compared to **0.78** hints/attempt when mastery is ≤2 and **0.78** when mastery is 10—consistent with scaffolding being used most during productive struggle rather than on trivial wins or complete failures.

Attempts with moderate hints have far fewer "complete failure" outcomes: the ≤2/10 rate drops from ~13% at 0–1 hints to ~3–5% at 2–9 hints (observational; not causal).
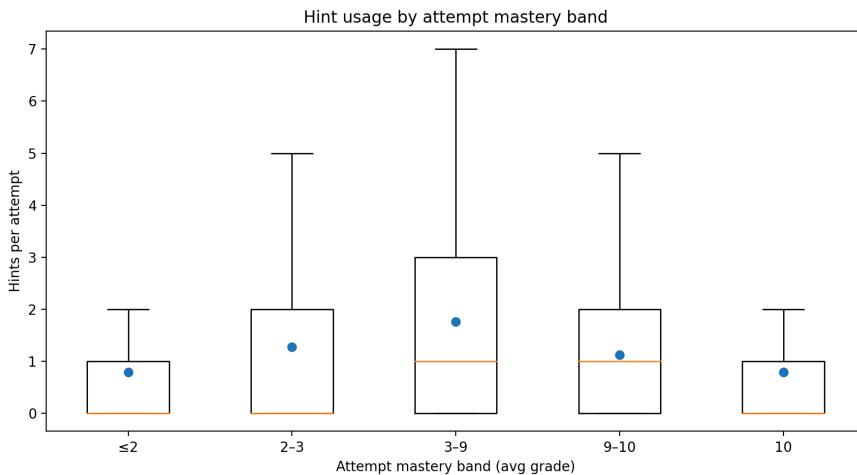
6. Results: gains persist when controlling for difficulty

A common alternative explanation for mastery improvement is that students switched to easier questions later. We address this by recomputing early-vs-late mastery changes within each difficulty band (globalDifficulty 1–3). Mastery gains remain positive within each band, including the hardest band: among students with ≥10 difficulty-3 attempts (n=47), mean mastery gain is **+3.02**, and hints/attempt declines by **0.33** on average.
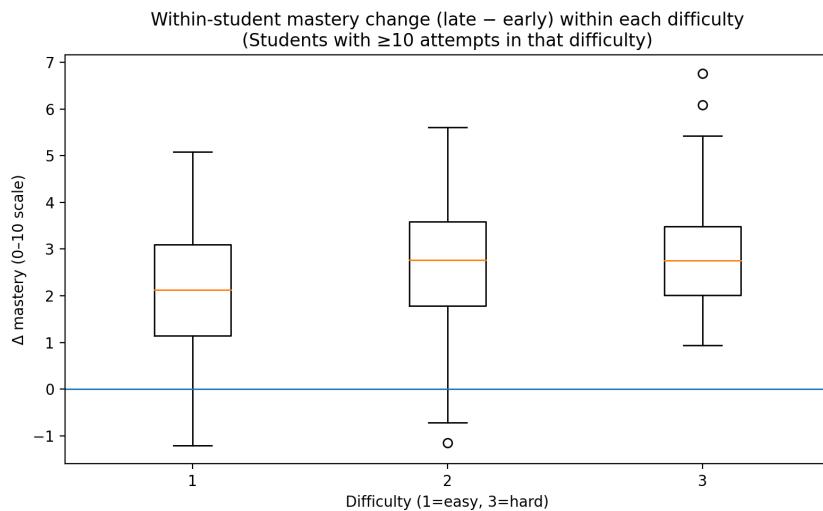

Figure 6. Within-student mastery change (late − early) within each difficulty band (students with ≥10 attempts in band).
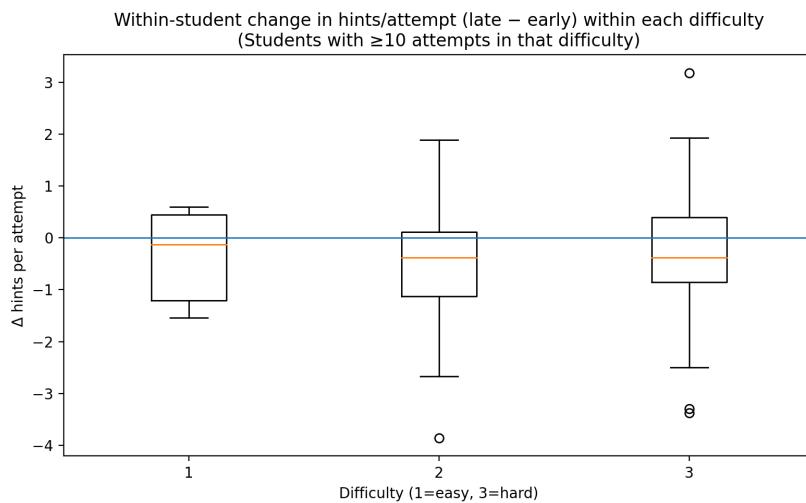

Figure 7. Within-student change in hints/attempt (late − early) within each difficulty band (students with ≥10 attempts in band).
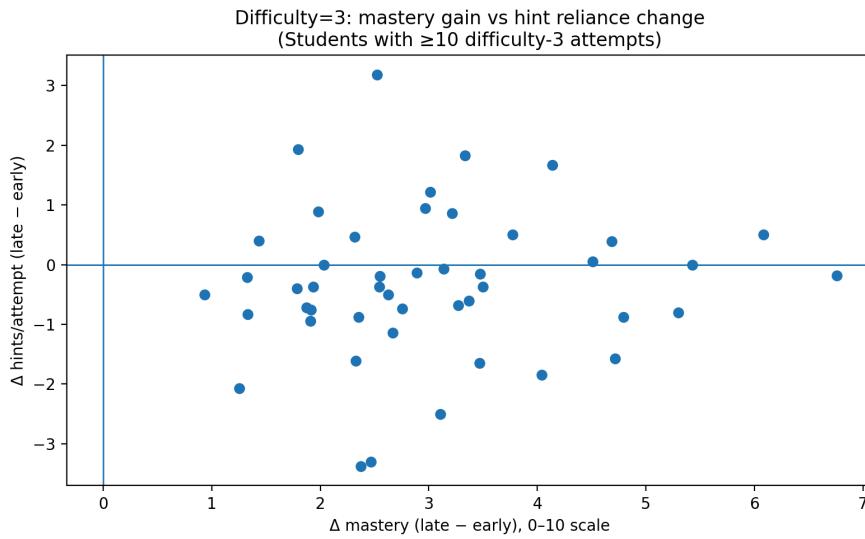
Figure 8. Difficulty=3: mastery gain vs hint reliance change (students with ≥10 difficulty-3 attempts).

7. Qualitative findings: what explains persistence and adoption

Roundtables with students and parents highlight mechanisms and constraints that complement the quantitative findings. Students describe AI tutoring as a judgment-free space for repeated questions, enabling persistence. Students and parents prefer stepwise coaching over direct final answers. Trust hinges on accuracy; confident wrong answers harm adoption. Flow friction (writing experience and question ingestion) can negate benefits even when pedagogy is strong. Motivation is influenced by goals and deadlines, but students want autonomy.

8. Limitations

This evidence is observational: it shows strong within-student improvement patterns but does not isolate causality. Mastery grades are platform measures, not externally administered standardized tests. HintCount measures quantity of support but not hint type, timing, or perceived helpfulness. Difficulty is coarse (1–3): it is useful for control but not a full item-response model. Finally, early drop-off is meaningful (24.1% of students stop after 1–5 attempts); results for sustained users may not generalize without improving the first-week experience.

9. Recommendations for policy-grade evaluation

To validate impact for wider adoption, we are running 12–16 week controlled pilots with cluster randomization. Primary outcomes to measure include externally administered assessment gains, curriculum-aligned mastery growth, and equity effects by baseline proficiency. Secondary outcomes

to measure include return after failure, time-to-solve and abandon rates, hint usage patterns as an independence signal, and teacher workload measures.

Appendix: Concrete Examples of Student Struggle and Tutor Scaffolding

This appendix addresses a common question from: what does 'stepwise hinting' look like in practice? Below are two anonymized interaction traces drawn from real student attempts. Student identifiers are removed. To keep the appendix readable, excerpts are lightly trimmed; the ordering reflects the recorded sequence of turns.

Example A — Micro-error correction in physics (6 minutes)
(Physics; difficulty 2; mastery 9.5; 10 hints)

Timeline (mm:ss)

- 00:00 — Student asks for the next step.

- 00:09 — Tutor prompts asking to compute time-of-flight first.

- 00:41 — Student requests validation; persists.

- 00:54 — Tutor prompts the plan again.

- 02:05 — Tutor confirms time and points out a decimal slip in the vertical drop.

- 05:39 — Student requests validation.

- 05:55 — Tutor confirms the corrected drop (~6 m) and reinforces the method.

What this illustrates
A student had the right approach but was blocked by a single arithmetic/decimal error. Stepwise hints unlocked progress without giving away the entire solution path.

Appendix Sidebar: Example B — Persistence under frustration in physics (17 minutes)
(Physics; difficulty 3; mastery 8.25; 8 hints)

Timeline (mm:ss)

- 00:00 — Student asks tutor to check work.

- 00:46 — Tutor confirms setup, flags an error.

- 02:33 — Student insists alternate method is valid; frustration increases.

- 03:03 — Tutor explains why canceling a common denominator term-by-term changes the value; asks for revision.

- 06:02 — Student asks tutor to check work.

- 06:16 — Tutor confirms the form written matches the standard approach.

- 16:17 — Student requests validation. Persists.

- 16:48 — Tutor confirms correct velocities and final energy loss result.

What this illustrates
The tutoring remains calm and rigorous even when the student is irritated. This supports the "judgment-free" mechanism: students continue engaging through confusion rather than disengaging.

Hint usage by attempt mastery band
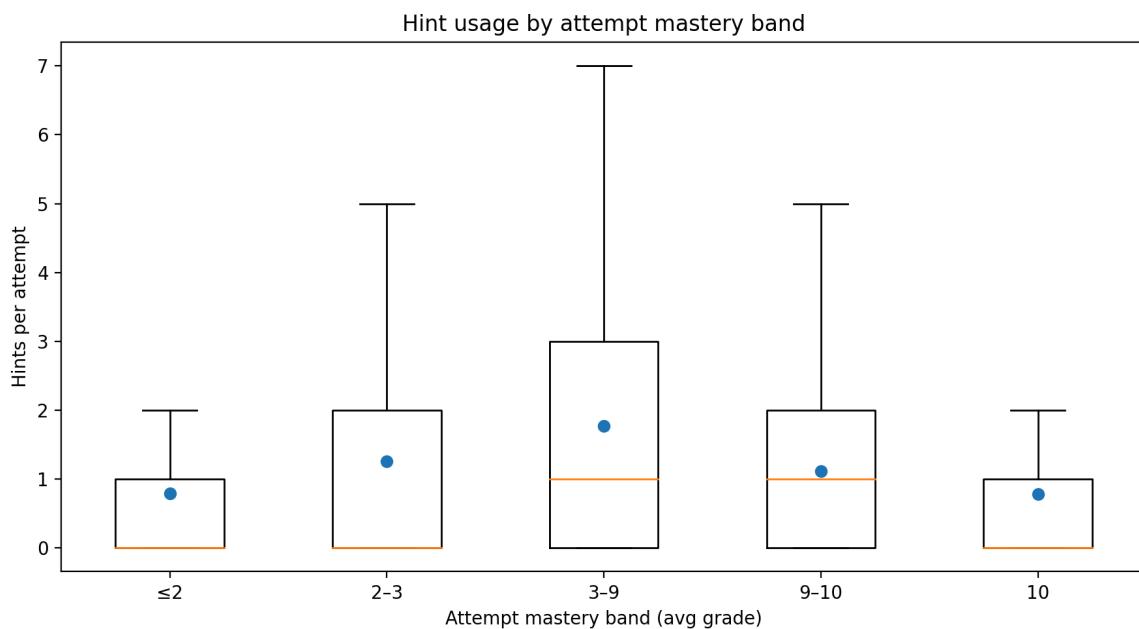
Low-start concept sequences (first grade ≤2): change from 1st→2nd exposure
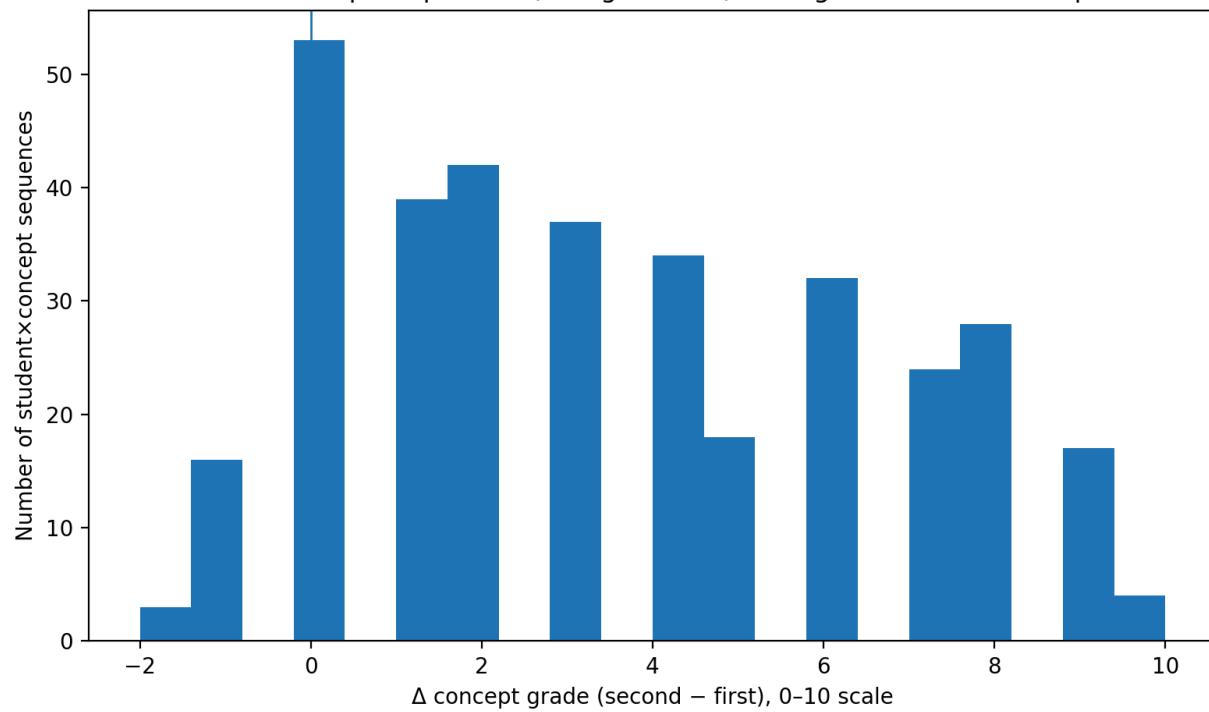
Attempt mastery distribution by difficulty (percent of attempts)
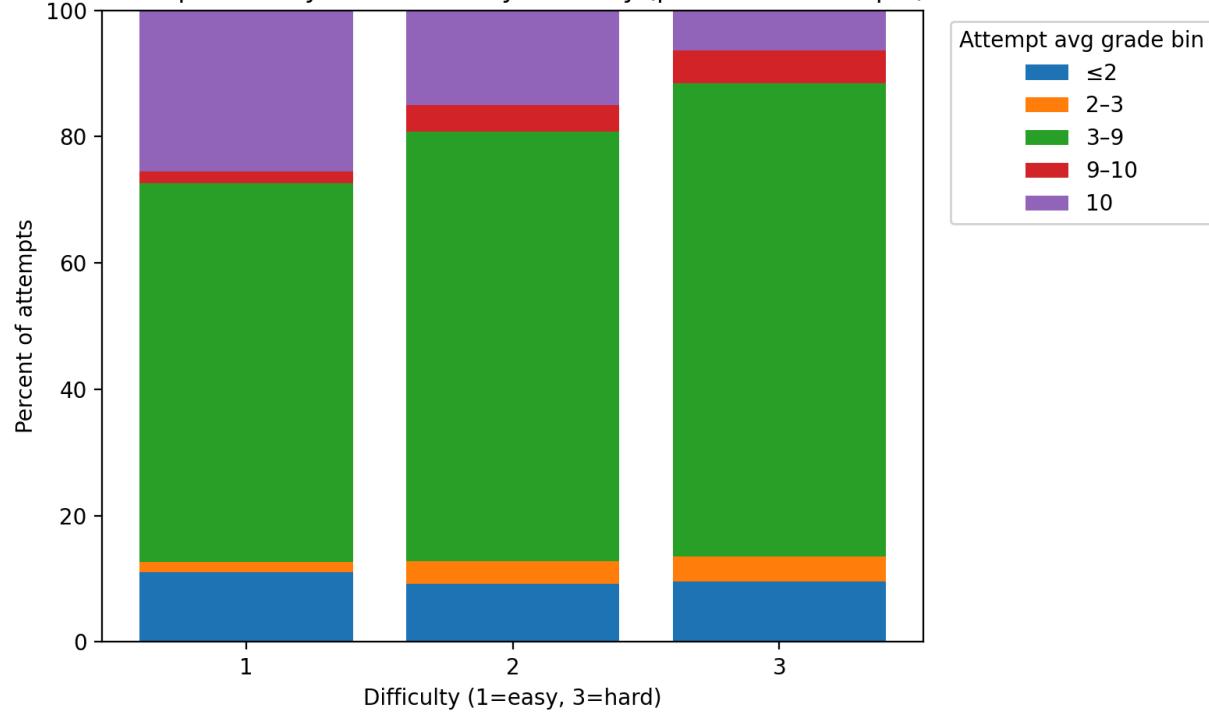
Example learning curve (Student A), rolling mean of mastery
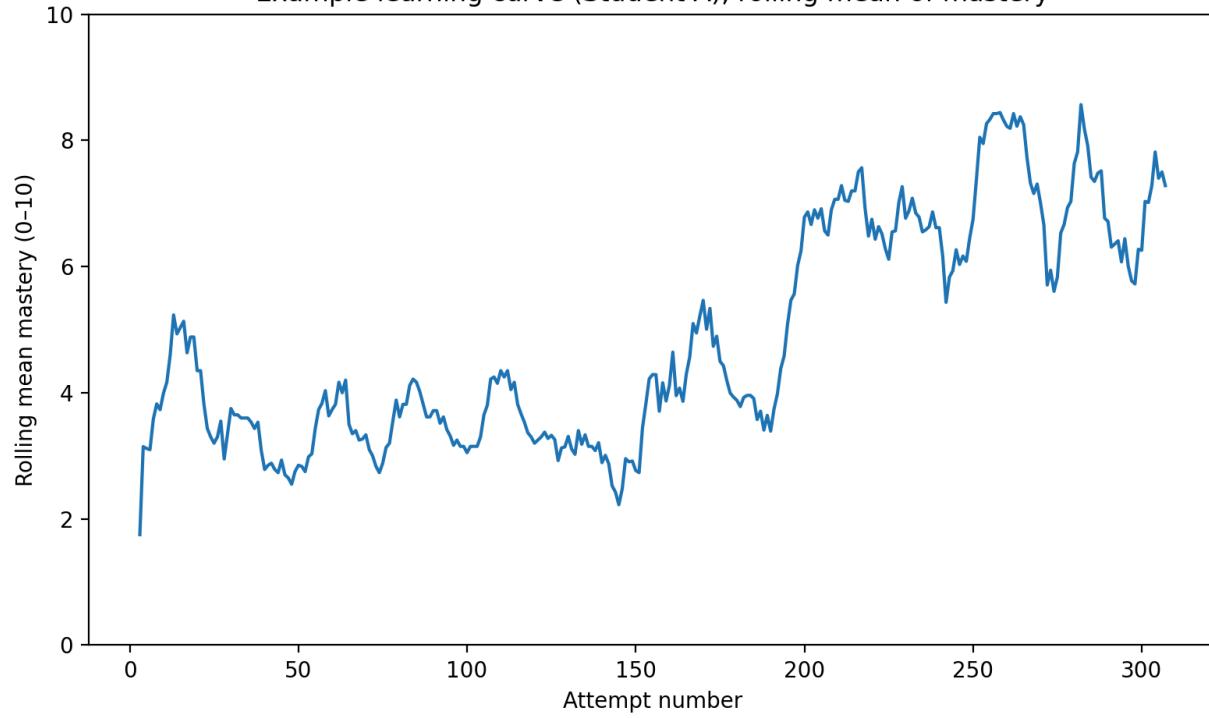
Example learning curve (Student B), rolling mean of mastery
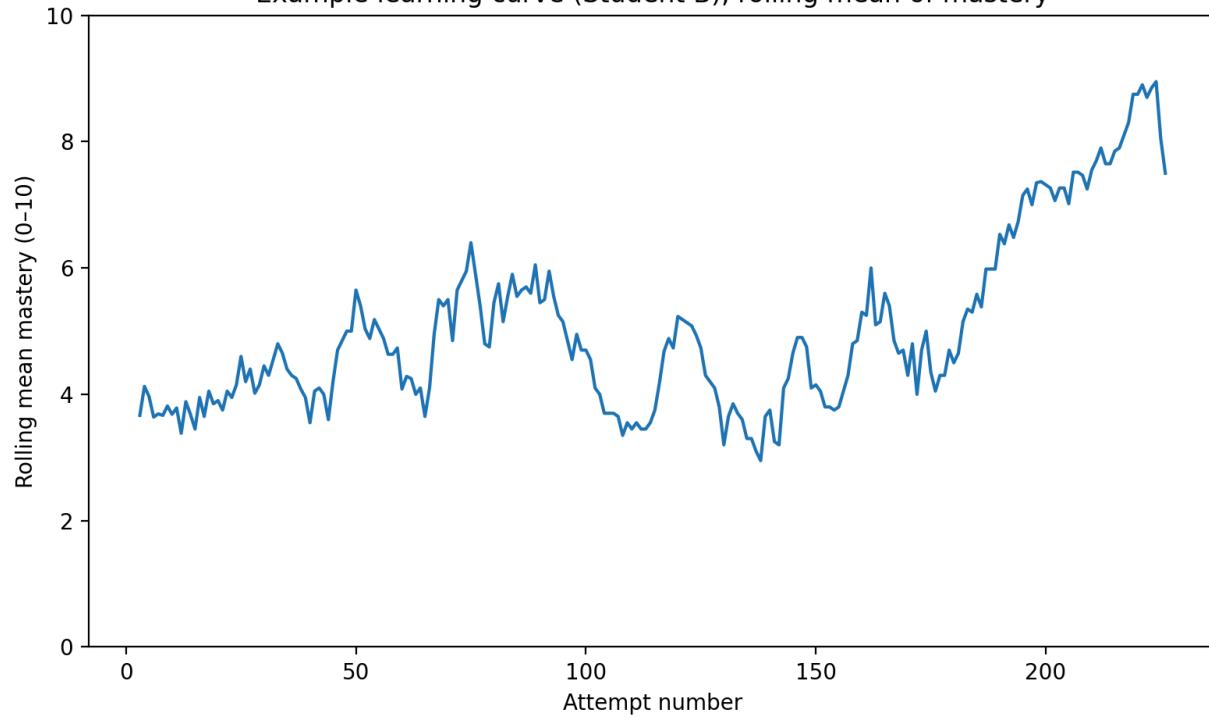
Example learning curve (Student C), rolling mean of mastery
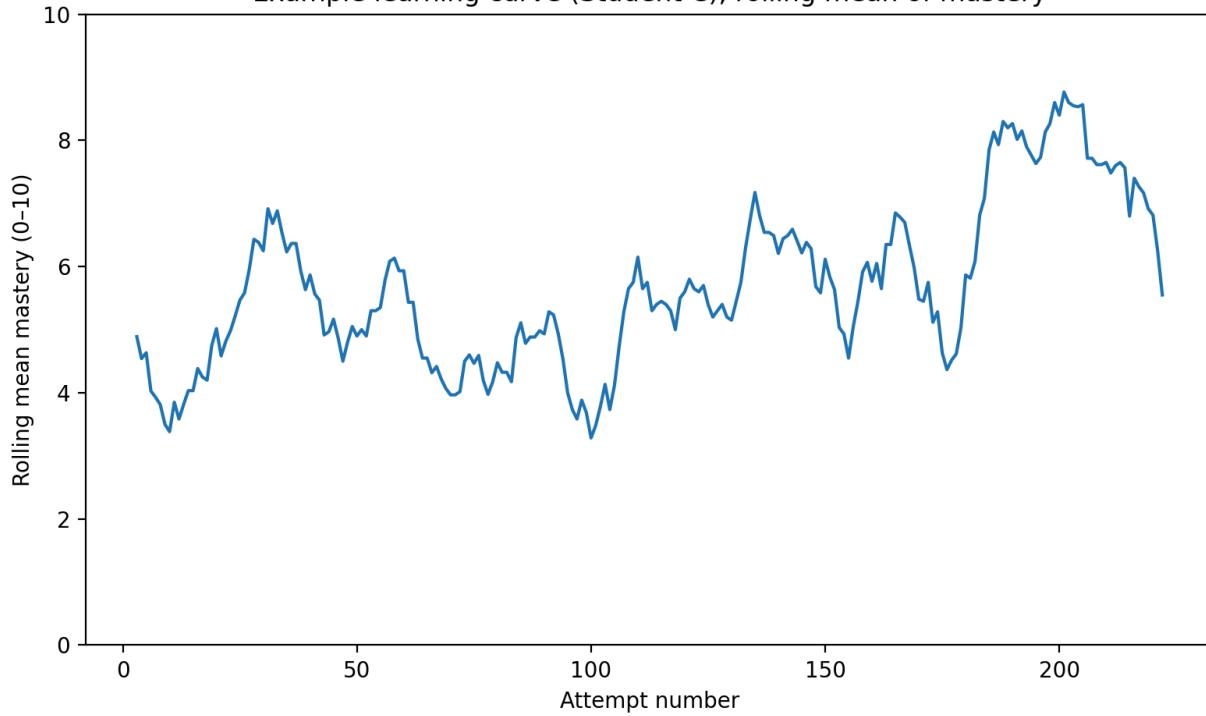
Example learning curve (Student D), rolling mean of mastery
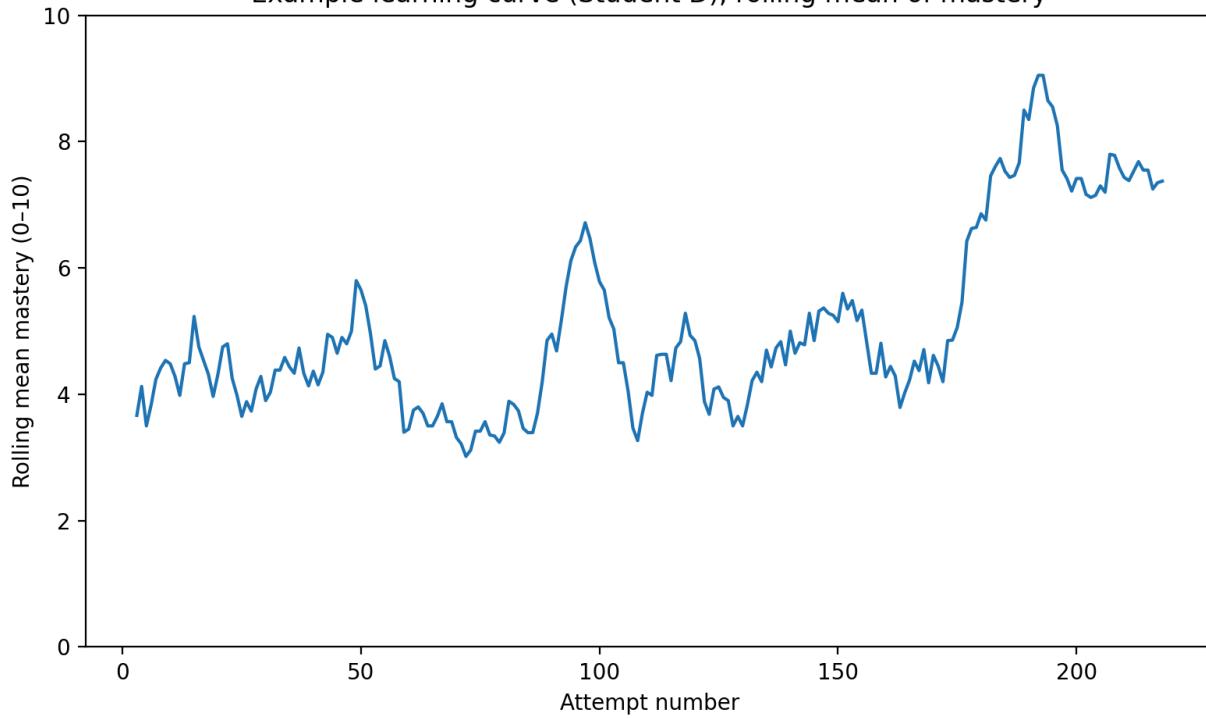
Within-student change: mastery gain vs hint reliance (students with ≥10 attempts)
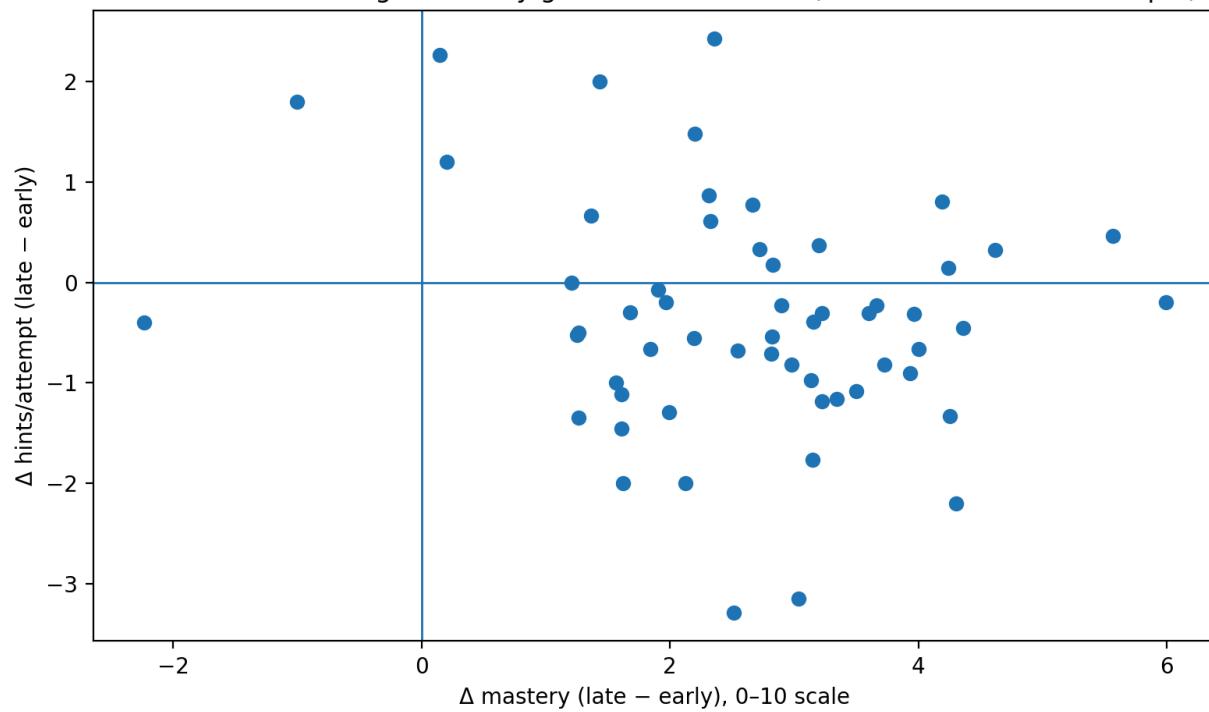
Mean attempt mastery by hint count

Rate of very low mastery (≤2/10) by hint count

Example student: mastery and hints/attempt over time (rolling means)

Within-student mastery change (late − early) within each difficulty
(Students with ≥10 attempts in that difficulty)

Within-student change in hints/attempt (late − early) within each difficulty
(Students with ≥10 attempts in that difficulty)

Difficulty=3: mastery gain vs hint reliance change
(Students with ≥10 difficulty-3 attempts)