Quiz errors do not rise with lesson position in mobile micro-learning: a preregistered analysis of 39,867 outcomes

Dr. Philip Sergelius[1], Dr. Max Hänze[1]

[1] NerdSip, Reinbek, Germany. Correspondence: philip@nerdsip.com

Preregistration: https://osf.io/8xnga · Data freeze: 2026-07-20 · Brief research report, 4 September 2026

Abstract. Many micro-learning products are built on the belief that attention fades across a course, so important material should come first. We preregistered a test of this belief and ran it on data from a mobile learning app. A course on this app is 3 to 10 short lessons on one topic, taken in order, often over several days. Each lesson is about 130 words shown a screen at a time, followed by one easy three-option question that must be answered correctly to move on. Missing that question on the first try is our outcome. The dataset holds 39,867 outcomes from 2,590 learners, 994 courses and 5,169 questions. A lesson's position in its course did not predict errors (odds ratio 0.978 per position, 95% CI 0.936 to 1.021). The last lesson of a course was missed as often as the first, and the interval rules out an effect large enough to justify front-loading. Learners completed each successive lesson about 1.5% faster with no detected loss of accuracy. Errors rose slightly with depth inside a single sitting (OR 1.026), but the effect is smaller than the smallest we registered as meaningful. Front-loading material within a short course is not supported by these data.

Keywords: micro-learning; mobile learning; learning analytics; item position effects; preregistration

Plain-language summary. Course designers are told to put the important material first, because people supposedly pay less attention as a course goes on. We checked this on 39,867 quiz answers from a phone learning app. People got the last lesson of a course right as often as the first one. They also got a little faster as they went, without making more mistakes. Putting the key material first does not help here.

1. Introduction

Micro-learning breaks instruction into pieces small enough to fit into a spare minute: a short text, one question, a session on a phone in a queue [1]. On the platform studied here those pieces are grouped into courses of 3 to 10 lessons on one topic, taken in a fixed order. A common design rule for such products is to put the most important material first, on the reasoning that attention drops as a course goes on. If that reasoning is wrong, the rule costs content for nothing.

The closest evidence comes from tests and exams. Position effects on accuracy are real but small per item. Using admission records from roughly 15 million Brazilian students with random question order, Reyes [2] finds that one extra position lowers the chance of a correct answer by 0.08 percentage points, an error odds ratio of about 1.004 per item. Low-stakes assessments give similar magnitudes, between 1.005 and 1.017 per item [3,4]. Vida, Bolsinova and Brinkhuis [5] analysed 599,519 responses from 90

university exams with randomized item order. Students answered later items faster, and position did not help predict accuracy. In an exam, however, a question's position, its depth in the sitting and the length of the test are one and the same thing.

In a mobile app they come apart. A learner may take lesson 1 on Monday and lesson 5 on Thursday, or twenty lessons across four courses in one evening. We preregistered three hypotheses on the Open Science Framework before extracting any data [6,7] (registration https://osf.io/8xnga, accepted 2026-07-20 at 20:13 UTC; extraction began at 20:46 UTC the same evening): that errors rise with a lesson's position in its course, that they rise with depth in a sitting, and that time per lesson changes with position. The questions in this app are comprehension checks, not tests: three options, easy distractors, unlimited retries, and a 4.4% base error rate. We therefore call the outcome a quiz error and do not claim to have measured attention.

2. Methods

2.1 The app and what it records

NerdSip is a mobile learning app (iOS and Android). A course contains 3 to 10 lessons. In the analysed sample nearly all courses have 3, 5, 7 or 10. A lesson is a short text (median 134 words) shown one paragraph-sized screen at a time, followed by a one-sentence takeaway and one three-option question with an explanation. Retries are unlimited and never block progress, though a retried question earns less experience.

The question is a gate: a lesson cannot be completed until it is answered correctly. About one completed lesson in ten has no answer on record, because the app also offers a way to finish a lesson without the question; those lessons are not in the sample.

2.2 Observation window and two changes during it

Per-lesson completion timestamps were added on 2026-05-24. Timed measures therefore cover 2026-05-24 to the freeze on 2026-07-20. The position analysis also uses outcomes recorded before that date, which have an outcome but no timestamp; these are 41.5% of the sample. The same release also stopped learners from answering questions they had skipped more than 14 days earlier. The error rate in the analysed sample is 3.75% before that date and 4.82% after it (odds ratio 1.75). Every model therefore carries an indicator for this change, and we report what it does. A second change, a swipe-card reading interface, reached the app stores on 2026-07-06. It covers only the final two weeks and only updated devices, and weekly error rates show no step at that date.

2.3 Sample

Of 4,209 courses, 2,674 were eligible: 3 to 10 lessons, excluding "deep-dive" follow-on courses whose first lesson begins mid-session by construction. We excluded administrator accounts and each course's creator on their own course. The sample is every first quiz outcome per learner and question in eligible courses: **39,867 outcomes from 2,590 learners, 994 courses and 5,169 questions**, 84.2% in English and 15.8% in German. One 9-lesson course with nine outcomes from one learner is left out of all displays. Three staff-written tutorial courses whose aggregate error rates were seen before registration provide 8.0% of outcomes; they stay in by preregistered decision, with a check that excludes them.

Learner identifiers were replaced with salted pseudonyms on the server before extraction; no names, e-mail addresses, lesson texts or question texts were read. The records keep creation time, timezone and tier, so they are pseudonymous personal data under the GDPR. The platform's privacy policy lists completed lessons, quiz results and time spent among the data stored under the user contract, and names legitimate interest as a basis for internal analysis; it does not mention research or publication specifically. This analysis rests on legitimate interest (Art. 6(1)(f) GDPR) in aggregate product research, and only aggregates are published. No institutional review board oversees this industry study. Every displayed cell contains at least 10 distinct learners, enforced in the figure code.

2.4 Measures

Quiz error is 1 if the learner ever answered the question wrong and 0 if the first answer was correct. **Position** is the lesson's order in its course, from 1. A **sitting** is a run of lessons that a learner completes with no break longer than 30 minutes between any two of them. It counts lessons across all courses, not just one. **Sitting depth** is how many lessons deep the learner is in that run: the first lesson of a sitting is depth 1, the next is depth 2, and so on. A break longer than 30 minutes ends the sitting and the count starts again at 1. A learner who finishes four lessons of one course and then three of another without a long break reaches depth 7. Only lessons with a timestamp can be placed in a sitting, so depth is defined for the timed part of the data. The 30-minute cutoff was preregistered; cutoffs of 15 and 60 minutes give the same depth result (OR 1.029 and 1.025 against 1.026). The **interval** is the time since the learner's previous completion within a sitting. It is not reading time. The timestamp is written when the question is answered correctly, so the interval includes reading, answering, retries and idling. A wrong answer therefore lengthens its own interval (median 72.0 s against 66.8 s), which makes any same-lesson test of pace against accuracy circular; the registered version of that test is reported in the supplement and not interpreted. Intervals at position 1 mostly contain a switch between courses (median 150.9 s against 65.0 s) and are flagged and controlled. In the interface used for most of the window, a correct answer also started a ten-second timer that finished the lesson if the learner did not tap first, so some intervals carry up to ten extra seconds.

2.5 Analysis

Confirmatory models are mixed-effects regressions with random effects for learner, course and question, fitted in R 4.6.1 with lme4 [8,9]. Following the preregistration, a confirmatory claim needs two-sided $p < .005$ with the predicted sign [10], and the smallest effect we treat as meaningful is an odds ratio of 1.10 per position step. Because a few learners contribute a large share of the data (the heaviest 1% hold 24% of outcomes), we refit the main models without the heaviest learners and report the range. For the timing model we also refit after dropping intervals shorter than 15 seconds, too fast to read a lesson, and longer than 10 minutes, which are idle time. Descriptive intervals are Wilson intervals [11]. All registered checks, including one that could not be run because the extract lacks the needed field, and the full deviation log are in the supplement.

3. Results

3.1 Lesson position does not predict errors

The overall error rate was 4.37% (1,744 of 39,867). An answer was on record for 88.5% to 92.5% of completed lessons at every position; the rest were finished without the question and are not in the sample. Figure 1 shows the error rate by position, one line per course length. No line rises.

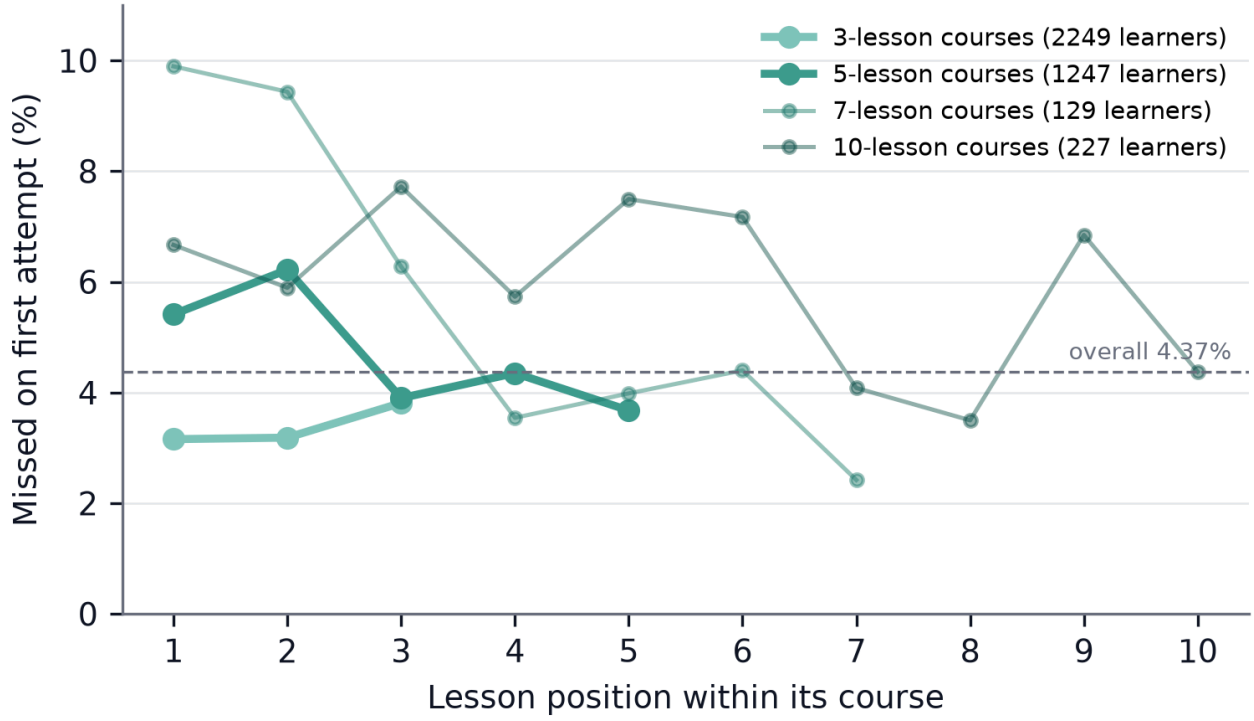

Figure 1. Share of questions missed on the first attempt, by the lesson's position in its course, one line per course length. Within each length the line is flat. The lines sit at different heights because courses differ from one another (Section 3.4), not because of position. Every point rests on at least 10 distinct learners.

The model agrees. Error odds did not rise with position: OR 0.978 per position, 95% CI 0.936 to 1.021, p = .308. The interval excludes the smallest effect we registered as meaningful (1.10), and the achieved precision gives 80% power at α = .005 to detect an OR of 1.084 or larger. The result is the same without the tutorial courses (0.972), in English only (0.982), without each course's final lesson (1.007), for learners with at least three answers (0.978), and on the timed half of the data (0.978). A registered check for curvature found none (linear p = .97, quadratic p = .78). Adding the instrumentation indicator and subscription tier moves the estimate further from decay, not toward it (0.945, 95% CI 0.906 to 0.986). Removing any one of the ten heaviest learners leaves it between 0.937 and 0.953; removing the heaviest 5% gives 0.922 (p = .11). None of these reach the confirmatory bar, so we read them as a stable null, not as improvement across a course.

Two things keep this from being an equivalence claim. Compounded over a ten-lesson course, the upper bound of 1.021 allows an error rate of 5.22% at lesson 10 against 4.37% at lesson 1, a rise of 0.85 percentage points. And the per-item effects found in large exam studies, 1.004 to 1.017, all sit inside our interval. This study does not contradict them. It rules out decay of the size a front-loading rule would need. It is also a result about learners who keep going: 96.6% of learners continued to the next lesson

after a correct answer and 91.0% after a wrong one, so later positions hold slightly fewer of the learners who make mistakes. Finally, 42% of outcomes come from 3-lesson courses, where decay has little room; the 7- and 10-lesson lines in Figure 1 are flat too, but rest on 129 and 227 learners.

3.2 Learners speed up

The interval between completions shortens as a learner moves through a course. On the registered data it falls 2.2% per position (t = −7.15), controlling for course switches, word count and language. Part of that comes from intervals too short to contain any reading: intervals under 15 seconds rise from 1.8% of the total at position 2 to 4.5% to 6.1% at positions 7 to 10, which is learners tapping through rather than reading faster. Dropping intervals under 15 seconds and over 10 minutes gives 1.6% per position (95% CI 1.1% to 2.1%, t = −6.18); dropping everything under 30 seconds gives 1.3%. The effect does not depend on the fact that wrong answers lengthen intervals: it is unchanged with a covariate for whether the question was missed (t = −5.23) or with missed questions removed (t = −4.38), and it holds when each learner is given their own slope (t = −5.96). Among the 383 learners with enough timed lessons for an individual slope, 67% speed up. Accuracy meanwhile does not move (Section 3.1). These are two separate estimates, not a joint test, so we say acceleration with no detected change in accuracy, not acceleration without cost. The pattern resembles what Vida and colleagues [5] found inside timed exams, which suggests it is not specific to time pressure.

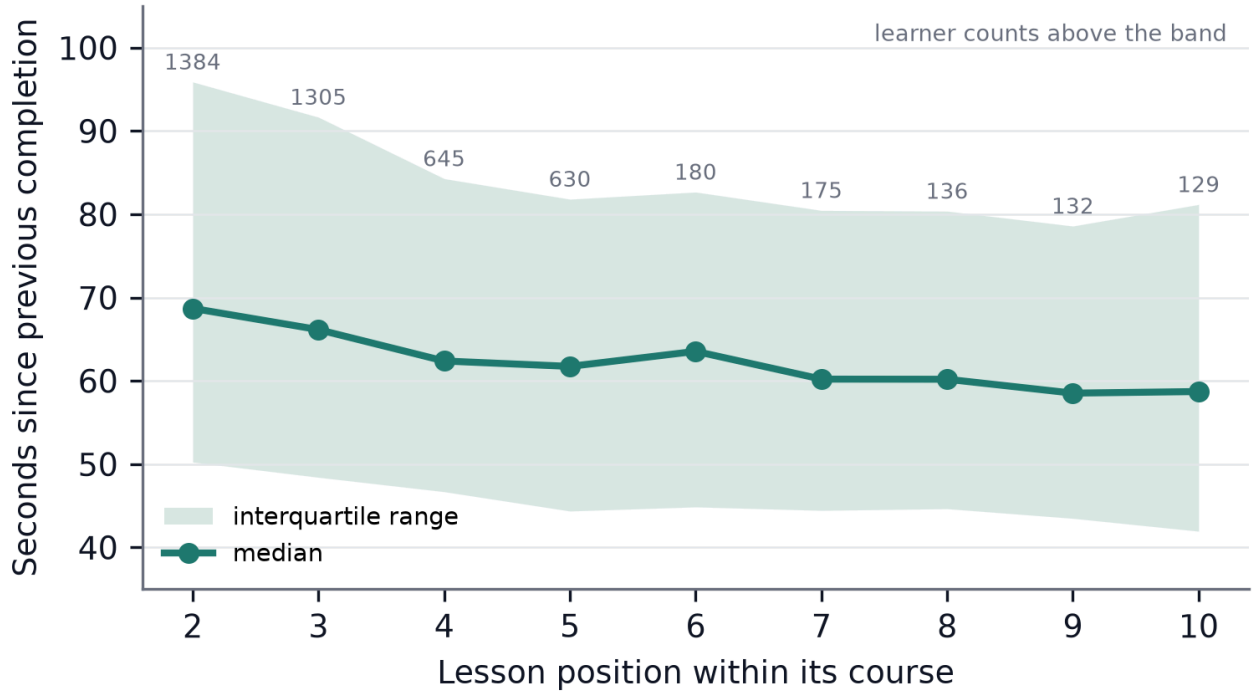

Figure 2. Median seconds between a lesson's completion and the previous completion in the same sitting, by the lesson's position in its course, with the interquartile range. Position 1 is not plotted: its interval is mostly a switch between courses (median 151 s). The interval includes answering the question and, for most of the window, up to ten seconds of an auto-advance timer. It is not a validated reading time. Every point rests on at least 10 distinct learners.

3.3 Depth within a sitting: detectable, but below the bar

Error odds rose with depth inside a sitting: OR 1.026 per lesson, 95% CI 1.008 to 1.043, p = .0036. This meets the registered p-threshold and holds within question (1.035) and within learner (1.028). It does not meet the registered size threshold: the whole interval lies below 1.10. A registered curvature check shows the rise is confined to the tail. The linear term is near zero and the quadratic term is positive (p = .0039), so errors are flat across the depths most learners reach and bend up only for very long sittings. That tail is thin. Only 65 learners ever reached fifteen lessons in one sitting, and such depths are 3.1% of timed outcomes. Removing any of three individual learners moves the p-value above .005, and a learner-weighted estimate has a wide interval (0.90 to 1.95). We report the association as detectable, smaller than meaningful, and resting on few people.

3.4 Courses differ, and course length is not the reason

The lines in Figure 1 sit at different heights: courses with more lessons have higher error rates, and the gap is already there at lesson 1, so it is not something that builds up over a course. It is also not length itself. Length is chosen by whoever generates the course, and the association lives between creators, not within them. Creators who make longer courses have higher error rates (OR 1.117 per lesson of a creator's average course length, p < .001), but within a creator's own courses, longer ones are not missed more often (OR 1.048, 95% CI 0.966 to 1.136, p = .26). Error rates also vary with when a course was generated, from 3.4% for courses made in March and April 2026 to 5.5% for those made in May to July, though controlling for that leaves the length estimate unchanged. The between-course variation in this dataset belongs to the courses and to who made them, not to position or length. Course length stays in the models as a control variable; it is not a finding.

Table 1. Headline estimates. Logistic models give odds ratios per unit. The timing model gives the change in the interval per lesson. All models carry random effects for learner, course and question.

Estimate	Value [95% CI]	p	Status
Lesson position → error odds	0.978 [0.936, 1.021]	.308	preregistered; not confirmed; interval excludes 1.10
Position → interval (trimmed)	−1.6% per lesson [−2.1, −1.1]	<.001	preregistered; supported; −2.2% on untrimmed data
Sitting depth → error odds	1.026 [1.008, 1.043]	.0036	preregistered; p met; below the 1.10 size threshold

4. Discussion

Across nearly forty thousand embedded questions, where a lesson sat in its course did not predict whether its question was missed, at a precision that rules out the effect a front-loading rule needs. Learners sped up as they went and did not lose accuracy. Depth within a sitting showed a small rise confined to very long sittings and to few learners. For the design of short courses, these data give no reason to move important material earlier.

Limitations. Every estimate is an association from one platform's telemetry. The outcome is coarse: a first-attempt miss on an easy question, with no measure of attention itself. The sample is conditional in two ways the data cannot undo: lessons finished without answering the question are not in the sample and are more common on paid tiers, and a question answered later, outside the lesson, is stored exactly like one answered in it. Learners who make mistakes are slightly more likely to stop, so the null

describes those who continue. The interval is not reading time. Most of the data is from 3- and 5-lesson courses. The speed-accuracy literature warns against reading pace as effort [12,13], and we do not. Generalization beyond one gamified consumer app, two languages and easy comprehension checks is untested.

The obvious next study assigns course length and session length rather than observing them, which a platform that generates its own courses can do. A second question this dataset raised, why error rates vary threefold with when a course was generated, is about the stability of AI-generated questions rather than about learners, and deserves its own report.

5. Conclusion

In 39,867 comprehension outcomes from a mobile micro-learning app, a lesson's position in its course did not predict whether its question was missed. Learners completed each successive lesson about 1.5% faster with no detected change in accuracy. Front-loading material within a short course is not supported by these data. The platform studied is at https://nerdsip.com.

Data and code availability

The preregistration, aggregate result tables, figure data, extraction query source and analysis code accompany the OSF project (https://osf.io/px46t). Row-level telemetry is pseudonymous personal data and is not shared.

Ethics statement

Retrospective analysis of pseudonymous product telemetry generated during normal use under the platform's Terms of Service and Privacy Policy. GDPR legal basis: legitimate interest. Identifiers were replaced with salted pseudonyms on the server before extraction. Published results are aggregates with at least 10 distinct learners per displayed cell. The privacy policy covers the collection of this learning data but does not name research use specifically; the authors regard aggregate, pseudonymous analysis of it as within the platform's legitimate interest.

Competing interests

The authors are the founders of NerdSip, the platform studied. The preregistration, the frozen dataset, the disclosed prior data observation, the published deviation log and the reporting of unsupported hypotheses are the safeguards offered against this conflict.

AI assistance disclosure

Analysis code, figures and manuscript drafting were produced with the assistance of Claude (Anthropic). The authors reviewed and approved all analytic decisions and take full responsibility for the content.

Funding

Internal (NerdSip). No external funding.

References

1. Lee, Y.-M. (2023). Mobile microlearning: a systematic literature review and its implications. *Interactive Learning Environments*, 31(7), 4636–4651.
2. Reyes, G. (2023, revised 2024). *Cognitive endurance, talent selection, and the labor market returns to human capital.* arXiv:2301.02575.

3. Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, 41(2), 115–129.
4. Ong, T. Q., & Pastor, D. A. (2022). Uncovering the complexity of item position effects in a low-stakes testing context. *Applied Psychological Measurement*, 46(7), 571–588.
5. Vida, L. J., Bolsinova, M., & Brinkhuis, M. J. S. (2021). Speeding up without loss of accuracy: item position effects on performance in university exams. In *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 454–460).
6. Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS*, 115(11), 2600–2606.
7. van den Akker, O. R., Weston, S., Campbell, L., et al. (2021). Preregistration of secondary data analysis: a template and tutorial. *Meta-Psychology*, 5, 2625.
8. R Core Team. (2026). *R: A language and environment for statistical computing* (Version 4.6.1). R Foundation for Statistical Computing.
9. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
10. Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10.
11. Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101–133.
12. Domingue, B. W., Kanopka, K., Stenhaug, B., et al. (2022). Speed–accuracy trade-off? Not so fast. *Journal of Educational and Behavioral Statistics*, 47(5), 576–602.
13. Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill. *Journal of Educational Psychology*, 106(3), 608–626.