Targeted Practice in an Online Primary School Platform: A Large-Scale Embedded Evaluation

Jasper Naberman[1], Merel Das[1,2], Matthieu Brinkhuis[2], Sergey Sosnovsky[2]
[1] Futurewhiz
[2] Universiteit Utrecht
[1] Email: jasper@futurewhiz.com

ABSTRACT: Online learning platforms are in need of scalable, non-intrusive ways to demonstrate their learning impact. This paper presents and attempts to address this problem by conducting an embedded quasi-experimental study that demonstrates how on-topic practice improves students' subsequent performance using weekly personalized review quizzes, which resurface items a student previously missed. In Squla, a commercial Dutch practice platform for primary-school students, we compare performance after on-topic practice with off-topic practice in samples matched on engagement covariates and analyze accuracy with a generalized linear mixed model. From a dataset of 7,154 students (38,638 answers), on-topic practice yielded modest but consistent gains in Mathematics (72% vs. 69%, $p < .001$) and Language (73% vs. 70%, $p = .03$), with no significant effects in Spelling & Grammar or Reading Comprehension. This paper offers a reusable blueprint for continuous, actionable and non-intrusive evaluation of learning effectiveness in production environments.

Keywords: embedded evaluation, primary education, online practice, learning impact

1 INTRODUCTION

Digital technologies have been widely adopted in contemporary education, creating a growing need for scalable methods to evaluate their effectiveness in fostering learning (Blumenstein, 2020). Randomized controlled trials (RCTs) remain the methodological gold standard, yet in commercial settings they can be resource-intensive, disrupt authentic usage, and struggle to match the rapid cadence of product iteration (Bojinov & Gupta, 2022; Larsen et al., 2024).

Prior large-scale learning analytics work shows that iterative practice and measurement in production systems can shape learning experiences and guide product decisions (Brinkhuis et al., 2018; Portnoff et al., 2021). To address this need for non-intrusive, embedded evaluation in a concrete setting, we examine Squla, a Dutch online practice platform for primary school students. Each week, the platform delivers personalized review quizzes composed of items students answered incorrectly in the preceding four weeks. This creates an implicit adaptive loop that operationalizes retrieval practice. Taking these low-stakes tests not only measures knowledge but also strengthens it through the testing effect (Roediger & Karpicke, 2006). The feature facilitates an in situ comparison of on-topic (recently practiced, related skill) and off-topic (unrelated skill) practice organized as an embedded quasi-experimental evaluation.

This embedded design is applied to ask whether targeted on-topic practice leads to measurable learning gains with the help of Squla. Because the absence of randomization introduces potential bias, we match samples on engagement covariates before fitting a mixed-effects model. In doing so, we offer a practitioner-oriented blueprint for continuous, evidence-driven product evaluation that is rigorous enough to inform decisions, yet feasible for teams operating in real-world conditions.

2 METHODS

To address the methodological question for scalable, non-intrusive evaluation, we looked at Squla, a platform for primary school students aged 3 to 12, via a weekly personalized review quiz that surfaces items a student previously answered incorrectly within the past four weeks. The quizzes were offered in four subjects (Mathematics, Language, Spelling & Grammar, Reading Comprehension) to grades 1 through 8 from June to mid-August 2024, and contained five items. Students could take a review quiz in any subject where they had made at least five mistakes over the prior four weeks; items were randomly sampled without replacement from the student's pool of mistakes within the subject. The content of the school subjects on Squla adheres to the Dutch national educational standards.

We contrasted performance on review quiz items following on-topic versus off-topic practice (Figure 1). Topics were defined at the category level within a subject. An answer was labeled on-topic if the student had practiced ≥ 10 items (the standard quiz length in Squla) in that category between the original mistake and the review quiz; otherwise it was off-topic. Because a review quiz can draw from multiple categories, a single student's answers may contribute to both groups depending on the question. This design thus exploits naturally occurring contrasts in routine use by students.

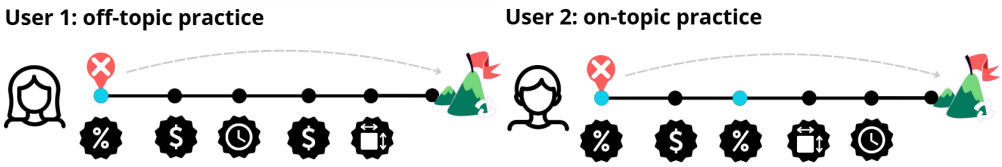

Figure 1: Example timelines of students in the off-topic and on-topic practice groups. Each circle represents a Mathematics quiz taken by a student, with the final quiz being the review quiz. Icons underneath the circles indicate the Mathematics category to which the quiz belongs.

Students were not randomized to practice conditions, so we used sample matching to reduce potential bias due to different levels of engagement (Stuart, 2010). We defined three engagement covariates: practiceBefore (items answered across the platform in the month before the study), practiceDuring (items answered in the relevant subject between a mistake and a review quiz), and practiceDays (days elapsed between a mistake and a review quiz). We applied Coarsened Exact Matching (CEM), choosing bin counts via Sturges' rule for practiceBefore and practiceDays, and 50 bins for practiceDuring. Matching was conducted per subject and per grade to respect distributional differences. Balance diagnostics indicated small standardized mean differences and variance ratios near 1; matching reduced the sample from 9,205 to 7,154 students.

After matching, we modeled binary answer correctness with a generalized linear mixed model with a logit link, including a random intercept at the student level. Let y_{ij} be 1 if student i answered item j correctly and $\pi_{ij} = \mathrm{P}(y_{ij} = 1 \mid b_i)$. We estimate

$$y_{ij} \sim \mathrm{Bernoulli}(\pi_{ij}), \qquad \mathrm{logit}(\pi_{ij}) = \alpha + \tau T_{ij} + \boldsymbol{\gamma}^{\top}\boldsymbol{S}_{ij} + \boldsymbol{\delta}^{\top}(T_{ij}\boldsymbol{S}_{ij}) + b_i, \qquad b_i \sim \mathcal{N}(0, \sigma_b^2).$$

Here $T_{ij} \in \{0, 1\}$ indicates on-topic, $\boldsymbol{S}_{ij}$ is a vector of subject dummies with Mathematics as the reference ($\gamma_{Math} = \delta_{Math} = 0$), and the b_i intercept captures between-student heterogeneity. Model selection was done using likelihood-ratio tests. Diagnostics indicated adequate model fit and model assumptions being met, confirming the model's validity.

3 RESULTS

This analysis aimed to determine whether prior practice on a topic influenced performance on this topic in the review quiz. After sample matching, we analyzed 38,638 answers given by 7,154 students (19,319 answers per practice group). Per-subject totals were 24,098 answers (Mathematics); 7,098 answers (Language); 3,042 answers (Spelling & Grammar) and 4,400 answers (Reading Comprehension). Engagement covariates were well balanced post-matching. These results, in combination with the sample size, suggest the matching-based quasi-experimental design is rigorous enough to inform decisions in a production environment.

Estimated marginal means (EMMs) on the probability scale indicated that on-topic practice was associated with small but consistent improvements in three subjects. In Mathematics, answer accuracy increased from 69% after off-topic practice to 72% after on-topic practice, and in Language from 70% to 73%. For Spelling & Grammar, the point estimates increased from 68% to 71% (non-significant), while Reading Comprehension showed a slight decline from 63% to 61% (also non-significant). Table 1 reports the on-topic and off-topic accuracy estimates per subject with 95% confidence intervals, percentage-point differences, corresponding odds ratios (ORs), and significance.

Table 1: EMMs on the probability scale, showing the estimated accuracy probability of a correct response (acc.) for on-topic and off-topic practice for each subject. Combined with ORs for contrasts of the EMMs. 'On' stands for on-topic practice, and 'off' for off-topic practice.

Subject	Off-Topic Acc. % [95% CI]	On-Topic Acc. % [95% CI]	Δ %-points	Odds Ratio (on/off) [95% CI]	p value
Mathematics	69 [67, 70]	72 [71, 73]	+3	1.16 [1.10, 1.25]	$< .001$
Language	70 [68, 72]	73 [71, 74]	+3	1.14 [1.01, 1.28]	.03
Spelling & Grammar	68 [65, 70]	71 [68, 74]	+3	1.19 [0.99, 1.43]	.06
Reading Comprehension	63 [60, 65]	61 [58, 63]	−2	0.92 [0.79, 1.05]	.21

Effect sizes on the OR scale tell a consistent story. The odds of a correct answer were higher after on-topic practice in Mathematics (on/off OR $= 1.16$, $p < .001$) and Language (OR $= 1.14$, $p = .03$), whereas the effects in subjects Spelling & Grammar (OR $= 1.19$, $p = .06$) and Reading Comprehension (OR $= 0.92$, $p = .21$) were not statistically significant. Absence of effects in Spelling & Grammar may be due to the smaller sample size, or greater performance variability. For Reading Comprehension, categories in Squla are organized in texts (such as stories about animals or hobbies) rather than discrete skills, so on-topic practice may not reinforce those skills, helping to explain the lack of effect.

4 DISCUSSION & CONCLUSION

This study demonstrates an example of an embedded quasi-experimental design that generates actionable evidence in a commercial platform without disrupting students' experience. On-topic practice preceding the review quiz was associated with small but consistent answer accuracy gains in Mathematics and Language. Effects were not statistically significant in Spelling & Grammar and Reading Comprehension. These results suggest that targeted practice would be more beneficial in hierarchically organized domains. Beyond the specific estimates, the study illustrates the practicality

of continuous, embedded evaluation for guiding product decisions at scale, prioritizing features for potential subsequent randomized tests and mitigating bias through matching.

Several limitations qualify these findings. As the participants were not randomized, unobserved confounding cannot be ruled out despite our sample matching based on relevant covariates; comparability checks mitigate but do not eliminate this risk. Sample representativeness may be imperfect because participation depended on engaging with the review quiz. Our on-topic definition relied on category membership; broad or cross-grade categories may dilute true skill alignment, particularly in Reading Comprehension. Finally, Spelling & Grammar had a relatively low sample size, which limited the sensitivity to detect smaller benefits.

Future work should tighten skill alignment, especially for Reading Comprehension, by grouping items hierarchically by comparable skills rather than text topics, enabling a reanalysis of existing data and a stronger test of on-topic practice. Extending the design temporally and to varying features, and examining heterogeneity by topic complexity or other factors, would strengthen external validity. Additionally, replicating this embedded approach across platforms and learning contexts can establish best practices for non-intrusive evaluation in industry settings. Overall, embedded quasi-experiments offer a scalable pathway to assess and improve educational technology in real-world contexts.

DECLARATION OF CONFLICT OF INTEREST

Jasper Naberman is employed by Futurewhiz, the developer of Squla. Merel Das was an intern at Futurewhiz during this study. Futurewhiz did not influence the analysis or interpretation of results.

REFERENCES

Blumenstein, M. (2020). Synergies of learning analytics and learning design: A systematic review of student outcomes. *Journal of Learning Analytics*, *7*(3), 13-32. https://doi.org/10.18608/jla.2020.73.3

Bojinov, I., & Gupta, S. (2022). Online experimentation: Benefits, operational and methodological challenges, and scaling guide. *Harvard Data Science Review*, *4*(3). https://doi.org/10.1162/99608f92.a579756e

Brinkhuis, M. J. S., Savi, A. O., Hofman, A. D., Coomans, F., van der Maas, H. L. J., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics*, *5*(2), 29-46. https://doi.org/10.18608/jla.2018.52.3

Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., & Stevens, N. T. (2024). Statistical challenges in online controlled experiments: A review of A/B testing methodology. *The American Statistician*, *78*(2), 135-149. https://doi.org/10.1080/00031305.2023.2257237

Portnoff, L., Gustafson, E., Rollinson, J., & Bicknell, K. (2021). Methods for language learning assessment at scale: Duolingo case study. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*. International Educational Data Mining Society. https://research.duolingo.com/papers/portnoff.edm21.pdf

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1-21. https://doi.org/10.1214/09-STS313