# Platform-Embedded Measurement of Online Learning: Evidence From a Question Order Swap Experiment

**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 products could benefit from scalable, non-intrusive ways to assess effectiveness during real use. In an online education platform, we embedded a minimal position swap in question sets. In a randomized controlled trial, the first and last questions swapped places to measure potential learning within practice sessions. We modeled binary response correctness with a generalized linear mixed model including random effects for student and content. Based on 53,035 answers (6,337 students), the same questions were more often correct at the end than at the start (first-regular: 35% vs. 32%, OR = 0.89, $p = .005$; last-regular: 24% vs. 21%, OR = 0.85, $p < .001$). The effect (~3 %-points) is modest but robust under real-world circumstances and requires no extra testing burden. Thus, this methodology is presented as a blueprint for continuous, in situ evaluation of learning.

**Keywords**: learning analytics, embedded evaluation, secondary education, online practice

## 1 INTRODUCTION

For learner-autonomous platforms, a central challenge is to measure learning effectiveness continuously and credibly in production environments, without disrupting students. Classroom RCTs and pre-post studies remain valuable, yet they are resource-intensive, slow to iterate, and often misaligned with self-directed usage. A complementary strand of work shows that platform-embedded assessments (tests or quasi-experiments that run during ordinary use) can surface decision-grade signals from log data alone (de Witte et al., 2015; Chen & Guthrie, 2019; Portnoff et al., 2021).

StudyGo is a Dutch online practice platform for secondary education (ages 12-18) offering textbook-aligned content across subjects. Among its features, practice question sets are short, frequent, and closely tied to chapter structure, making them ideal for in situ learning evaluation within a single study session. If repeated exposure to similar questions facilitates learning (Roediger & Karpicke, 2006), then a given question should be answered more accurately at end of a question set than at the start. Therefore, we unobtrusively swap first/last questions for a randomized subset of students, inside question sets, to measure learning within a practice session. Crucially, to attribute position differences to learning rather than topic heterogeneity within a question set, we screen for topic-cohesive question sets using filtering based on learning curves (Martin et al., 2011).

Beyond statistical evidence, the contribution for practitioners is a lightweight, reusable blueprint: how to introduce a small, low-risk randomized perturbation to question order; how to pre-filter content and focus analysis on topic-cohesive question sets; and how to analyze the effect so the signal remains interpretable under real-world circumstances. This approach is intended as a reliable supplement to other, more rigorous methodologies aimed at providing causal inferences. It offers continuous, non-intrusive evaluation that suits platforms at scale.

## 2 METHODS

We evaluated the effectiveness of StudyGo by introducing a minimal position swap intervention in existing practice question sets. For a randomized subset of students (1:1 allocation) we swapped the position of the first and last question in a set, while other students received the regular order. As shown in Figure 1, under the regular order, question 1 appears first and question $n$ appears last. Under the swapped order these two questions trade positions. This method yields four analytic cells that let us compare the same questions, without confounding question difficulty, in different positions per variation: first-regular, last-regular, first-swapped, and last-swapped. The core estimand is the within-question positioning effect during ordinary usage, i.e., in situ and without added burden to students.

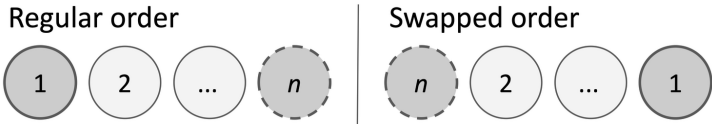


**Figure 1: Illustration of the regular and swapped order question sets in the study design.**

To ensure position differences reflect learning rather than topic heterogeneity within a question set, we retained only topic-cohesive sets using a learning curve screening (error rate vs. opportunity to practice) (Martin et al., 2011). For each set we fit a power law curve, with $P = BN^{-\alpha}$ where $P$ is error rate and $N$ is the question position index (opportunity to practice). We required a negative slope ($\alpha > 0$) and, based on Jarantow et al. (2023), a good non-linear fit measured by residual standard error (RSE). Based on the empirical distribution of RSE values, question sets with RSE $\leq 0.15$ were included.

For the statistical model, answer accuracy was coded as binary where correct = 0 and incorrect = 1. Let $y_i \in \{0,1\}$ denote the response for answer $i$, with $\pi_i = Pr(y_i = 1)$. We modeled the log-odds of an incorrect answer with a generalized linear mixed model (GLMM) with a logit link:

$$y_i \sim \text{Bernoulli}(\pi_i), \qquad \text{logit}(\pi_i) = \alpha + \beta_1 \text{PosLast}_i + \beta_2 \text{VarSwapped}_i +$$

$$\beta_3 (\text{PosLast} \times \text{VarSwapped})_i + b_{u[i]}^{(\text{student})} + b_{t[i]}^{(\text{type})} + b_{s[i]}^{(\text{stream})} + b_{k[i]}^{(\text{topic})} + b_{q[i]:k[i]}^{(\text{question:topic})}.$$

For the fixed effects, here $PosLast \in \{0,1\}$ indicates last (1) vs. first (0) position; $VarSwapped \in \{0,1\}$ indicates swapped (1) vs. regular (0) ordering experimental variation. The interaction $\beta_3$ allows the position effect to differ for questions that are normally first vs. normally last. For the random effects, intercepts for student, question type, stream (educational track), topic, and question nested in topic capture heterogeneity at relevant levels. Model selection was done using likelihood-ratio tests. Diagnostics indicated adequate model fit and assumptions being met, confirming the model's validity.

The fixed-effects contrast first-regular vs. last-swapped compares the same question when it appears at the start vs. end for questions normally first. The last-regular vs. first-swapped contrast does the same for questions normally last. If practice within a set contributes to learning, the end position should yield more accurate answers.

## 3 RESULTS

The final analytic set comprised 53,035 answers from 6,337 secondary education students across 718 topic-cohesive practice sets (with seven questions on average) that passed the learning curve screening (RSE $\leq 0.15$) and session filters (completed in one sitting, no skips, no $< 1$ second

responses). A randomization check on the unchanged middle questions showed no ability imbalance between order conditions (Wilcoxon signed-rank test, $\boldsymbol{p} = .16$). We modeled the log-odds of an incorrect answer with the GLMM described above. The model fit the data well: there was no overdispersion, homogeneous variances, and Q-Q plots of random effects were approximately normal. AIC/BIC favored this specification over simpler alternatives. Both main effects and the interaction were significant (last vs. first: $\boldsymbol{\beta} = -0.68$, SE $= 0.07$, $\boldsymbol{p} < .001$; swapped vs. regular: $\boldsymbol{\beta} = -0.52$, SE $= 0.07$, $\boldsymbol{p} < .001$; interaction (position × variation): $\boldsymbol{\beta} = 1.09$, SE $= 0.14$, $\boldsymbol{p} < .001$), indicating that the position effect differs for questions that are normally first vs. normally last. Taken together, these coefficients imply that moving a question from start to end reduces errors for both groups: for questions that are normally first the end vs. start contrast is $\boldsymbol{\beta_1} + \boldsymbol{\beta_2} + \boldsymbol{\beta_3}$ (OR ≈ 0.89), and for questions that are normally last it is $\boldsymbol{\beta_1} - \boldsymbol{\beta_2}$ (OR ≈ 0.85). Thus, both groups are less often wrong at the end, with a slightly larger reduction for questions that are normally last.

Table 1 reports estimated marginal means (EMMs) on the probability scale for the four (position, order) cells, together with the two within-question contrasts that hold question content constant. Questions that are normally first show a lower error rate when they appear at the end (last-swapped) than at the start (first-regular). Questions that are normally last show the same pattern when compared at the starting position (first-swapped) vs. ending position (last-regular). In both cases the difference is about 3 percentage points, with odds ratios (ORs) indicating a reliably lower odds of error at the end of a question set.

**Table 1: EMMs on the probability scale and key contrasts. Percentages are EMMs for the probability of an incorrect answer. ORs compare the odds of an incorrect answer across the two positions for the same questions (end vs. start).**

| Question set (by normal position) | Start (position, order) [95% CI] | End (position, order) [95% CI] | Δ %-points | Odds Ratio [95% CI] | $p$ value |
|---|---|---|---|---|---|
| Normally first questions | 35 (first, regular) [27, 43] | 32 (last, swapped) [25, 40] | −3 | 0.89 [0.82, 0.97] | .005 |
| Normally last questions | 24 (first, swapped) [18, 31] | 21 (last, regular) [16, 27] | −3 | 0.85 [0.78, 0.93] | < .001 |

Random effects captured meaningful heterogeneity that would otherwise mask the position signal: variances (SD) on the log-odds scale were 0.29 (0.54) for students, 0.29 (0.54) for questions (nested within topics), and 0.38 (0.62) for topics, with smaller variability for streams 0.03 (0.18) and question types 0.04 (0.21). Together with the significant fixed effects and within-question contrasts, these results indicate that, after screening for topic-cohesion and accounting for learner and content differences, questions are answered more accurately at the end of a practice set than at the start, providing in situ evidence of within-session learning.

## 4 DISCUSSION & CONCLUSION

StudyGo's embedded position-swap experiment provides in situ evidence of within-session learning. Holding question content constant and modeling substantial student- and content-level heterogeneity, the same questions were answered more accurately at the end of a set than at the start (∼3 %-point improvement; ORs ∼0.85 - 0.89). Interpreting the magnitude, the effect is modest, as expected for short sets, but useful enough for product decisions. Even under real-world

circumstances and without extra testing burden, practice on StudyGo yields measurable, immediate knowledge gains. In addition, this study's value is methodological for practitioners. A minimal randomization pairs with a learning curve screen to focus analysis on topic-cohesive sets, while a GLMM absorbs variation from students, topics, streams, question types, and questions within topics. The result is a lightweight, repeatable blueprint for continuous monitoring that complements traditional pre/post or classroom trials.

Several limitations qualify the claims. Generalizability is bounded by the question set screening (and other quality filters). Difficulty often increases or decreases across a set; swapping only first/last positions mitigates but does not fully disentangle learning from difficulty progression. Lastly, outcomes in this design reflect immediate accuracy, not long-term knowledge retention.

Future work could test durability by linking end-of-set gains to later performance. In addition, more than two positions could be varied to separate learning from difficulty progression directly. In parallel, the same blueprint can be adapted to other platform features that present repeated, short-term practice opportunities.

Overall, a minimal intervention embedded in ordinary use, coupled with principled pre-filtering and mixed-effects modeling, yields a clear, interpretable signal of within-session learning on StudyGo. While modest in size, the effect is operationally meaningful and demonstrates a scalable way to measure effectiveness continuously in learner-autonomous products.

## DECLARATION OF CONFLICT OF INTEREST

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

## REFERENCES

Chen, Z. & Guthrie, J. (2019). Measuring the effectiveness of learning resources via student interaction with online learning modules. *arXiv*. http://dx.doi.org/10.48550/arXiv.1903.08003

De Witte, K., Haelermans, C., & Rogge, N. (2015). The effectiveness of a computer-assisted math learning program. *Journal of Computer Assisted Learning, 31*(4). https://doi.org/10.1111/jcal.12090

Jarantow, S. W., Pisors, E. D., & Chiu, M. L. (2023). Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. *Current Protocols, 3*(6). https://doi.org/10.1002/cpz1.801

Martin, B., Mitrovic, A., Koedinger, K. R., & Mathan, S. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction, 21*, 24-283. https://doi.org/10.1007/s11257-010-9084-2

Portnoff, L., Gustafson, E., Rollinson, J., & Bicknell, K. (2021). Methods for language learning assessment at scale: Duolingo case study. Paper presented at the International Conference on Educational Data Mining. 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