NBER WORKING PAPER SERIES

THE ECONOMICS OF ALGORITHMIC PERSONALIZATION: EVIDENCE FROM AN EDUCATIONAL TECHNOLOGY PLATFORM

Keshav Agrawal
Susan Athey
Ayush Kanodia
Shanjukta Nath
Emil Palikot

Working Paper 34950
http://www.nber.org/papers/w34950

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2026

We are grateful to Deepak Agarwal, Nikhil Saraf, and their team at Stones2Milestones for collaboration on this project. The Golub Capital Social Impact Lab at Stanford Graduate School of Business provided funding for this research. We thank Rina Park, Jazon Jiao, Vikram Dixit Kumarakswamy, Anjali Adukia, Michele Belot, and participants of the AI in Social Science 2025 conference for helpful comments and discussion. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Economics of Algorithmic Personalization: Evidence from an Educational Technology Platform

Keshav Agrawal, Susan Athey, Ayush Kanodia, Shanjukta Nath, and Emil Palikot

NBER Working Paper No. 34950
March 2026
JEL No. C21, C93, I21, L81, M30

ABSTRACT

Can personalized recommendations improve engagement in educational technology? We design, test, and scale a collaborative filtering system for Freadom, an English-learning app for Indian children. A randomized controlled trial (RCT) with 7,750 students shows that personalization, deployed in a single content section, increases engagement by 60% in that section and by 14% app-wide. We then exploit an eligibility threshold in a regression discontinuity design (RDD) to track effects over five months of deployment. For user cohorts receiving personalization during deployment, RDD estimates exceed RCT benchmark by a factor of 2.5, opposite of the “voltage drop" typically observed in policy scale-ups. This provides evidence that, for algorithmic interventions, RCT estimates may be lower bounds on scaled impact rather than upper bounds. However, personalization benefits are front-loaded. Gains concentrate in users’ first weeks, with diminishing returns thereafter. This pattern, combined with the sharp decline in predicted match quality as users exhaust their best content matches, suggests that content availability rather than algorithmic sophistication becomes the binding constraint.

Keshav Agrawal
Stanford University
kasince2k@gmail.com

Shanjukta Nath
University of Georgia
shanjukta.nath@uga.edu

Susan Athey
Stanford University
Graduate School of Business
Department of Economics
and NBER
athey@stanford.edu

Emil Palikot
Northeastern University
Business
Marketing
and Stanford Universit
emil.palikot@gmail.com

Ayush Kanodia
Stanford University
akanodia@stanford.edu

A randomized controlled trials registry entry is available at https://osf.io/ynmer/overview

1 Introduction

Education apps have among the lowest retention rates of any software category.[1] Completion rates for online courses average 3-6 percent, a figure that did not improve despite years of investment in course development (Reich and Ruipérez-Valiente, 2019). Users download learning apps, try a few lessons, and leave. This limits the potential of educational technology (EdTech) to improve learning outcomes, particularly in developing countries where EdTech has been promoted as a scalable solution to educational inequality (Muralidharan et al., 2019; Rodriguez-Segura, 2022).

Personalized recommendation algorithms offer a potential solution. These systems have generated enormous value for entertainment platforms by matching users with content suited to their individual preferences (Gomez-Uribe and Hunt, 2015; Holtz et al., 2020; Zielnicki et al., 2025). But translating these gains to education faces two challenges. First, learners may not have heterogeneous preferences for educational content, or, if they do, those preferences may be difficult to infer from limited interaction data. Second, even if heterogeneous preferences exist, they may not drive behavior. In educational settings, learners often follow curricula, learning paths, or schedules set by teachers and parents rather than choosing content freely based on their own tastes. If such guidance dominates learner choice, matching on preferences adds little value.

Furthermore, EdTech platforms face practical decisions such as when to deploy personalized recommendations, for which users, and what evidence to rely on. Suppose an experiment shows that personalization increases engagement by 14 percent. Should the platform expect a 14 percent gain in deployment or something different? Two literatures offer conflicting guidance. The policy evaluation literature documents that treatment effects typically decline during scale-up, referred to as the "voltage drop" phenomenon (List, 2022). This suggests the experimental estimate is an upper bound on what the platform should expect. By contrast, the machine learning literature documents that algorithms improve with scale. More users generate more data, enabling better predictions (Halevy et al., 2009). If algorithmic learning dominates implementation frictions, the experimental estimate may be a lower bound. Which force wins is an empirical question. To answer it, researchers need to measure and compare treatment effects both in a controlled experiment and during actual deployment

We study this in the context of Freadom, a mobile application serving approximately 6,400 daily active users. These users are children aged 3-12 learning to read in English in In-

[1]Industry benchmarks suggest only 2-4 percent of users remain active after 30 days. See Business of Apps, "Education App Benchmarks", 2025.

dia. The app provides almost 3,200 short illustrated stories organized into sections—horizontally scrollable slates. Before our intervention, editors selected content for all these sections on a weekly basis. We designed a personalized recommendation system using collaborative filtering, setting an eligibility threshold at 60 story interactions. Users below this threshold receive editorial recommendations, while users above receive personalized recommendations based on their interaction history. This threshold creates a sharp discontinuity that enables causal inference during scaled deployment. We evaluate the system in two phases: first, a randomized experiment measuring the impact on user engagement for two weeks; and second, a regression discontinuity analysis tracking effects over five months of scaled deployment.

We conducted a randomized controlled trial with 7,750 users. Treatment users received personalized recommendations in one content section of the app; control users continued receiving recommendations curated by content specialists (editorial system). We find that personalization increases engagement in the personalized section by 60% (SE = 17%) and app-wide engagement by 14% (SE = 7%). The increase in overall usage indicates users substitute toward app engagement from other activities rather than merely reallocating time across app sections.

Second, we analyze the scaled deployment for the subsequent five months, during which all users with 60+ interactions received personalized recommendations in the personalized section. As users continue engaging with the app, they cross the eligibility threshold at different times, creating a continuous natural experiment. We exploit this sharp discontinuity to estimate local average treatment effects (LATE) by comparing users just above versus just below the 60-interaction threshold using the local randomization methods for regression discontinuity developed in Cattaneo et al. (2015, 2018). We validate the RDD design in three ways: first, by confirming that users just below the threshold exhibit outcomes indistinguishable from control group users in the experimental sample. Second, we show that the mean user engagement just below and above the threshold is comparable in the week prior to personalized recommendations. Third, there is no discontinuity at the threshold in a placebo section that never received personalized recommendations. Pooled RDD estimates exceed experimental benchmarks by a factor of 2.5 ($p < 0.001$), though we document substantial week-to-week variation.

This RCT-to-RDD comparison addresses a central question in the science of scaling: Do treatment effects observed in controlled experiments persist when interventions scale to real-world deployment? The "voltage drop" literature suggests they typically shrink. We find the opposite—scaled effects exceed experimental benchmarks—suggesting that for algorithmic interventions, the standard calculus may be reversed. We document three additional patterns that illuminate when and why personalized recommendations scale successfully.

First, users with more interaction history benefit more from personalization, consistent with the data requirements of collaborative filtering algorithms. In the RCT, heavy users experience substantially larger gains (24 percentage point difference, $p < 0.01$) than light users. This heterogeneity reflects a fundamental feature of matrix factorization. The algorithm learns preferences better when more historical data is available.

Second, treatment effects vary substantially over time. During the first month of deployment, we estimate gains comparable to the RCT. Effects then increase substantially during the mid-deployment period, though weekly estimates exhibit considerable volatility—particularly during school holidays and the Diwali period, when baseline app usage drops and several weeks show no detectable gains from personalization.

Third, personalization benefits are front-loaded: users experience the largest engagement gains during their first week of exposure, with marginal returns declining sharply thereafter. We trace this pattern to content inventory constraints. The predicted engagement declines steeply across story ranks: within the top 20 stories—corresponding to roughly three weeks of usage for an average user—predicted engagement drops by 20 percent. This indicates that the stock of high-quality recommendations depletes rapidly. Unlike entertainment platforms with continuously refreshed catalogs, Freadom's library of 3,200 stories, partitioned across eight grade levels, imposes binding constraints on sustained gains in personalization.

This study provides evidence on how data accumulation and content depth jointly determine personalization effectiveness, with implications for platform economics at a moderate scale. Our findings highlight a dual investment challenge: algorithmic performance requires sufficient user interaction histories, while sustained engagement gains depend on content library depth. The complementarity between these inputs—and the binding constraint imposed by limited content inventories—informs how smaller platforms compete and prioritize between algorithm development and content acquisition.

Two important caveats warrant emphasis. First, we measure engagement rather than learning directly. While we do not observe test scores or other learning outcomes, prior evidence suggests engagement gains often translate to learning outcomes. For instance, Escueta et al. (2020) reviews 19 RCTs that increased engagement with EdTech products and finds that 12 improved learning outcomes. We show that treated users spend more time reading and complete more difficult stories, but the relationship between engagement and learning depends on what students substitute away from and whether additional content consumed has diminishing returns. Second, our analysis spans only five months post-deployment. While we document content inventory constraints that limit long-run benefits, longer-term monitoring would strengthen these findings.

Despite these limitations, our findings challenge conventional wisdom about scaling. For

interventions that learn from data, experimental estimates may understate rather than overstate scaled impact. This suggests that platforms evaluating personalization should view RCT results as a lower bound, not an upper bound, when algorithms will continue learning post-deployment. However, we identify the content inventory constraint as a new channel through which voltage drops can occur. As personalization algorithms surface the best-matched content, they deplete the stock of high-quality recommendations, causing benefits to decline over time. For platforms with finite content libraries, algorithmic improvement and content exhaustion work in opposite directions, and content availability may ultimately turn into a binding constraint.

2 Related Literature

Existing evidence on personalized content recommendation systems in education comes primarily from small studies that bundle personalization with user interface changes. Drachsler et al. (2008) recruited 250 university students to study the effect of showing personalized course material recommendations versus no recommendations. While this study is a randomized experiment, it bundles two changes: adding a user interface element and personalizing recommendations. Ruiz-Iniesta et al. (2018) develop and test a recommendation system on the digital education platform, Smile and Learn, in an observational study. Their treatment combined a new user interface component with an introduction of a collaborative filtering algorithm. They find substantial increases in consumption of recommended items. However, their treatment confounds navigation simplification with personalization.

More recently, several studies have examined adaptive learning platforms that personalize content based on student ability. Muralidharan et al. (2019) document significant learning gains (0.37 SD after 4.5 months) from adapting content difficulty to match student performance levels. Kumar and Mehra (2024) develop an algorithm that identifies student learning needs and generates personalized homework questions accordingly. In a field experiment with 240 middle school students in India, they find that students receiving personalized homework score 0.16 SD higher on standardized exams, with the largest gains for medium-ability students. Oreopoulos et al. (2024) shows that improving teachers' ability to use Khan Academy's adaptive teaching modules positively impacts students whose classrooms had higher average engagement, but null effects for students with low engagement. However, their bundled intervention makes it difficult to isolate the impact of algorithmic personalization from teacher coaching.

Outside of education, personalized recommendation systems have generated large value at major technology platforms. Linden et al. (2003) describe how Amazon's item-to-item

collaborative filtering increases cross-selling at scale. Holtz et al. (2020) provide experimental evidence from Spotify showing that personalized recommendations increase engagement, though with tradeoffs in content diversity. Gomez-Uribe and Hunt (2015) and Zielnicki et al. (2025) study the value of personalized recommendation systems at Netflix. Gomez-Uribe and Hunt (2015) estimate that the recommender system was worth over $1 billion annually in 2015, with the majority of content consumption driven by personalized recommendations. Zielnicki et al. (2025) provide experimental evidence that Netflix's advanced recommender system generates 4-12% higher engagement than simpler alternatives. However, these platforms operate at massive scale, Netflix serves over 200 million subscribers, Spotify over 500 milion users, raising questions whether personalization techniques can function effectively in moderate-data environments typical of educational technology.

Our work is the first RCT-based study in EdTech that isolates the effect of personalization of content type on user engagement from other app changes. We examine preference-based personalization using collaborative filtering, matching users with the most engaging content based on their interaction histories. This complements prior work on ability-based adaptation by demonstrating that preference heterogeneity, not just skill heterogeneity, can be exploited to increase engagement. Critically, our setting allows us to test whether personalization techniques developed for platforms with hundreds of millions of users can function effectively at moderate scale (approx. 6,400 daily active users), answering a question of direct relevance to educational technology practitioners. Additionally, our findings contribute to behavioral economics by demonstrating that heterogeneity in content preferences, rather than ability or skill alone, substantially determines engagement in educational settings. Unlike traditional interventions relying on external incentives or nudges, personalization leverages intrinsic interest, suggesting that preference-based matching can serve as an alternative mechanism for sustaining engagement.

A central challenge in policy evaluation is assessing whether treatment effects observed in controlled experiments persist when interventions scale to real-world deployment. A robust literature documents that they typically do not. Most promising ideas experience "voltage drops" as small pilots expand to large-scale implementation (List, 2022; Brandon et al., 2022; Vivalt, 2020). Multiple factors contribute to the observed voltage drops including context dependence, implementation quality, general equilibrium effects, and population representativeness (Banerjee et al., 2017; Al-Ubaydli et al., 2020).

This "voltage drop" literature, however, focuses primarily on static interventions such as cash transfers, educational curricula, and behavioral nudges, where the intervention itself is fixed during scale-up. A separate literature documents that algorithmic systems improve with scale. Ardalani et al. (2022) and Zhang et al. (2024) document scaling laws for recom-

mendation models, showing quality follows power-law improvements in data and compute. Schaefer and Sapi (2023) provide causal evidence of data network effects in search engines. Peukert et al. (2024) demonstrate that algorithmic recommendations require sufficient user-level data to outperform human curation, with performance improving as data accumulates.

These two literatures have largely talked past each other because they study different types of interventions. We argue that when the intervention is an algorithm that learns from deployment data, the sign of the scaling gradient may reverse. More users generate more interaction data, enabling better predictions and larger treatment effects—a countervailing force that can offset or dominate traditional voltage drops. In our setting, RDD estimates during scaled deployment exceed experimental benchmarks by a factor of 2.5, consistent with algorithmic improvements from accumulated data. This finding has important implications for how policymakers and platform managers should interpret experimental evidence on algorithmic interventions: rather than applying standard voltage drop discounts, they may need to consider that the intervention might improve with scale.

Rigorous evaluation of whether effects persist at scale is often challenging. Experimental evaluation of deployed interventions requires holdout samples, and such designs can be highly informative; Brynjolfsson et al. (2025) uses a nine-year holdout to measure cumulative welfare effects of Facebook's advertising system. However, maintaining long-term holdouts is both costly and operationally difficult. Dmitriev et al. (2016) document pitfalls of long-running experiments, including cookie churn and survivorship bias. Developing strategies to reduce these holdout costs remains an active area of research (Mercer and Meakin, 2025; Johnson et al., 2017).

We propose an alternative strategy that exploits a discontinuity induced by an eligibility threshold for personalization. Users need 60 interactions before receiving personalized recommendations, creating a sharp threshold that serves dual purposes: ensuring the algorithm has sufficient data for accurate predictions and enabling causal identification during scaled deployment. Several papers combine RCT and RDD evidence to validate regression discontinuity designs in contexts ranging from political to development economics (Buddelmeyer and Skoufias, 2004; Green et al., 2009; Chaplin et al., 2018). We take this approach further: rather than using an RCT to validate an RDD retrospectively, we validate the RDD's local-randomization assumption during our randomized experiment, then apply it prospectively to the scaled deployment period. This design enables "continuous evaluation". As new users cross the eligibility threshold each week, the RDD provides ongoing quasi-experimental variation without requiring any users to be permanently excluded from treatment (holdout).

Finally, while we document that aggregate treatment effects are larger during scaled deployment than in the initial experiment, we identify an important constraint limiting long-

run benefits. Unlike entertainment platforms with continuously refreshed catalogs—Netflix adds hundreds of titles monthly; Spotify hosts 70 million songs—educational platforms often operate with static content libraries. Freadom's 3,200 stories, partitioned across grade levels, create binding inventory constraints. We document that personalization benefits are front-loaded: users experience the highest engagement gains during their first week of exposure, with marginal returns declining thereafter. This reflects content inventory depletion rather than algorithmic failure—the algorithm continues to optimize, but the set of high-quality matches shrinks as users exhaust preferred content. This content constraint relates to the voltage effect framework, but through a different mechanism: rather than implementation drift degrading treatment quality, declining returns occur because the opportunity set for effective personalization shrinks. For platforms considering personalization investments, this suggests algorithmic improvement and content development are complements.

3 Empirical Setting

3.1 The Freadom Application and User Base

Stones2Milestones (S2M) is an education technology company operating in India. Freadom is their flagship mobile and web application. Freadom serves children aged 3-12 learning to read in English, addressing a critical educational challenge in India. This is a setting where many children lack access to age-appropriate English reading materials at home, particularly in households where parents have limited English proficiency (Treffers-Daller et al., 2022; Roy, 2004). During our study period (July to December, 2021), the app had approximately 6,400 daily active users.

Users access Freadom through three primary channels. Business-to-Business (B2B) partnerships with schools account for 35% of users, where schools recommend the app as supplementary learning material. Business-to-Consumer (B2C) users represent 40%, primarily parents who independently download the app for their children. The remaining 25% are paid subscribers who access premium features.[2]

Freadom's content library contains 3,163 short illustrated stories created by S2M's pedagogical team and educational publishers specializing in children's content. Stories are age-appropriate, curriculum-aligned, and designed with explicit educational goals. Each story is self-contained, typically 2-5 minutes in length, and includes both narrative content and comprehension quizzes. Stories are tagged with metadata including grade level, content cat-

[2]Paid users are initially acquired as B2B or B2C users and later upgrade to the paid version. The historical variables in the dataset do not allow us to distinguish whether these premium users were upgraded from B2B or B2C.

egory (e.g., animals, sports, adventure, science), and difficulty level. S2M operates under the premise that maximizing engagement with this curated content helps learners achieve their educational objectives.

Freadom organizes content into sections, each displaying a horizontal slate that typically contains 15 stories. Users scroll vertically across sections and swipe horizontally within a section to browse stories. Clicking on a story reveals a description page with an illustration and summary; users then decide whether to start reading or return to browsing. During our study period, the most prominent sections included *Recommended Story*, *Trending Now*, *Top Stories*, and *Today for You*. Before our intervention, all sections used non-personalized algorithms: editorial curation, popularity-based ranking, or recency-based selection. Appendix A1 provides screenshots of the interface.

3.2 User Engagement Patterns

When a story appears in a slate, users progress through several decision stages: whether to click and view the story's description page, whether to start reading after viewing the description, and whether to complete the story after starting. User-story interactions follow a characteristic funnel pattern. Approximately 30% of story impressions result in users clicking to view the description page. Of users who view a description, approximately 50% proceed to start reading the story. Of users who start a story, approximately 53% complete it, with many users completing most of the story and dropping out in the last pages. Overall, only about 8% of story impressions result in completion, while the majority of interactions involve exploration and information-gathering as users browse multiple stories before selecting which to read fully.

This exploration behavior motivates our primary outcome metric construction. Counting only story completions would miss both the information-gathering phase and stories started but not finished. We construct a *Story Engagement* metric that assigns partial credit to viewing (0.3) and starting (0.5) stories, with full credit (1.0) for completion. We show robustness to alternative metrics including story completions, reading time, and completion of difficult stories. This is also the metric described in our pre-analysis plan.

Users in our sample exhibit substantial heterogeneity in engagement patterns and platform tenure. Table 1 reports summary statistics for key usage metrics. We focus on the sample of users who launched the app between July 2020 and July 2021 (the period prior to our RCT).

The median user completes 0.0 stories per week, while the 90th percentile completes 2.0 stories each week throughout the year. Furthermore, the median user interacts with

Table 1: User Base Summary Statistics

Characteristic	Mean	Median	90th Percentile	SD
Stories completed per week	0.70	0.00	2.00	2.70
Story interactions per week	7.20	1.50	17.30	21.90
Weeks active on app	3.80	2.00	8.00	5.10
Cumulative story interactions	43.90	3.00	72.00	250.30

Note: Summary statistics for active users during July 2020 - July 2021. Sample includes all users who launched the app at least once during this period. Story interactions include skipping, viewing, starting, or completing a story.

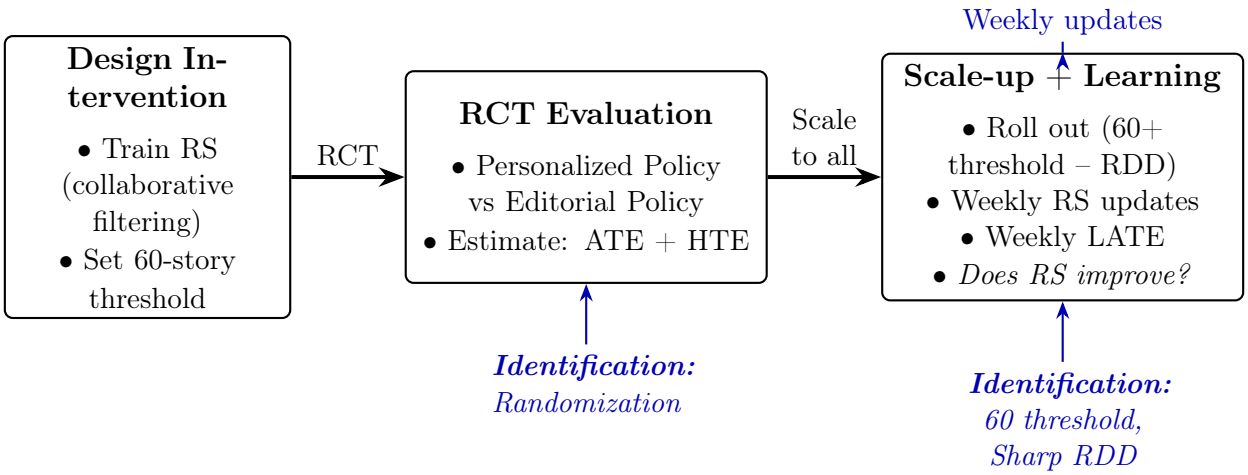
only 1.5 stories per week, whereas the 90th-percentile user interacts with over 17 stories per week. These facts underscore two important dynamics; first, many users are active for just a few weeks before churning from the app. Specifically, an average user is active for 3.8 weeks (a median of 2 weeks). Second, the app usage is highly skewed; with some users being mostly inactive and others interacting with and completing many stories (the standard deviation of the number of story interactions per week is 22). This heterogeneity in interaction histories creates substantial variation in the quantity of preference data available to the personalization algorithm. Users with longer tenure and richer engagement histories provide the algorithm with more information to learn their preferences, while newer or lighter users present a greater cold-start challenge.

3.3 Research Design Overview

Figure 1 summarizes our research design. We proceed in three stages. First, we developed a collaborative filtering recommendation system and established a 60-interaction eligibility threshold based on prediction accuracy. Second, we conducted a randomized controlled trial comparing personalized recommendations against editorial curation, estimating average and heterogeneous treatment effects. Third, we scaled personalized recommendations to all eligible users while continuing to monitor effects through a regression discontinuity design that exploits the eligibility threshold.

The 60-interaction threshold serves a dual purpose. First, it ensures the algorithm has sufficient data for accurate predictions, and second, it creates a sharp discontinuity enabling causal identification during scaled deployment. The two approaches combined allow us to establish a proof-of-concept experimentally, scale to all eligible users, and continue evaluating effects causally, eliminating the need to maintain costly holdout groups.

Figure 1: Research Design Pipeline and Empirical Strategy



4 Conceptual Framework

Consider a set of users indexed by $i \in \{1, \ldots, N\}$ observed over discrete time periods $t \in \{1, \ldots, T\}$, where each period corresponds to one week. Let $\mathcal{J} = \{1, \ldots, J\}$ denote the set of available stories. Let $\mathcal{H}$ denote the space of all possible interaction histories. Each history $h \in \mathcal{H}$ is a matrix with J rows (one per story) and four columns:

$$h = \begin{pmatrix} j_1 & s_1 & e_1 & u_1 \\ j_2 & s_2 & e_2 & u_2 \\ \vdots & \vdots & \vdots & \vdots \\ j_J & s_J & e_J & u_J \end{pmatrix} \tag{1}$$

where for each story $j \in \mathcal{J}$: j is the story ID, $s_j \in \{0, 1\}$ indicates whether the story appeared in a slate shown to the user, $e_j \in \{0, 1\}$ indicates whether the user interacted with the story (clicked, started, or engaged) and $u_j \in \{0, 0.3, 0.5, 1\}$ is the engagement level, with $u_j = 0$ if $e_j = 0$. Thus, $\mathcal{H} = (\mathcal{J} \times \{0, 1\} \times \{0, 1\} \times \{0, 0.3, 0.5, 1\})^J$. Let H_{it} denote the random variable representing user i's interaction history prior to week t, taking values $h_{it} \in \mathcal{H}$.

Let $X_{it} : \mathcal{H} \to \mathbb{Z}_{\geq 0}$ be the function that maps a history to the cumulative number of interactions $X_{it}(h_{it}) = \sum_{j=1}^{J} e_j$ where e_j is the interaction indicator for story j in history h_{it}. This count serves the dual purpose of eligibility for recommendation system and the running variable in our regression discontinuity design.

Outcome Definition. Let U_{ij} denote user i's engagement with story j, measured as a normalized score between 0 and 1 that combines reading time, completion status, and interaction depth. Let $\mathcal{S}_{it}$ denote the set of stories user i interacts with during week t. The weekly outcome used in our analysis is $Y_{it} = \sum_{j \in \mathcal{S}_{it}} U_{ij}$. This represents total engagement aggregated across all stories consumed during the week. The set $\mathcal{S}_{it}$ is determined endogenously by user behavior, comprising the stories the user chooses to engage with among those presented in the slate.

Recommendation Policy. Let $\mathcal{R}$ denote the set of all possible rankings of the J stories. A ranking $r \in \mathcal{R}$ is an ordered list $r = (j_{(1)}, j_{(2)}, \ldots, j_{(J)})$ where $j_{(k)}$ denotes the story at rank k. A policy is a function:

$$\pi : \mathcal{H} \to \mathcal{R} \tag{2}$$

that maps interaction histories to rankings over stories. Let R_{it} denote the random variable representing the ranking user i receives in week t, with realization r_{it}. In other words, for user i in week t with history h_{it}, the policy generates ranking $r_{it} = \pi(h_{it})$. We consider two policies. The editorial policy π_t^E produces a common ranking curated by content specialists, independent of individual histories. Formally, $\pi_t^E(h) = r_t^E$ for all $h \in \mathcal{H}$, where r_t^E is a fixed ranking in week t. The personalized policy π_t^P uses collaborative filtering to generate user-specific rankings based on interaction histories. The personalized policy is retrained weekly. The subscript t denotes the algorithm version in period t.

App Implementation. For user i in week t, one of the two policies generates a ranking of all stories at the start of the week. The app then applies standard rules to translate this ranking into what the user sees throughout the week. These rules include deduplication (removing previously started stories), slate size (15 stories visible at a time), staleness rules (removing top stories if not engaged with), and daily updates. The rules are equivalent across all policies. The treatment is the ranking itself. The app mechanics determine how this ranking translates into user experience.

Potential Outcomes. Potential outcomes are defined over rankings. For user i in week t, let $Y_{it}(r)$ denote the potential engagement which is the total engagement user i would exhibit in week t if presented with ranking $r \in \mathcal{R}$. The distribution of R_{it} depends on treatment assignment. Under control ($D_i = 0$): $R_{it} = \pi_t^E(h)$ for all h, yielding the fixed editorial ranking r_t^E. Under treatment ($D_i = 1$): $R_{it} = \pi_t^P(H_{it})$, a user-specific ranking that depends on the random history H_{it}. The observed outcome is $Y_{it} = Y_{it}(R_{it})$.

4.1 Estimands

Our primary estimand in period t is the average treatment effect of personalized recommendations relative to editorial curation:

$$\tau_t = \mathbb{E}[Y_{it}(\pi_t^P(H_{it}))] - \mathbb{E}[Y_{it}(r_t^E)], \tag{3}$$

where $\pi_t^E(H_{it}) = r_t^E$ for all i. This is the difference in expected outcomes when rankings are generated by the personalized policy (which depends on random user histories) versus the fixed editorial ranking. Further, for a subpopulation $\mathcal{G} \subseteq \{1, \ldots, N\}$, the conditional average treatment effect is:

$$\tau_t(\mathcal{G}) = \mathbb{E}[Y_{it}(\pi_t^P(H_{it})) - Y_{it}(r_t^E) \mid i \in \mathcal{G}]. \tag{4}$$

4.2 Identification

We employ two complementary strategies: a randomized controlled trial for establishing proof-of-concept, followed by a regression discontinuity design to estimate the value of the recommendation system in scaled deployment over five months.

Randomized Controlled Trial. During the experimental period t^*, we randomly assigned eligible users to treatment ($D_i = 1$) or control ($D_i = 0$), stratifying on grade level, acquisition channel, and prior engagement quartile.

Assumption 4.1 (Randomization)**.** Treatment assignment is independent of potential outcomes under any ranking:

$$D_i \perp\!\!\!\perp \{Y_{it^*}(r) : r \in \mathcal{R}\}. \tag{5}$$

Under this assumption, the average treatment effect is identified:

$$\tau_{t^*} = \mathbb{E}[Y_{it^*}(\pi_{t^*}^P(H_{it^*}))] - \mathbb{E}[Y_{it^*}(r_{t^*}^E)] = \mathbb{E}[Y_{it^*} \mid D_i = 1] - \mathbb{E}[Y_{it^*} \mid D_i = 0]. \tag{6}$$

where the equality follows from Assumption 4.1.

Regression Discontinuity Design. Following the experiment, personalized recommendations became available to all users satisfying an eligibility criterion. As defined above, for user i in period t, let X_{it} denote cumulative story interactions at the beginning of period t. Users receive personalized recommendations when their interaction count reaches the

threshold $c = 60$. Users are unaware of this threshold. Treatment assignment is deterministic: $D_{it} = \mathbf{1}\{X_{it} \geq c\}$. We adopt the local randomization framework provided in (Cattaneo et al., 2015, 2018). Define a window $\mathcal{W}_h = [c - h, c + h)$ around the threshold.

Assumption 4.2 (Local Randomization)**.** Within the window $\mathcal{W}_h$, the running variable is as-if randomly assigned with respect to potential outcomes:

$$Y_{it}(r) \perp\!\!\!\perp X_{it} \mid X_{it} \in \mathcal{W}_h, \quad \forall r \in \mathcal{R}. \tag{7}$$

This assumption requires that, conditional on being near the threshold, the exact interaction count is uninformative about potential outcomes. Since policies affect outcomes only through the rankings they generate, this assumption ensures valid causal inference at the threshold. The local average treatment effect is:

$$\begin{aligned}\tau_t^{RDD} &= \mathbb{E}[Y_{it}(\pi_t^P(H_{it})) \mid X_{it} \in \mathcal{W}_h] - \mathbb{E}[Y_{it}(\pi_t^E) \mid X_{it} \in \mathcal{W}_h] \\ &= \mathbb{E}[Y_{it} \mid D_{it} = 1, X_{it} \in \mathcal{W}_h] - \mathbb{E}[Y_{it} \mid D_{it} = 0, X_{it} \in \mathcal{W}_h]\end{aligned}$$

We adopt the local randomization framework rather than continuity-based RDD for three reasons. First, the 60-story threshold is mechanical and unknown to users, making the local randomization assumption—that the exact interaction count within a narrow window around the threshold is orthogonal to potential outcomes— plausible. The threshold itself has no direct effect on outcomes beyond determining treatment assignment. Second, with relatively few users crossing the threshold each week, randomization inference provides exact finite-sample p-values without relying on large-sample approximations required by continuity-based methods. Third, this approach avoids the auxiliary specification choices inherent to continuity-based RDD (kernel selection, polynomial order, optimal bandwidth algorithms), focusing identification entirely on the credibility of local randomization within a pre-specified window validated through placebo tests and using the RCT data. (Cattaneo et al., 2015).

Further, the finite-sample randomization inference (Abadie et al., 2014; Imbens and Rubin, 2015) approach provides exact finite-sample validity. This ensures that the method does not rely on parametric distributional assumptions and asymptotic approximations. We select bandwidth h using placebo tests on experimental data, choosing the largest bandwidth for which users just below threshold exhibit outcomes statistically indistinguishable from randomized control users (Section 7.1).

Estimand Comparison. The RCT identifies τ_{t^*}, the ATE for eligible users during the experimental period. The RDD identifies τ_t^{RDD}, the LATE for users at the threshold margin

in period t. These differ in target population and, for $t \neq t^*$, algorithm version. Comparing them assesses whether experimental findings transport to scaled deployment. Moreover, comparing RDD estimates across periods reveals how effects evolve as the algorithm improves.

5 Developing the Recommendation System

This section describes how we designed the personalized recommendation system. We made three key decisions: i) which prediction model to use, ii) which users should be eligible for personalized recommendations, and iii) where in the app to deploy the system.

5.1 The Personalization Opportunity at Moderate Scale

The central empirical question motivating our analysis is whether personalization techniques developed for large technology platforms can function effectively with the data volumes typical of educational technology app.

We adopt collaborative filtering via matrix factorization as our personalization approach. This method became the dominant paradigm in recommendation systems following the Netflix Prize competition, where matrix factorization substantially outperformed alternatives and formed the core of the winning solution (Koren et al., 2009; Bell and Koren, 2007). We favor collaborative filtering over content-based alternatives because of the structure of available data. Freadom's infrastructure precisely tracks user-story interactions, which stories users view, start, and complete, providing rich behavioral data on millions of interactions. User characteristics, by contrast, are sparse. We observe grade level and acquisition channel, but key determinants of reading preferences such as prior reading experience, English proficiency, and intrinsic interests are unobserved. Collaborative filtering exploits the data we have in abundance rather than relying on covariates we lack, learning preference patterns directly from revealed choices.

Scale matters because collaborative filtering requires sufficient user-story interactions to identify systematic preference patterns (Koren et al., 2009). The method decomposes the interaction matrix into low-dimensional latent factors capturing underlying preferences. With sparse data, the algorithm cannot distinguish genuine preference heterogeneity from noise. This requirement creates a fundamental constraint. Platforms must accumulate substantial interaction histories before personalization becomes feasible, and users with limited histories cannot receive reliable recommendations. Our setting allows us to examine whether a moderate-scale educational platform generates sufficient data to support effective personal-

ization, and how treatment effects vary with the richness of user histories.

5.2 Modeling Development and Design Choices

We observe a dataset of user-story interactions over time, where each observation U_{ij} records user i's engagement with story j. Stories never shown to a user do not enter the dataset. Our objective is to predict engagement for user-story interactions that have not yet been observed, enabling recommendations of stories that users are likely to enjoy but have not yet encountered.

We compare two prediction models. The first is a popularity-based model estimated via two-way fixed effects (TWFE):

$$\mathbb{E}[U_{ij}] = \sigma(\beta_0 + \psi_i + \gamma_j), \tag{8}$$

where ψ_i is a user fixed effect, γ_j is a story fixed effect, and $\sigma(\cdot)$ is the sigmoid function. This model captures systematic differences in user engagement levels and the story's overall popularity. Conditional on user baseline engagement, stories are ranked by γ_j: their average engagement across all users. The TWFE model thus provides identical rankings to all users.

The second is a collaborative filtering model based on matrix factorization (Koren et al., 2009; Rendle, 2010):

$$\mathbb{E}[U_{ij}] = \sigma(\theta_i'\lambda_j + \beta_0 + \psi_i + \gamma_j), \tag{9}$$

where θ_i is a latent preference embedding per user i and λ_j is a latent embedding per story j. The key addition is the interaction term $\theta_i'\lambda_j$, which captures user-story match quality beyond overall popularity and baseline engagement. Users with similar embedding vectors share similar preferences; stories with similar embeddings appeal to similar audiences. This model generates personalized rankings that vary across users.

Both models include user and story fixed effects, so the performance difference between the two models is driven by the latent embeddings capturing preference heterogeneity. If collaborative filtering substantially outperforms TWFE, this constitutes evidence of exploitable preference heterogeneity, the necessary condition for personalization to add value.

We estimate both models on observed interactions from history, minimizing mean squared error (MSE) with L_2 regularization. The objective functions for the two approaches differ in their regularization, as collaborative filtering has additional parameters relative to TWFE. Formally, for collaborative filtering, the following objective is minimized:

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{S}} (U_{ij} - \hat{U}_{ij})^2 + \lambda_\theta \sum_i \|\theta_i\|^2 + \lambda_\gamma \sum_j \|\lambda_j\|^2 + \lambda_\psi \sum_i \psi_i^2 + \lambda_\beta \sum_j \gamma_j^2, \tag{10}$$

where $\mathcal{S}$ denotes the set of observed user-story interactions in the training data. We do not impute values for unobserved pairs. They simply do not enter the loss function. Estimation proceeds via stochastic gradient descent using the Adam optimizer (Adam et al., 2014). Regularization prevents overfitting for users or stories with sparse interaction histories.

After estimation, we can predict engagement for any user-story pair, including pairs never observed in the training data. User i's embedding θ_i is learned from her interactions with stories in her history; story j's embedding λ_j is learned from interactions by users who engaged with that story. The latent factor structure enables information sharing: if user i has never interacted with story j, but other users with similar embeddings to θ_i engaged heavily with story j, then λ_j will align on the same latent dimensions as θ_i, yielding high predicted engagement $\theta_i'\lambda_j$ even though we never observed that specific pair.

Model Selection. We evaluate model performance using historical data from July 2020 to July 2021. The dataset contains all logged user-story interactions during this period, comprising almost 9 million observations across approximately 170,000 users. We split the data temporally. The interactions from the first eleven months form the training set, while the final month serves as the held-out evaluation set. This temporal split mirrors the prediction task in deployment, where models trained on historical data must predict engagement with future interactions.

We compare three specifications: (i) a constant-only model that predicts the training set mean for all observations, providing a lower bound on performance; (ii) the TWFE model from equation (8); and (iii) the collaborative filtering approach from Equation (9). Appendix A2 provides details of the model estimation. Table 2 reports mean squared error on the held-out evaluation set.

Table 2: Prediction Model Comparison

Model	**MSE**
Mean (constant only)	0.131
Two-way fixed effects	0.102
Collaborative Filtering	0.096

Note: Mean squared error on held-out evaluation set. The mean model predicts the training set average for all observations. The TWFE model includes user and story fixed effects. The collaborative filtering model adds $k = 20$ latent dimensions for users and stories.

The TWFE model has a 22 percent lower MSE relative to the constant-only baseline, indicating that user and story fixed effects capture substantial variation in engagement.

Collaborative filtering yields a further 6 percent reduction relative to TWFE. This indicates that user-story match quality varies systematically beyond overall popularity and baseline engagement—precisely the preference heterogeneity that personalization aims to exploit.

Threshold Design. The performance of collaborative filtering depends on the richness of interaction histories. Users with few prior interactions provide limited data for estimating preference embeddings; similarly, stories with few interactions yield unreliable story embeddings. We therefore restrict personalized recommendations to users and stories exceeding minimum interaction thresholds.

To select appropriate thresholds, we evaluate model performance across different minimum history requirements. We train the collaborative filtering model on the whole training set, then compute MSE on evaluation subsets defined by varying user and story interaction counts. Table 3 reports results.

Table 3: Model Performance by Interaction History

	Minimum Story Interactions		
Minimum User Interactions	20	60	100
20	0.097	0.093	0.093
60	0.096	0.093	0.093
100	0.096	0.093	0.093

Note: Mean squared error by minimum interaction history. Rows indicate minimum prior interactions per user; columns indicate minimum prior interactions per story. Model trained on full training set; MSE computed on evaluation set subsets satisfying the indicated thresholds.

Performance improves as interaction histories lengthen, with the largest gains occurring between 20 and 60 interactions. Beyond 60 interactions, improvements are negligible. This pattern reflects the learning curve of collaborative filtering: initial interactions are highly informative about user preferences, but marginal returns diminish as embeddings stabilize.

Based on this analysis, we select a threshold of 60 prior story interactions for user eligibility. This threshold balances predictive accuracy against sample coverage. At the time of the experiment, approximately 15 percent of active users met this criterion, representing the platform's most engaged segment. For stories, over 90 percent of the content library exceeded 60 interactions, so the story threshold imposes minimal restrictions on available content.

The threshold creates a fundamental scaling constraint. New users and users with limited engagement histories cannot receive personalized recommendations until they accumulate

sufficient interaction data. This is the cold-start problem inherent to collaborative filtering. We do not attempt to solve this problem in our implementation; users below the threshold continue receiving editorial recommendations. The heterogeneity analysis in Section 6 examines how treatment effects vary with interaction history, providing evidence on how personalization gains scale in this setting with data availability.

Section Selection. Freadom organizes content into multiple sections, each displaying a slate of 15 stories. Sections vary in prominence as those appearing higher on the screen receive more traffic. We deploy personalized recommendations in the *Recommended Story* section for three reasons.

First, this section was originally intended for personalized recommendations, as its name suggests, though it had been operating with editorial curation prior to our intervention. Second, the section occupies a prominent position in the app, ensuring sufficient user traffic. Third, users of this section were disproportionately likely to meet our 60-interaction threshold, reflecting its popularity among engaged users. Other sections, including *Trending Now*, *Top Stories*, and *New Release*, continued using their existing algorithms throughout the experiment, providing a natural benchmark for assessing substitution across sections.[3]

5.3 Deployment

The *Recommended Story* section contains a slate of 15 stories. Each week, the algorithm generates predicted engagement for all eligible user-story pairs and assigns each user a slate of 15 top-ranked stories. Slates update daily as users engage with content. The control group receives editorial recommendations with identical slate size and refresh logic. Appendix A3 provides full implementation details.

6 Randomized Controlled Trial

We evaluate the personalized recommendation system through a randomized experiment. The objective is to compare user engagement under the existing editorial-based system versus the personalized recommendation system.

[3]Initially we selected an additional section, *Today for You*, and intended to deploy personalized recommendations on both sections. However, during the deployment, we encountered engineering and coordination challenges and decided to proceed with only one personalized section.

6.1 Experimental Design

We randomized 7,750 users with at least 60 prior story interactions into treatment and control groups, stratifying on student grade, user category, and quartile of past engagement. The treatment group received personalized recommendations in the *Recommended Story* section; the control group continued receiving editorial recommendations. The user interface remained identical across groups; only the displayed stories varied. Other app sections were unchanged, and users were not informed of the experiment. The experiment ran from July 22 to August 4, 2021, powered to detect a minimum effect of 0.08 standard deviations on *Total Story Engagement*.[4]

During the experimental period, 3,023 users launched the app at least once, of whom 525 viewed at least one story in the *Recommended Story* section.[5] These users generated 77,942 story interactions. Our analysis includes all users who launched the app, regardless of whether they engaged with the *Recommended Story* section. This choice reflects two considerations: users can observe the first story in the section without clicking, and the intervention may affect the extensive margin of section usage. Appendix A4 confirms covariate balance between treatment and control groups.

6.2 Engagement Outcomes

We examine two categories of outcomes: section-specific effects within *Recommended Story* and app-wide effects across all sections. The latter is important because changes to one section may induce substitution or complementarities in usage elsewhere.

Our primary outcome is *Total Story Engagement*, defined as the sum of *Story Engagement* values across all user-story interactions during the experimental period (defined as Y_{it} in section 4). Our secondary outcome is *Total Stories Completed*, the count of stories read to completion. We also measure total reading time, both within the *Recommended Story* section and across the entire app, as well as the number of difficult stories completed app-wide.

Table 4 presents summary statistics. Panel A reports outcomes for the *Recommended Story* section; Panel B reports app-wide outcomes.

6.3 The Impact of Personalization on Engagement

Table 5 presents our main results: average treatment effects of the personalized recommendation system on user engagement, estimated via difference-in-means. We report effects on

[4]Appendix A7 provides a summary of the pre-analysis plan for this experiment.

[5]The gap between randomized and active users reflects continuous attrition, as students drop off the platform over time.

Table 4: Summary Statistics of Experimental Outcomes

Outcome	Group	Mean	75th %ile	90th %ile	95th %ile	Max
Panel A: Recommended Story Section						
Total Story Engagement	Control	0.28	0	0.51	1.6	13.6
	Treatment	0.45	0	1.3	3	23.9
Total Stories	Control	0.15	0	0	1	11
	Treatment	0.27	0	1	2	21
Total Reading Time	Control	1.04	0	0	7.5	78.5
	Treatment	1.94	0	7.25	11	148.5
Panel B: All Sections						
Total Story Engagement	Control	3.79	4.47	11.4	17.77	81.5
	Treatment	4.32	5.5	13.5	19.02	63.7
Total Stories	Control	1.89	2	6	9	41
	Treatment	2.25	3	7	11	42
Total Reading Time	Control	12.65	14.5	40	65.32	273
	Treatment	15.15	15	46	73.12	365
Number of Difficult Stories	Control	1.87	2	6	9	41
	Treatment	2.24	3	7	11	41

Note: This table reports summary statistics for outcome variables by treatment status. Panel A presents outcomes measured within the Recommended Story section where personalized recommendations were deployed. Panel B presents outcomes measured across all sections of the app. Sample includes all users who launched the app during the experiment period ($N = 3{,}023$).

four outcomes: *Total Story Engagement* and *Stories Completed*, each measured both app-wide (columns 1–2) and within the *Recommended Story* section (columns 3–4). For each outcome, we present the raw treatment effect, its standard error, and the effect expressed as a percentage of the control group mean.

Personalized recommendations substantially increased engagement. Within the *Recommended Story* section, treated users exhibited 60 percent higher *Total Story Engagement* (S.E. 18%) and completed 77 percent more stories (S.E. 24%) relative to the control group. Importantly, these section-specific gains did not come at the expense of engagement elsewhere. App-wide *Total Story Engagement* increased by 14 percent (S.E. 7%)) and *Stories Completed* by 19 percent (S.E. 8%). The positive app-wide effects indicate that users did not substitute away from other, non-personalized sections of the app. Rather, personalization induced users to consume more content overall — both within the treated section and across the platform.

Figure 2 provides further evidence against substitution across sections. The figure plots treatment-control differences in mean *Total Story Engagement* for each of the app's major sections, with 95 percent confidence intervals (CI). The *Recommended Story* section, where personalization was implemented, shows a large, statistically significant increase in engage-

Table 5: Average Treatment Effects of Personalized Recommendations

	Outcomes			
	Total Engagement (All)	Stories Completed (All)	Total Engagement (RS)	Stories Completed (RS)
ATE	0.522*	0.359**	0.171***	0.119***
SE	(0.27)	(0.16)	(0.052)	(0.037)
ATE % baseline	13.767*	18.979**	60.265***	77.478***
SE	(7.117)	(8.465)	(18.212)	(24.139)
No. obs. treated	1376	1376	1376	1376
No. obs. control	1430	1430	1430	1430

*Note: Average treatment effects of personalized recommendations estimated via difference-in-means. Standard errors in parentheses. ATE % baseline expresses effect as percentage of control mean. *** p<0.01, ** p<0.05, * p<0.10.*

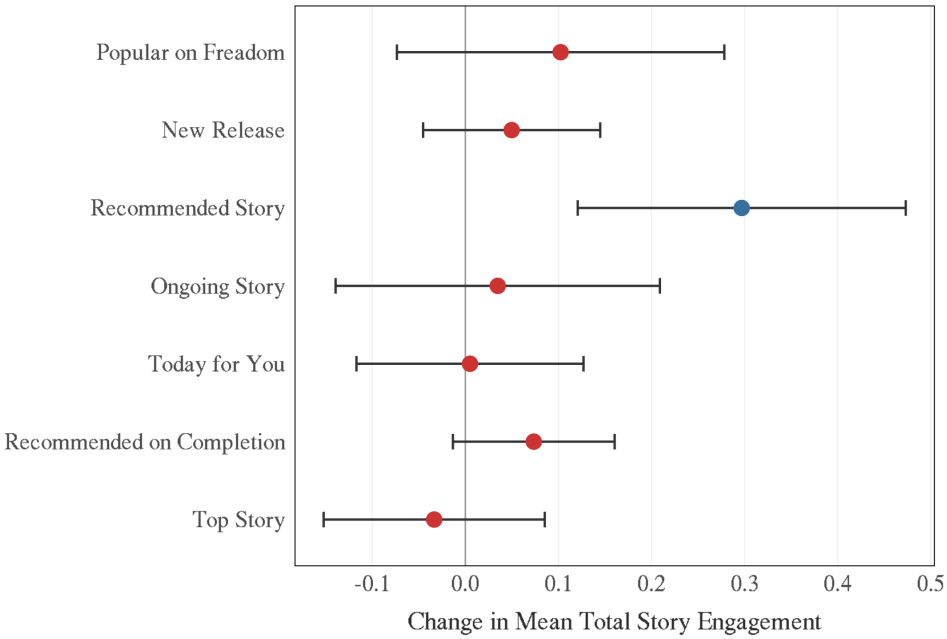
ment at a 5% level of significance. Crucially, 95% CIs for engagement in all other sections include 0. If anything, point estimates for sections such as *Popular on Freadom* and *Recommended on Completion* are positive, though imprecisely estimated. This pattern reinforces our interpretation that personalized recommendations expanded overall engagement rather than merely reallocating attention across sections.

Figure 3 displays the cumulative distribution functions of *Total Story Engagement* in the *Recommended Story* section for treatment and control groups. The treatment distribution first-order stochastically dominates the control distribution: at every level of engagement, a smaller fraction of treated users falls below that threshold. The divergence is particularly pronounced at the lower end of the distribution. Approximately 87 percent of control users had zero engagement with the section, compared to 82 percent of treated users. This five percentage point difference on the extensive margin, inducing users to engage with the section at all, accounts for a substantial share of the overall treatment effect. The distributions also diverge at higher engagement levels, indicating that personalization increased engagement on the intensive margin as well. We formally test for distributional differences using the Wilcoxon rank-sum test, rejecting the null hypothesis of no location shift for both section-specific engagement ($p < 0.001$) and app-wide engagement ($p = 0.05$).

6.4 Effects on Learning Inputs

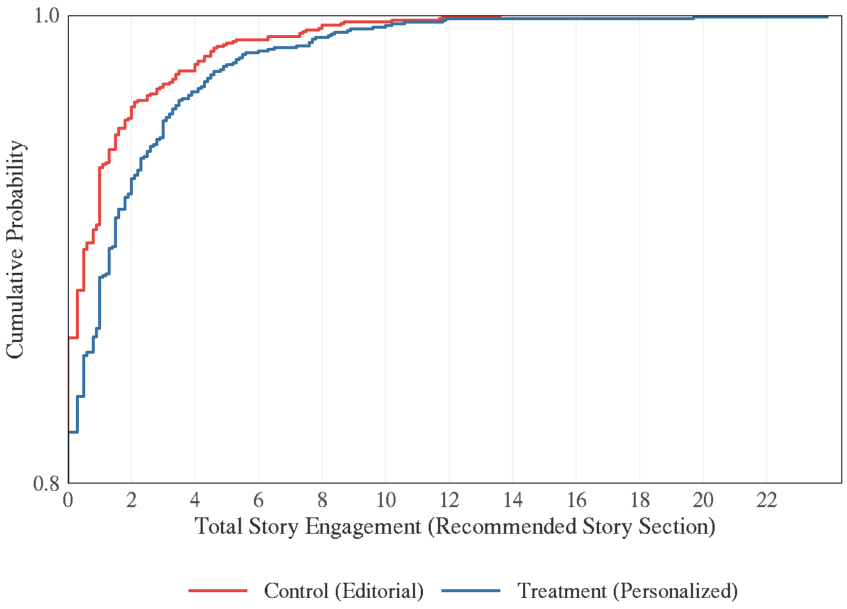
While we do not directly measure learning outcomes, we can examine treatment effects on outcomes that plausibly serve as inputs to learning: time spent reading and exposure

Figure 2: Average Treatment Effects for Main Sections of the App



Note: Treatment-control differences in mean Total Story Engagement *by app section, with 95 percent confidence intervals. The* Recommended Story *section (blue) received personalized recommendations; all other sections (red) maintained editorial recommendations throughout the experiment.*

Figure 3: Distribution of Engagement in the Recommended Story Section



Note: Cumulative distribution functions of Total Story Engagement *in the* Recommended Story *section by treatment status. The y-axis begins at 0.8 to highlight variation in the upper portion of the distribution.*

to challenging content. These measures also provide a useful check on the recommendation algorithm. If personalization increased engagement by steering users toward shorter or easier stories, we might question whether the engagement gains translate into educational value. The results in Table 6 suggest this is not the case.

We construct two outcome measures. *Total Reading Time* is calculated by summing the estimated reading duration of all stories a user completed during the experiment, where the publisher provides reading duration for each story. We measure this both within the *Recommended Story* section and app-wide. *Number of Difficult Stories* captures exposure to challenging content. Each story is assigned a grade level by the publisher. We classify a story as difficult for a given user if its grade level exceeds the user's grade level. The outcome is the count of such stories completed during the experiment, measured app-wide.

Table 6: Average Treatment Effects of Personalized Recommendations on Learning Inputs

	Outcomes		
	Total Reading Time (All)	Total Reading Time (RS)	No. of Difficult Stories (All)
ATE	2.501**	0.899***	0.369**
SE	(1.104)	(0.265)	(0.16)
ATE % baseline	19.781**	86.713***	19.713**
SE	(8.733)	(25.534)	(8.563)
No. obs. treated	1376	1376	1376
No. obs. control	1430	1430	1430

*Note: Average treatment effects of personalized recommendations estimated via difference-in-means. Standard errors in parentheses. ATE % baseline expresses effect as percentage of control mean. *** p<0.01, ** p<0.05, * p<0.10.*

Table 6 presents average treatment effects on these learning-related outcomes. Personalization significantly increased reading time: treated users spent 87 percent more time reading in the *Recommended Story* section (S.E. 26%) and 20 percent more time reading app-wide (S.E. 9%). Treated users also completed 20 percent more difficult stories (S.E. 9%). These results indicate that the engagement gains documented in the previous section were not driven by the algorithm recommending shorter or less challenging content. If anything, personalization led users to spend more time reading and to engage with more difficult material. These patterns are consistent with increased learning inputs.

6.5 Heterogeneous Treatment Effects

We examine treatment effect heterogeneity across levels of prior engagement. This analysis is important for understanding the scalability of the personalized recommendation system along two dimensions. First, our experimental design required users to have at least 60 prior story interactions, excluding a substantial portion of the user base. If treatment effects are similar for users closer to this threshold, the personalization system could potentially be extended to users with shorter histories by lowering the eligibility threshold. We anticipate that the algorithm improves as it accumulates more data on user preferences, but quantifying the gradient of treatment effects across engagement levels informs decisions about threshold selection. Second, we examine whether personalization attracts new users to the *Recommended Story* section, users who previously did not engage with this part of the app.

We define three binary indicators of prior engagement. *Heavy Engagement* equals one if a user's pre-experiment *Total Story Engagement* falls above the sample median, and zero otherwise. *Heavy Completion* equals one if a user's pre-experiment count of completed stories exceeds the sample median. *New RS User* equals one if a user had zero story interactions in the *Recommended Story* section during the two weeks prior to the experiment, indicating no recent engagement with the section that would later receive personalized recommendations.

Table 7 presents conditional average treatment effects (CATEs) for each subgroup. Panel A reports effects on *Total Story Engagement* in the *Recommended Story* section; Panel B reports effects on *Stories Completed*. For each subgroup definition, we report the control group mean (Baseline) and the CATE separately for users with the indicator equal to one versus zero. The final column tests whether CATEs differ significantly across subgroups.

The results reveal substantial heterogeneity. Heavy engagement users exhibit CATEs roughly three to six times larger than light users: for *Total Story Engagement*, the CATE is 0.298 (SE = 0.089) for heavy users versus 0.052 (SE = 0.052) for light users, a statistically significant difference at 5% level of significance ($p < 0.05$). Similar patterns emerge for *Heavy Completion*. These findings are consistent with the collaborative filtering algorithm performing better when it has richer data on user preferences. However, light users still experience positive, albeit smaller, and imprecisely estimated, treatment effects. This suggests that lowering the inclusion threshold would yield gains, though potentially lower than the estimates reported here.

The results for new section users are particularly encouraging for scalability. Users with no recent engagement in the *Recommended Story* section exhibit statistically significant treatment effects: a CATE of 0.063 (SE = 0.035) for *Total Story Engagement* and 0.042 (SE = 0.019) for *Stories Completed*. In percentage terms, these effects are large relative to the low baseline engagement of this group. Personalization successfully attracted users

Table 7: Heterogeneous Treatment Effects: Total Engagement and Stories Completed

	Subgroup = 1		Subgroup = 0		
	Baseline	CATE	Baseline	CATE	Diff. Subgroups
Panel A: Total Engagement (Recommended Story Section)					
Heavy engagement	0.389 (0.047)	0.298 (0.089)	0.174 (0.035)	0.052 (0.052)	0.246 (0.103)
Heavy completion	0.416 (0.051)	0.251 (0.089)	0.162 (0.032)	0.076 (0.051)	0.175 (0.103)
New RS user	0.053 (0.021)	0.063 (0.035)	0.493 (0.052)	0.255 (0.09)	-0.192 (0.097)
Panel B: Stories Completed (Recommended Story Section)					
Heavy engagement	0.221 (0.032)	0.199 (0.066)	0.084 (0.024)	0.044 (0.034)	0.155 (0.074)
Heavy completion	0.247 (0.036)	0.169 (0.067)	0.068 (0.02)	0.059 (0.032)	0.111 (0.074)
New RS user	0.015 (0.009)	0.042 (0.019)	0.28 (0.037)	0.18 (0.067)	-0.139 (0.069)

Note: Estimates of average treatment effects conditioned on subgroups. Panel A presents effects on Total Story Engagement in the Recommended Story section; Panel B presents effects on Stories Completed. Columns 2 and 3 describe the values for subgroups when the indicator takes the value 1, while columns 4 and 5 describe the values when the indicator takes the value 0. The last column shows the difference in the effectiveness of personalized recommendations between the two subgroups. Columns 2 and 4 are the average outcomes in the control groups (Baseline), and columns 3 and 5 are the estimates of the conditional average treatment effect (CATE). Standard errors are in parentheses.

who otherwise would not have engaged with the section, demonstrating that the system can expand its reach beyond existing users. While the absolute magnitude of effects is larger for established section users, the ability to draw in new users is critical for scaling personalization across the platform.

7 Scaled Deployment and Regression Discontinuity

Following the experiment, the personalized recommendations system was scaled to all eligible users beginning August 5, 2021. This section evaluates whether the treatment effects estimated in the experiment persist during real-world deployment. To do that we use a regression discontinuity design that exploits the 60-interaction eligibility threshold. Our RDD analysis compares users in their first week of personalized recommendations to users who remain just below the eligibility threshold in that week. Specifically, for each week t during the scaled deployment period, we construct treatment and control groups as follows:

Treatment group: Users who (i) had fewer than 60 cumulative interactions at the start of week t, (ii) crossed the 60-interaction threshold during week t, and (iii) fall within the bandwidth above the threshold (60 to $60 + h$ interactions) at the end of week t. These users receive personalized recommendations for the first time in week $t + 1$.

Control group: Users who (i) had fewer than 60 cumulative interactions at the start of week t, (ii) remained below 60 interactions throughout week t, and (iii) fall within the bandwidth below the threshold ($60 - h$ to 59 interactions) at the end of week t. These users continue receiving editorial recommendations in week $t + 1$.

We measure engagement in week $t + 1$. Critically, users who cross the threshold are evaluated only in their first week of personalized recommendations and are excluded from subsequent weeks' RDD samples. This ensures that neither the treatment nor control group has prior exposure to personalized recommendations, mirroring the RCT design where both groups compare first exposure to personalization versus editorial recommendations.

As users continue engaging with the app, different cohorts cross the eligibility threshold each week, creating continuous quasi-experimental variation. The 60-interaction threshold is exogenous, built into the app's design, and unknown to users.

The recommendation system itself evolved during the deployment period. We retrained the collaborative filtering model weekly on all accumulated interaction data, enabling periodic improvement. This means the "treatment" (personalized recommendations) is not fixed but iteratively updated over time as the algorithm learns from expanding data. We estimate local average treatment effects week by week using the local randomization RDD framework discussed in Section 4.

7.1 Validation Using Experimental Data

Before analyzing the scaled deployment, we use the experimental data to validate the RDD design and to select the optimal bandwidth.

Using data from the RCT period, we construct three user groups: *Treatment*, *Control RCT*, *Control RDD*. The former two are from the experiment and consists of users who had at least 60 story interactions before the experiment. The *Control RDD* group consists of users with fewer than 60 interactions but more than $60 - h$, where h is the bandwidth. These users were ineligible for the RCT but fell just below the eligibility threshold, making this group and equivalent of the control group in the RDD design. To validate the RDD design, the difference between the mean outcomes of the two control groups should be small and the difference between treatment and *Control RDD* group to resemble the experimental ATE. Table 8 presents the validation analysis where we vary the bandwidths between 20 and 60 in increments of 5. The effects are estimated using the local randomization RDD estimator in Cattaneo et al. (2015).

For bandwidths ≤ 35, the tests comparing *Control RCT* vs *Control RDD* yield p-values ranging from 0.169 to 0.440. This means that under the null hypothesis of no difference

Table 8: Placebo Tests: Total Engagement

	Control RCT vs Control RDD				Treatment vs Control RDD			
BW	N(L)	N(R)	Est.	P-val	N(L)	N(R)	Est.	P-val
20	1616	344	0.080	0.169	1616	360	0.131	0.029
25	2182	407	0.049	0.440	2182	417	0.147	0.105
30	2951	469	0.046	0.371	2951	493	0.205	0.016
35	3872	533	0.047	0.303	3872	541	0.209	0.001
40	4999	598	0.082	0.073	4999	587	0.226	0.000
45	6555	645	0.085	0.046	6555	629	0.250	0.000
50	8522	702	0.103	0.024	8522	675	0.248	0.000
55	11298	748	0.139	0.001	11298	707	0.256	0.000

Note: This table reports placebo tests using local randomization RDD with varying bandwidths around the threshold of 60 story interactions. The Control RDD group consists of users who were outside the RCT (neither Treatment nor Control RCT) and fall below the threshold. For each bandwidth w, the window is $[60 - w, 60 + w)$. Estimates represent difference in means between groups. P-values are computed using randomization inference under Fisher's sharp null hypothesis of no treatment effect. $N(L)$ corresponds to the number of observations below the threshold and $N(R)$ corresponds to the number of observations above the threshold.

between these groups, we would observe effect estimates as large as or larger than what we saw in 17–44% of random permutations. These relatively high p-values provide no evidence against the null of no difference, consistent with local randomization holding in narrower windows. For wider bandwidths (> 35), the p-values decrease below 0.05, indicating that such large differences would be unlikely under random assignment alone, suggesting potential violations of the local randomization assumption.

For the *Treatment* vs *Control RDD* comparison, we observe a different pattern. Across all bandwidths except one, p-values fall below 0.05, and many approach zero. Under the null hypothesis of no difference between these groups, we would observe effect estimates this large in fewer than 5% of random permutations. These results are consistent with two interpretations: either personalization increases engagement, or users just above and below the threshold differ systematically in ways that affect outcomes.

Taken together, these results support the validity of our RDD design. The comparison of two control groups shows that users just below the threshold are similar to users just above who received editorial recommendations, at least within a bandwidth of 35. This rules out the possibility that users above and below the threshold differ systematically in ways that affect engagement. The *Treatment* vs *Control RDD* comparison can therefore be attributed to personalization rather than pre-existing differences. We conclude that the *Control RDD* group constitutes a valid counterfactual for treated users and select a bandwidth of 35 for

the main analysis, the largest for which we have evidence supporting local randomization.

7.2 RDD Estimates from Scaled Deployment: Weekly Treatment Effects

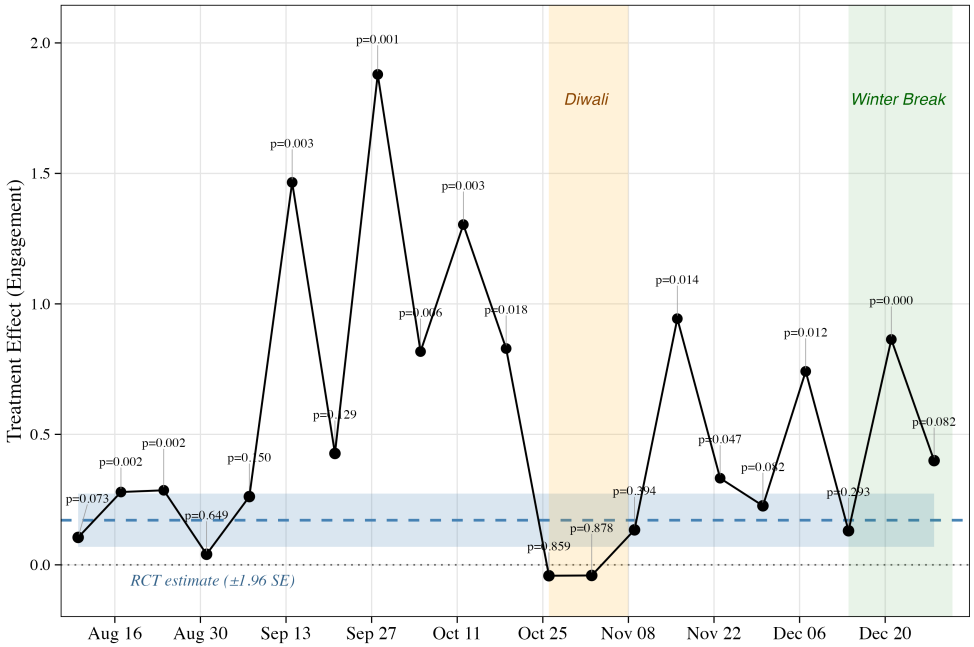
We now analyze the scaled deployment period (August 5 – December 31, 2021) using the regression discontinuity design validated in Section 7.1. Figure 4 presents local average treatment effects estimated separately for each week of the deployment period using local randomization RDD with a bandwidth of 35 interactions around the eligibility threshold. Each point represents the difference in mean *Total Story Engagement* in *Recommended Story* section between users with 60–94 prior interactions (treated) and users with 25–59 prior interactions (control) during that week. The dashed horizontal line and shaded region denote the RCT point estimate and 95 percent confidence interval, respectively, providing a benchmark for comparison.

The weekly estimates suggest three distinct phases. First, the several weeks following scale-up yield treatment effects broadly consistent with the experimental benchmark. Point estimates range from 0.1 to 0.3, within the 95% confidence intervals of the RCT estimate of 0.171 (SE = 0.052). This initial consistency suggests that the personalization system maintained its effectiveness as it transitioned from controlled experimentation to full deployment.

Second, beginning in mid-September, treatment effects increase substantially above the RCT benchmark. Point estimates rise to 0.8–1.9 during several weeks, with p-values below 0.01 in 5 out of the 7 weeks. This pattern is consistent with the recommendation system improving as it accumulated additional training data through weekly retraining: users crossing the eligibility threshold during this period encountered a more refined algorithm than participants in the initial RCT. Notably, these gains coincide with active school term periods when baseline engagement is highest, suggesting that personalization benefits are amplified when users interact more intensively with the platform.

Third, treatment effects decline substantially during late October and early November (indicated by orange shading). Point estimates fall to approximately zero, with p-values exceeding 0.85, providing no evidence of personalization benefits during this period. This decline coincides with the Diwali festival, when baseline app usage drops sharply as children's routines are interrupted. Lower overall engagement may contribute to reduced personalization benefits. Following this period, effects recover but exhibit greater volatility through December. Some weeks show significant positive effects comparable to the pre-holiday period ($p < 0.05$), while others yield null results ($p > 0.20$). The Winter Break period (green shading) similarly shows mixed results, though with less severe disruption than Diwali. This

Figure 4: Weekly Local Average Treatment Effects: Scaled Deployment Period



Note: Local average treatment effects of personalized recommendations on Total Story Engagement, estimated separately for each week. The dashed horizontal line denotes the RCT point estimate (0.171) with 95 percent confidence interval. Orange shading indicates the Diwali period; green shading indicates the Winter Break. Values above the dots show the p-values.

temporal heterogeneity underscores the importance of accounting for contextual factors, particularly the Indian school calendar, when evaluating sustained intervention effects.

To verify that our identification assumption holds throughout the deployment period, we conduct two checks. First, we compare the average *Total Story Engagement* of the treated (60-94) and control groups (25-59) in the week prior to crossing the 60-story interaction threshold. Results in section A5.1 shows that the treated and control groups exhibit comparable engagement patterns prior to crossing the threshold. This supports the validity of the regression discontinuity design. The absence of systematic pre-treatment differences between users just above versus just below the threshold indicates that crossing 60 interactions does not select for different user types on this outcome dimension.

Second, weekly placebo tests comparing treatment effect estimates in the *Recommended Story* section against the *Top Story* section, where editorial recommendations continued throughout. The *Top Story* section shows estimates clustered around zero in majority of weeks (mean effect: 0.04), while the *Recommended Story* section exhibits large, frequently significant effects (mean effect: 0.58). This confirms that the discontinuity at 60 interactions captures the effect of personalization rather than user maturation or selection at the

threshold. See Appendix A5.2 for details.

7.3 Pooled RDD Estimates

To characterize average treatment effects across the full deployment period and assess robustness to analytic choices, we pool observations across all weeks and estimate local average treatment effects for varying bandwidths. Table 9 presents these results.

Table 9: Pooled RDD Estimates: Scaled Deployment Period

Bandwidth	Estimate	p-value	N (Control)	N (Treated)	N (Total)
20	0.401	< 0.001	6885	1940	8825
25	0.407	< 0.001	8879	2103	10982
30	0.427	< 0.001	10952	2210	13162
35	0.424	< 0.001	13062	2295	15357
40	0.430	< 0.001	15302	2376	17678
45	0.426	< 0.001	17493	2461	19954
50	0.426	< 0.001	19487	2551	22038
55	0.428	< 0.001	21054	2610	23664

Note:Local average treatment effects estimated via local randomization. The running variable is prior cumulative story interactions minus 60. For each bandwidth h, the estimation window is $[60 - h, 60 + h)$. The outcome is Total Story Engagement in the Recommended Story section during the week following threshold crossing. Sample includes all user-week observations from August 5 – December 31, 2021.

The pooled estimates are stable across considered bandwidths. Point estimates range from 0.401 to 0.430, with no systematic pattern as the bandwidth expands from 20 to 55 interactions. All estimates are highly statistically significant ($p < 0.001$), reflecting the substantial sample sizes, over 15,000 user-week observations at our preferred bandwidth of 35, that accumulate over five months of deployment.

The pooled estimates substantially exceed the RCT benchmark of 0.171. At the preferred bandwidth of 35, the local average treatment effect is 0.424, approximately 2.5 times larger than the experimental ATE. This difference likely reflects algorithmic improvement through weekly retraining; users crossing the threshold later in the deployment period encountered a more refined recommendation system. However, we cannot rule out compositional differences: users who reached 60 interactions during scaled deployment may exhibit greater preference heterogeneity than those in the RCT, making personalization more effective for these later cohorts. Both mechanisms are consistent with the core finding that personalization benefits persist and strengthen during scaled deployment.[6]

[6]Two factors suggest this comparison may understate the improvement during deployment. First, the

7.4 Effects of Sustained Personalization

The preceding analysis examines treatment effects for users in their first week of receiving personalized recommendations. A distinct question concerns whether continued exposure to personalization generates additional benefits: do users who have received personalized recommendations for multiple weeks exhibit higher engagement than users with shorter exposure to personalization? Understanding these dynamic effects is necessary for assessing the long-run value of personalization systems and for predicting how treatment effects evolve as users accumulate experience with algorithmic curation.

Identification Strategy. To estimate the marginal effect of an additional week of personalization, we construct a comparison that exploits the staggered timing of threshold crossings. Consider users observed at calendar week t. Some users crossed the 60-interaction threshold in week t and have thus begun receiving personalized recommendations; others remained below 60 interactions in week t but crossed in week $t+1$. By comparing these groups' engagement in subsequent weeks, we can estimate the returns to additional personalization exposure.

We impose an additional condition on the comparison group: control users must have been within the bandwidth but below the threshold when the treatment group crossed (week t), and they crossed the threshold in the subsequent week ($t+1$). Specifically, at week t, *Treatment* group $X_{it} \in [60, 60+h)$ (crossed threshold, in bandwidth above) and *Control* group $X_{it} \in [60-h, 60)$ (in bandwidth below threshold, not yet eligible). At week $t+1$, the *Treatment* group already receives personalized recommendations, and the *Control* group $X_{i,t+1} \in [60, 60+h)$ now crosses the threshold and begins receiving personalized recommendations.

Assumption 7.1 (Staggered Threshold Crossing)**.** For users in the bandwidth at week t, the running variable is as-if randomly assigned with respect to future potential outcomes:

$$Y_{it+k}(r) \perp\!\!\!\perp X_{it} \mid X_{it} \in \mathcal{W}_h, \quad \forall r \in \mathcal{R} \;\&\; k \geq 1. \tag{11}$$

This assumption requires that, conditional on having an interaction count near the threshold at week t, the precise count is effectively random with respect to potential outcomes under any ranking in all future weeks $t+k$. This extends the local randomization assumption to

RDD estimation window excludes the heaviest users (those far above the threshold), who exhibit the largest treatment effects in our heterogeneity analysis (Table 7). The RDD therefore estimates effects for marginal users near the eligibility threshold, while the RCT averages over the full eligible population including heavy users. Second, the RCT measures cumulative engagement over an almost two-week period, while RDD estimates reflect weekly engagement.

future time periods, ensuring that users just above and just below the threshold at week t are comparable not only in week t but also in subsequent weeks.

For users who crossed at week t (treatment) and users who crossed at week $t+1$ (control), we compare engagement at horizon $k \geq 1$ weeks after the treatment group's crossing. At calendar week $t+k$, both groups receive rankings based on their accumulated histories $H_{i,t+k}$, but these histories differ. *Treatment* group $H_{i,t+k}$ reflects $k+1$ weeks of personalized exposure (weeks t through $t+k$) and *Control* group $H_{i,t+k}$ reflects k weeks of personalized exposure (weeks $t+1$ through $t+k$) plus editorial exposure in week t.

The policy π^P_{t+k} maps each user's history to a ranking: $r_{i,t+k} = \pi^P_{t+k}(H_{i,t+k})$. Let τ_i denote the week in which user i crossed the eligibility threshold. The estimand is the marginal effect of the $(k+1)$th week of personalization relative to the kth week:

$$\delta_k = \mathbb{E}[Y_{i,t+k}(\pi^P_{t+k}(H_{i,t+k})) \mid \tau_i = t] - \mathbb{E}[Y_{i,t+k}(\pi^P_{t+k}(H_{i,t+k})) \mid \tau_i = t+1]. \tag{12}$$

Under the staggered threshold crossing assumption, this is identified by:

$$\delta_k = \mathbb{E}[Y_{i,t+k} \mid D_{it} = 1, X_{it} \in \mathcal{W}_h] - \mathbb{E}[Y_{i,t+k} \mid D_{it} = 0, X_{it} \in \mathcal{W}_h]. \tag{13}$$

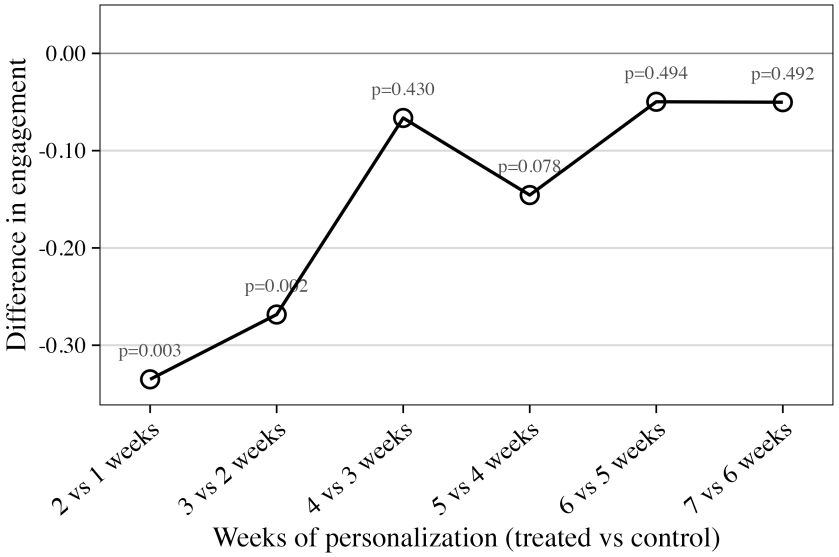
We estimate δ_k using the local randomization RDD framework, comparing mean outcomes for users in the bandwidth at week t who were just above versus just below the threshold.

Results. Figure 5 presents estimated marginal effects $\hat{\delta}_k$ for $k \in \{1, \ldots, 6\}$. We find that the marginal value of additional personalization exposure is negative for the first several weeks.

At horizon $k = 1$, comparing users with two weeks versus one week of personalization, we estimate $\hat{\delta}_1 = -0.35$ ($p = 0.003$). This negative estimate indicates that users in their second week of personalization exhibit lower engagement than users in their first week. Note, that the *Control* group in this comparison has just entered their first week of personalization, and exhibited the average increase in *Total Story Engagement* of approximately 0.4 (based on our RDD estimates). The effect remains negative and statistically significant at $k = 2$ ($\hat{\delta}_2 = -0.27$, $p = 0.002$), comparing three weeks versus two weeks of exposure. By $k = 3$, the marginal effect attenuates toward zero ($\hat{\delta}_3 = -0.05$, $p = 0.430$) and remains statistically insignificant at longer horizons.

These findings suggest that the benefits of personalization are front-loaded. Users experience the largest engagement gains during their initial exposure to algorithmic recommendations, with diminishing marginal returns to continued exposure. This pattern stands in apparent tension with the large positive effects documented in the cross-sectional RDD

Figure 5: Marginal Effects of Additional Personalization Exposure



Note: Estimates of marginal effects of an additional week of personalized recommendations on Total Story Engagement. Each point compares users with $k+1$ weeks of personalization (treated) to users with k weeks of personalization (control) at the same calendar time. P-values are displayed above each point. P-values are computed using randomization inference under Fisher's sharp null hypothesis of no treatment effect.

analysis, which compares users with personalization to users without.

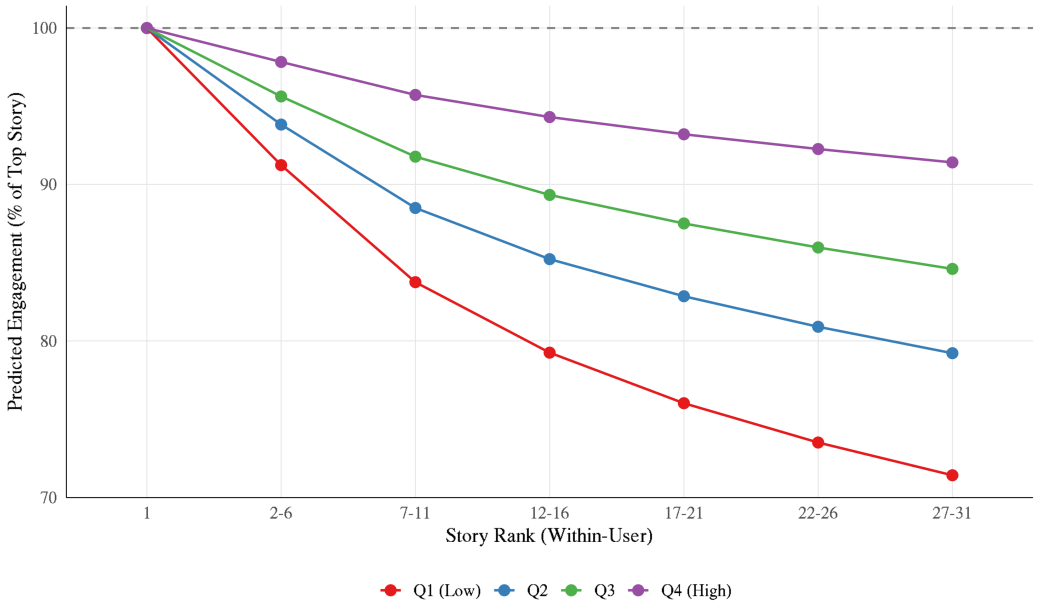
7.4.1 Potential Mechanism: Content Inventory Constraints.

Why do marginal returns to personalization decline so rapidly? We hypothesize that the pattern reflects content inventory constraints inherent to the platform's library structure. The recommendation algorithm ranks all available stories by predicted user-story match quality and presents the highest-ranked options. As users consume recommended content over successive weeks, they deplete the stock of high-quality matches, forcing the algorithm to draw from progressively lower-ranked stories.

Figure 6 provides evidence supporting this mechanism. We plot mean predicted engagement, the algorithm's estimated match quality, as a function of story rank within each user's personalized ordering. Users are organized into quartiles of overall predicted engagement. We trace a drop relative to the top ranked story in percent.

Predicted engagement declines steeply and convexly with rank: stories ranked 20 are predicted to lead to a 20% lower engagement than a top story. Crucially, the decline is highly nonlinear, with the steepest drops occurring among the highest-ranked stories. This convexity implies that the algorithm identifies a small set of well-matched stories for each user, but the quality of subsequent recommendations deteriorates rapidly once this initial

Figure 6: Algorithm-Predicted Match Quality by Story Rank



Note: Mean predicted engagement as a function of story rank within each user's personalized recommendation ordering. Users grouped into quartiles of predicted engagement across all ranks. For each user, stories are ranked by the collaborative filtering algorithm's predicted engagement score; rank 1 denotes the highest-predicted story - 100% predicted engagement.

inventory is exhausted. On average users interact with seven stories per week (see Table 1); thus, by the third week the engagement per story exposed in slate is predicted to drop by 20%.[7]

The content library of approximately 3,200 stories, while substantial for an educational platform of this scale, imposes binding constraints on sustained personalization. Moreover, stories must be grade-appropriate for each user, effectively partitioning the library across the eight grade levels served by the application. Producing new stories requires pedagogical expertise, age-appropriate vocabulary calibration, curriculum alignment, and professional illustration, a fundamentally slow process.

One alternative explanation for these diminishing returns is that the recommendation algorithm becomes less accurate for lower-ranked stories, systematically underpredicting their quality. We investigate this mechanism in Appendix A6 by examining model calibration across the distribution of within-user story ranks. We find that the algorithm is well-calibrated up to rank 30, with calibration errors centered near zero for all rank bins. In particular, we find no evidence that the algorithm underestimates engagement for lower-ranked stories. If anything, predicted and observed engagement tracks closely throughout

[7]We also expect the number of story interactions to drop as users are exposed to stories of lower quality.

the range.[8]

These findings carry important implications for the design and evaluation of personalization systems in moderate-scale environments. First, the front-loaded benefit structure suggests that personalization generates substantial value precisely where it is needed most: during initial user engagement when the risk of churn is highest. The large first-week effects documented in our cross-sectional analysis may be sufficient to meaningfully improve user retention even if marginal returns subsequently decline. Second, the content inventory constraint highlights a fundamental asymmetry between personalization in entertainment versus education. Recommendation algorithms are only as effective as the underlying content library permits. In settings with static or slowly evolving catalogs, the long-run returns to algorithmic curation are bounded by content availability rather than algorithmic sophistication. Third, these results underscore the importance of continuous content investment as a complement to personalization technology. Platforms seeking to sustain engagement gains must replenish their content libraries at rates commensurate with user consumption, a challenge that may require fundamentally different content strategies than those employed by large-scale entertainment platforms.

8 Conclusion

We designed, tested, and scaled a collaborative filtering recommendation system for an educational app for Indian children. Evaluation of the recommendation system is conducted by combining a randomized experiment, followed by a regression discontinuity analysis during five months of deployment.

Our key empirical findings are fourfold. First, experimental treatment effects exceed 60% in the personalized section and 14% app-wide. Second, these effects persist during scaled deployment, with pooled RDD estimates reaching 2.5 times experimental benchmarks as weekly retraining enables algorithmic improvement. Third, effects exhibit substantial temporal variation, with larger gains mid-deployment but volatility during holidays when baseline usage drops. Fourth, personalization benefits are front-loaded due to content inventory constraints—users experience largest gains during initial exposure as they exhaust high-quality matches.

[8]One reason the algorithm's predictions for lower-ranked stories may be biased is that users encounter these stories later in their cumulative app experience, when fatigue or declining novelty may reduce engagement independent of story quality. Additionally, users who persist long enough to reach lower-ranked content may differ systematically from those who churn earlier. The algorithm cannot distinguish genuine variation in match quality from these confounds. Nevertheless, as we show in A6, these effects don't seem to have a substantial effect.

We provide the first large-scale experimental evidence isolating preference-based personalization effects from interface changes in educational settings, complementing prior work on ability-based adaptation. We advance the scaling literature by documenting a rare case where treatment effects amplify rather than attenuate during deployment. We demonstrate how eligibility thresholds can provide perpetual evaluation through regression discontinuity, eliminating the need for costly long-term experimental holdouts.

These findings showcase an asymmetry in the economics of personalization between large and small platforms. Large entertainment platforms, operate with continuously refreshed content catalogs where algorithmic learning and content expansion reinforce each other. In contrast, moderate-scale educational apps face binding content constraints: producing high-quality stories requires pedagogical expertise, curriculum alignment, and professional illustration—which are costly and slow at lower scale. Our evidence demonstrates this asymmetry's consequences: substantial first-week gains show personalization generates value where platforms need it most (initial exposure when churn risk is highest), but front-loaded benefits and subsequent decay reflect content depletion as users exhaust high-quality matches. The implication is clear: algorithmic sophistication and content production capacity are complements, not substitutes. Small platforms considering personalization must ask not whether their algorithms can learn preferences—collaborative filtering works at moderate scale—but whether they can sustain content refreshment rates commensurate with consumption by their most engaged users.

Two design principles emerge for practitioners. First, strategic threshold selection can simultaneously ensure algorithmic performance and enable causal identification—our 60-interaction cutoff provided sufficient training data while creating a sharp discontinuity for RDD analysis. This enables continuous monitoring without costly experimental holdouts. Second, iterative improvement during deployment maintains quality at scale when algorithms learn from accumulated data.

These findings challenge conventional interpretations of experimental evidence on algorithmic interventions. For systems that learn from deployment data, experimental estimates may represent lower rather than upper bounds on scaled impact—reversing standard voltage-drop expectations documented for static interventions. Platforms evaluating personalization might consider RCT results as conservative benchmarks when algorithms will continue learning post-deployment, not as upper bounds requiring discount factors for implementation losses.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014). Finite population causal standard errors. Technical report, National Bureau of Economic Research Cambridge.

Adam, K. D. B. J. et al. (2014). A method for stochastic optimization. *arXiv preprint arXiv:1412.6980 1412*(6).

Al-Ubaydli, O., J. A. List, and D. Suskind (2020). 2017 klein lecture: The science of using science: Toward an understanding of the threats to scalability. *International Economic Review 61*(4), 1387–1409.

Ardalani, N., C.-J. Wu, Z. Chen, B. Bhushanam, and A. Aziz (2022). Understanding scaling laws for recommendation models. *arXiv preprint arXiv:2208.08489*.

Athey, S., A. Kanodia, and E. Palikot (2021). Personalized recommendation system for stones2milestones: Pre-analysis plan. https://github.com/EmilPalikot/S2M_PAP/blob/main/S2M_PAP.pdf. (2021). Archived on OSF in 2025: https://archive.org/details/osf-registrations-ynmer-v1.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Bell, R. M. and Y. Koren (2007). Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter 9*(2), 75–79.

Brandon, A., C. M. Clapp, J. A. List, R. D. Metcalfe, and M. Price (2022). The human perils of scaling smart technologies: Evidence from field experiments. Technical report, National Bureau of Economic Research.

Brynjolfsson, E., A. Collis, D. Deisenroth, H. Garro, D. Kutzman, A. Liaqat, and N. Wernerfelt (2025). The consumer welfare effects of online ads: Evidence from a nine-year experiment. *American Economic Review: Insights 7*(4), 447–462.

Buddelmeyer, H. and E. Skoufias (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA*, Volume 827. World Bank Publications.

Cattaneo, M. D., B. R. Frandsen, and R. Titiunik (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference 3*(1), 1–24.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2018). rdlocrand: Local randomization methods for rd designs. *R package version 0.3. URL: https://CRAN. R-project. org/package= rdlocrand*.

Chaplin, D. D., T. D. Cook, J. Zurovac, J. S. Coopersmith, M. M. Finucane, L. N. Vollmer, and R. E. Morris (2018). The internal and external validity of the regression discontinuity

design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management 37*(2), 403–429.

Dmitriev, P., B. Frasca, S. Gupta, R. Kohavi, and G. Vaz (2016). Pitfalls of long-term online controlled experiments. In *2016 IEEE international conference on big data (big data)*, pp. 1367–1376. IEEE.

Drachsler, H., H. Hummel, B. Berg, J. Eshuis, W. Waterink, R. Nadolski, A. Berlanga, N. Boers, and R. Koper (2008, 01). Effects of the isis recommender system for navigation support in self-organized learning networks. *Journal of Educational Technology and Society 12*.

Escueta, M., A. J. Nickow, P. Oreopoulos, and V. Quan (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature 58*(4), 897–996.

Gomez-Uribe, C. A. and N. Hunt (2015). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS) 6*(4), 1–19.

Green, D. P., T. Y. Leong, H. L. Kern, A. S. Gerber, and C. W. Larimer (2009). Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis 17*(4), 400–417.

Halevy, A., P. Norvig, and F. Pereira (2009). The unreasonable effectiveness of data. *IEEE intelligent systems 24*(2), 8–12.

Holtz, D., B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral (2020). The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 75–76.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Johnson, G. A., R. A. Lewis, and E. I. Nubbemeyer (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research 54*(6), 867–884.

Koren, Y., R. Bell, and C. Volinsky (2009). Matrix factorization techniques for recommender systems. *Computer 42*(8), 30–37.

Kumar, A. and A. Mehra (2024). Improving educational delivery in k-12 schools with personalization: Evidence from a randomized field experiment in india. *Available at SSRN 2756059*.

Linden, G., B. Smith, and J. York (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing 7*(1), 76–80.

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale.*

Crown Currency.

Mercer, J. and J. Meakin (2025). The power of asymmetric experiments @ Meta. Medium.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review 109*(4), 1426–1460.

Oreopoulos, P., C. Gibbs, M. Jensen, and J. Price (2024). Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. Technical report, National Bureau of Economic Research.

Peukert, C., A. Sen, and J. Claussen (2024). The editor and the algorithm: Recommendation technology in online news. *Management science 70*(9), 5816–5831.

Reich, J. and J. A. Ruipérez-Valiente (2019). The mooc pivot. *Science 363*(6423), 130–131.

Rendle, S. (2010). Factorization machines. In *2010 IEEE International conference on data mining*, pp. 995–1000. IEEE.

Rodriguez-Segura, D. (2022). Edtech in developing countries: A review of the evidence. *The World Bank Research Observer 37*(2), 171–203.

Roy, J. (2004). Redistributing educational attainment: Evidence from an unusual policy experiment in india. *Available at SSRN 630141*.

Ruiz-Iniesta, A., L. Melgar, A. Baldominos, and D. Quintana (2018). Improving children's experience on a mobile edtech platform through a recommender system. *Mobile Information Systems 2018*(1), 1374017.

Schaefer, M. and G. Sapi (2023). Complementarities in learning from data: Insights from general search. *Information Economics and Policy 65*, 101063.

Treffers-Daller, J., L. Mukhopadhyay, A. Balasubramanian, V. Tamboli, and I. Tsimpli (2022). How ready are indian primary school children for english medium instruction? an analysis of the relationship between the reading skills of low-ses children, their oral vocabulary and english input in the classroom in government schools in india. *Applied Linguistics 43*(4), 746–775.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

Zhang, B., L. Luo, Y. Chen, J. Nie, X. Liu, D. Guo, Y. Zhao, S. Li, Y. Hao, Y. Yao, et al. (2024). Wukong: Towards a scaling law for large-scale recommendation. *arXiv preprint arXiv:2403.02545*.

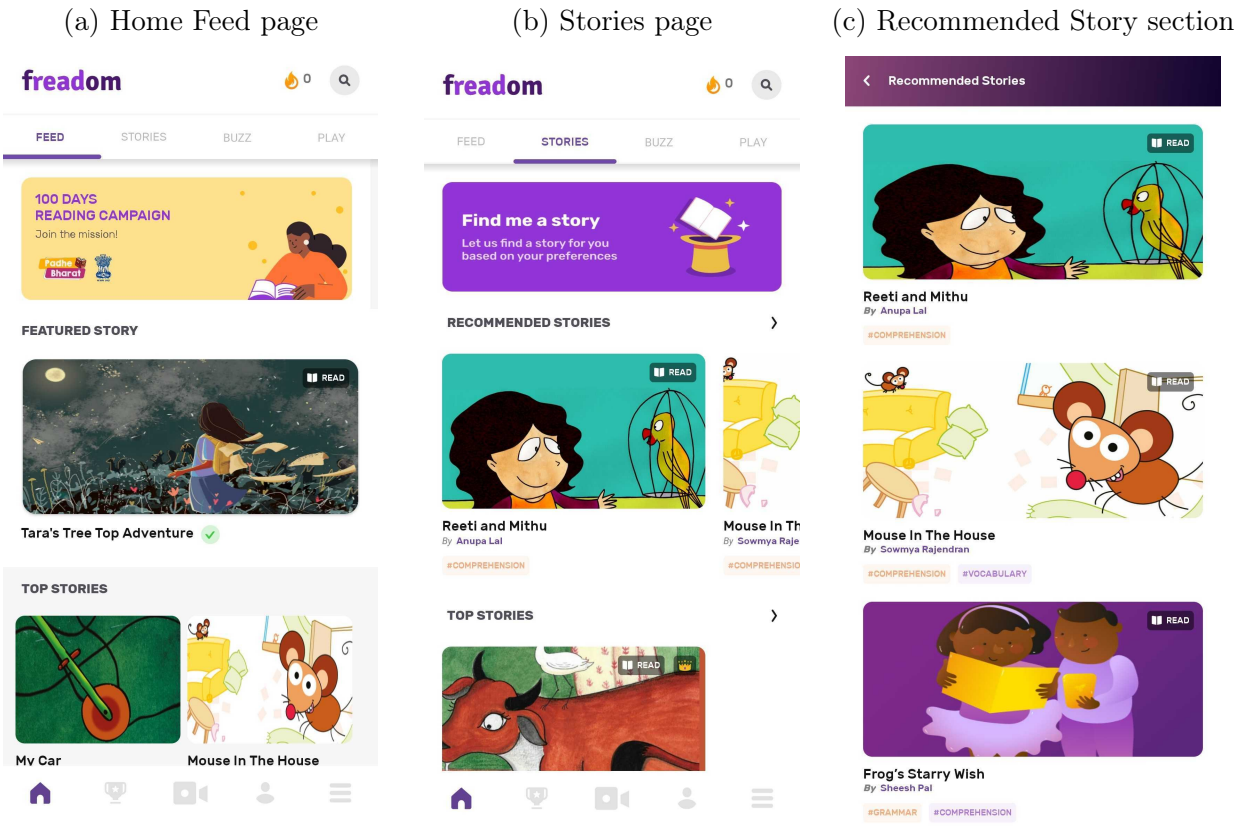
Zielnicki, K., G. Aridor, A. Bibaut, A. Tran, W. Chou, and N. Kallus (2025). The value of personalized recommendations: Evidence from netflix. *arXiv preprint arXiv:2511.07280*.

Appendix

A1 App Interface

Figure A1 shows screenshots from the Freadom interface. The home screen (Panel (a)) displays multiple content sections, each containing a horizontal slate of stories. Users navigate by scrolling vertically through sections and swiping horizontally within a section to browse stories. Panel (b) shows the Stories page, which lists all available sections. Panel (c) shows the *Recommended Story* section—the section where we deployed personalized recommendations—displaying a slate of story cards. When users click on a story, they see a description page with an illustration, title, and brief summary, allowing them to decide whether to start reading or return to browsing.

Figure A1: Freadom interface



Note: Panel (a) shows the home feed where users open the app, displaying multiple content sections. Panel (b) shows the Stories page listing all available sections. Panel (c) shows the Recommended Story *section with a horizontal slate of story cards. Each section displays 15 stories. Users swipe horizontally to browse within a section.*

A2 Collaborative Filtering Details

This appendix provides a detailed description of the collaborative filtering model and estimation procedure.

A2.1 Model Structure

Following the description in Section 5, the collaborative filtering (matrix factorization) model predicts engagement as:

$$\hat{U}_{ij} = \sigma(\theta_i' \lambda_j + \psi_i + \gamma_j + \beta_0), \tag{14}$$

where $\sigma(\cdot)$ is the sigmoid function and β_0 is a global intercept.

The interaction term $\theta_i' \lambda_j$ captures the idiosyncratic quality of user-story matches. Intuitively, each dimension of the latent space corresponds to an unobserved characteristic—such as preference for humor, interest in animals, or affinity for longer narratives. A user's embedding θ_i encodes her preferences along these dimensions; a story's embedding λ_j encodes its attributes. When preferences and attributes align, predicted engagement is high.

A2.2 Estimation

Estimation proceeds as follows:

1. Initialize all embeddings and intercepts with small random values.
2. Iterate over minibatches of observed interactions:
   (a) Sample a batch of observations $\{(i_b, j_b)\}$ from the training data.
   (b) Compute predictions $\hat{U}_{i_b j_b}$ using Equation (14).
   (c) Compute the batch loss: $\sum_b (U_{i_b j_b} - \hat{U}_{i_b j_b})^2$ and gradients with respect to θ_{i_b}, λ_{j_b}, ψ_{i_b}, and γ_{j_b}.
   (d) Update parameters using the Adam optimizer (Adam et al., 2014).
3. Continue until validation loss stops improving.

We tune the number of latent dimensions k and regularization parameters using a randomly held-out validation set. Once optimal hyperparameters are selected, we re-estimate the model on the full training data. In our implementation, we set $k = 20$ latent dimensions.

A2.3 Assumptions and Limitations

We assume a low-rank structure: user-story engagement can be well approximated by the interaction of k-dimensional latent factors. If true preferences are high-dimensional or exhibit complex nonlinearities, this approximation may perform poorly.

Two limitations follow directly. First, the model cannot generate reliable predictions for new users or new stories with no interaction history, known as the "cold-start problem". User i's embedding θ_i is only estimable if i appears in the training data with sufficient interactions; similarly for stories. This motivates our threshold requiring users to have at least 60 prior interactions before receiving personalized recommendations. Second, if the editorial recommendation system that generated the training data systematically excluded certain user-story pairs, the estimated embeddings may not generalize well to those pairs. We mitigate this concern by training on a large historical dataset spanning diverse recommendation regimes and by validating on temporally held-out data.

A3 System Implementation

The *Recommended Story* section displays a slate of 15 stories to each user. We retrain the model and generate predictions weekly. At the start of each week, the algorithm ranks stories by predicted engagement and sends it to the app slate management system.

Within each week, slates update dynamically. When a user starts a story, it is removed and replaced by the next-highest-ranked story. Additionally, if a user is active in the section for two consecutive days but does not complete any of the top three stories, we infer disinterest and remove those stories from the slate. Users who open the app observe top-ranked stories even without clicking, leaving no logged interaction; the staleness rule incorporates this implicit negative feedback.

Algorithm 1 describes how the personalized policy translates interaction histories into user recommendations. The algorithm operates in two steps. Step 1 constitutes the policy π_t^P: it maps the user's history h_{it} to a ranking r_{it} of all stories. Step 2 describes the app's slate management rules, which are equivalent for both editorial and personalized policies. These rules include deduplication (removing previously started stories), slate size constraints (15 visible stories), and the staleness rule (removing top stories if the user repeatedly views but does not engage with them). The editorial policy π_t^E follows the same algorithm structure, but Step 1 simply returns the fixed editorial ranking r_t^E rather than generating user-specific predictions.

Algorithm 1 Personalized Recommendation System

Require: User history h_{it}, trained collaborative filtering model m
Ensure: Daily slates $s_{i,d}$ for days $d \in \{1, \ldots, 7\}$ of week t

STEP 1: Ranking Generation (executed once at start of week)

for each story $j \in \mathcal{J}$ **do**
    Compute predicted engagement: $\hat{U}_{ij} \leftarrow m(h_{it}, j)$ using Equation (14)
end for
Sort stories by decreasing predicted engagement: $\hat{U}_{i,j_{(1)}} \geq \hat{U}_{i,j_{(2)}} \geq \ldots \geq \hat{U}_{i,j_{(J)}}$
Generate ranking: $r_{it} \leftarrow (j_{(1)}, j_{(2)}, \ldots, j_{(J)})$

STEP 2: Slate Management (executed daily within week)

Day 1:
$s_{i,1} \leftarrow$ top 15 stories from r_{it}: $(j_{(1)}, \ldots, j_{(15)})$

Days 2–7:
for each day $d \in \{2, \ldots, 7\}$ **do**
    Available $\leftarrow r_{it} \setminus$ Started$(h_{i,d-1})$ ▷ Remove started stories
    if User active on days $d-1$ and $d-2$ **and** completed none of top 3 from $s_{i,1}$ **then**
        Available $\leftarrow$ Available $\setminus \{j_{(1)}, j_{(2)}, j_{(3)}\}$ ▷ Staleness rule
    end if
    $s_{i,d} \leftarrow$ top 15 from Available
end for
return $\{s_{i,1}, \ldots, s_{i,7}\}$

A4 Covariate balance check

Table A1 presents a comparison of means of user characteristics across treatment and control groups. We find that difference between treatment and control are small and statistically insignificant at the 5% level of significance.

Table A1: Balance of covariates across treatment and control

covariate	Mean (treatment)	sd (treatment)	Mean (control)	sd (control)	p-value (Difference)
past Story Engagement	101.23	139.38	94.67	114.35	0.17
past stories	57.14	89.03	52.75	77.22	0.16
max streak	18.58	85.00	17.79	84.00	0.80
share B2B	0.31	0.46	0.29	0.46	0.46
share B2C	0.41	0.49	0.41	0.49	0.79
share paid	0.25	0.43	0.26	0.44	0.66
share grade 2	0.24	0.43	0.24	0.43	0.82
share grade 3	0.22	0.42	0.22	0.41	0.69

Note: Means of users' characteristics in treatment and control. Last column: p-value from a t-test for the difference in means. Category paid includes users from a paid fLive program and regular paying users; category B2B includes regular B2B customers and club 1br users, a B2B promotion.

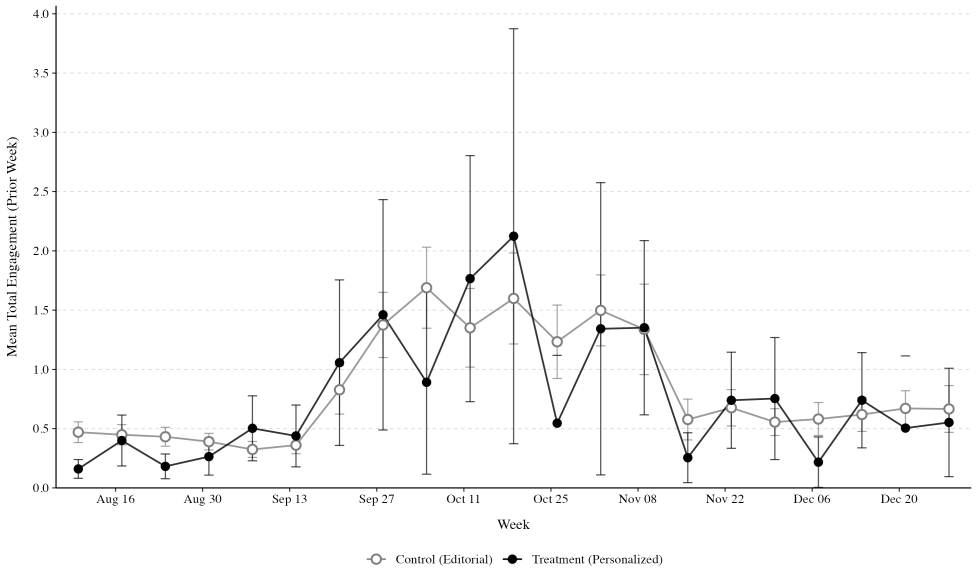
A5 Validation of the Regression Discontinuity Design

In this section, we first compare engagement of treatment and control user in the week before the week in which the treatment group crossed the 60 stories eligibility threshold. Second, we provide a placebo test, by considering a section of the app in which there are no personalized recommendations; thus, there is no change in the way recommendations are generated at the 60 story interactions.

A5.1 Pre-Treatment Balance Check

This analysis examines whether users just above versus just below the 60-interaction threshold exhibit comparable engagement levels prior to treatment assignment. We compare mean Total Story Engagement in the week before threshold crossing between treated users (60–94 interactions) and control users (25–59 interactions) throughout the deployment period. Figure A2 presents weekly comparisons of prior engagement by treatment status from August through December 2021.

Figure A2: Pre-Treatment Balance: Prior Week Engagement by Treatment Status



Note: Mean Total Story Engagement in the week prior to crossing the 60-interaction threshold, by treatment status. Treatment group (black filled circles): users with 60–94 interactions. Control group (gray hollow circles): users with 25–59 interactions. Sample includes all users who crossed the threshold during August–December 2021. Bars represent 95% confidence intervals. Treatment group received personalized recommendations; control group received editorial recommendations.

The treatment and control groups exhibit similar engagement patterns in the week prior to crossing the threshold. The two series track closely across all weeks, with 95% confidence intervals overlapping in nearly every period. Point estimates occasionally diverge. For instance, mid-October shows treatment-group prior engagement of 2.1 versus 1.6 for controls, but these differences are statistically insignificant. The pattern holds across both high-engagement periods (September–October) and low-engagement periods (November–December during holidays). This temporal stability in balance holds despite compositional changes in which users cross the threshold over time.

A5.2 Weekly Placebo Test: Top Story Section

The placebo tests in Section 7.1 validate our RDD design using experimental data, confirming that users just below the 60-interaction threshold provide valid counterfactuals during the RCT period. A natural question is whether this identification assumption continues to hold throughout the scaled deployment period, as user composition at the threshold evolved over time. We monitor this by examining whether discontinuities emerge at the 60-interaction threshold in sections of the app where personalized recommendations were never deployed.

Specifically, we compare weekly treatment effect estimates between the *Recommended Story* section (where personalization was active) and the *Top Story* section (where editorial recommendations continued throughout).

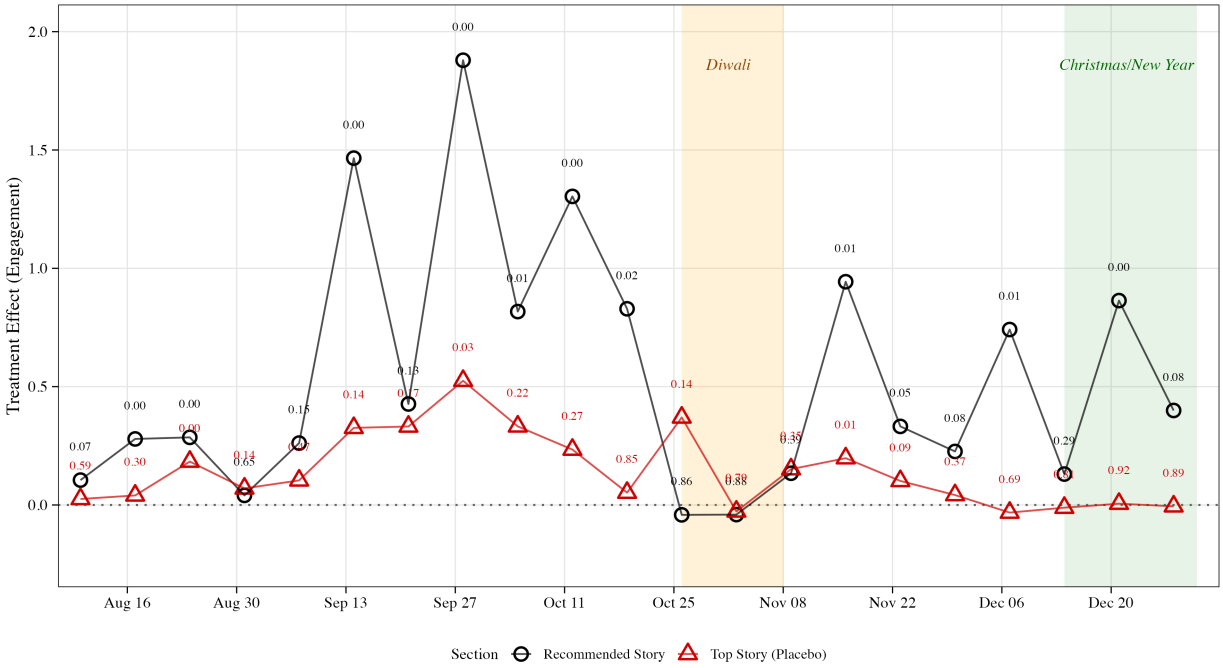
We construct parallel analysis datasets for both sections using identical procedures. For each section, we aggregate user-story interactions to the child-week level, computing *Total Story Engagement* as our outcome measure. We then apply the same local randomization RDD framework described in Section 7, comparing users with 60–94 prior interactions (treated) to users with 25–59 prior interactions (control) within each week. Crucially, the running variable cumulative story interactions (X_{it}) is measured app-wide and is therefore identical across sections. The only difference between analyses is the outcome: engagement in the *Recommended Story* section versus engagement in the *Top Story* section. If the discontinuity at 60 interactions reflects the causal effect of personalized recommendations, we should observe significant treatment effects only in the *Recommended Story* section. If instead the discontinuity captures user maturation, selection effects, or other confounders that operate at the threshold regardless of treatment assignment, we should observe similar patterns in both sections.

Figure A3 presents weekly local average treatment effect estimates for both sections over the scaled deployment period (August–December 2021). Black circles denote estimates for the *Recommended Story* section; red triangles denote estimates for the *Top Story* placebo section. The results reveal a stark contrast between sections. The *Recommended Story* section exhibits large, frequently significant treatment effects, with point estimates ranging from 0.1 to 1.9 and 13 of 20 weeks achieving statistical significance at conventional levels. In contrast, the *Top Story* section shows estimates clustered around zero. The mean treatment effect in the *Top Story* section is 0.04, compared to 0.58 in the *Recommended Story* section.

These findings provide support for the internal validity of our RDD design. The absence of a discontinuity in the *Top Story* section rules out several alternative explanations for our main results. If users who accumulate 60 interactions are fundamentally different, more engaged, more skilled at navigating the app, or more committed to learning, this should manifest in higher engagement across all sections, not only the section receiving personalized recommendations. Similarly, if users strategically adjust their behavior around the threshold, or if unobserved characteristics correlate discontinuously with the running variable, these patterns should affect outcomes in all sections equally. Both sections are observed over the same time period and subject to the same external shocks (Diwali, school holidays, Christmas break), so the differential response across sections cannot be attributed to calendar-time variation.

The placebo test also speaks to potential spillover effects. One might hypothesize that

Figure A3: Weekly Placebo Test: Recommended Story vs. Top Story Section



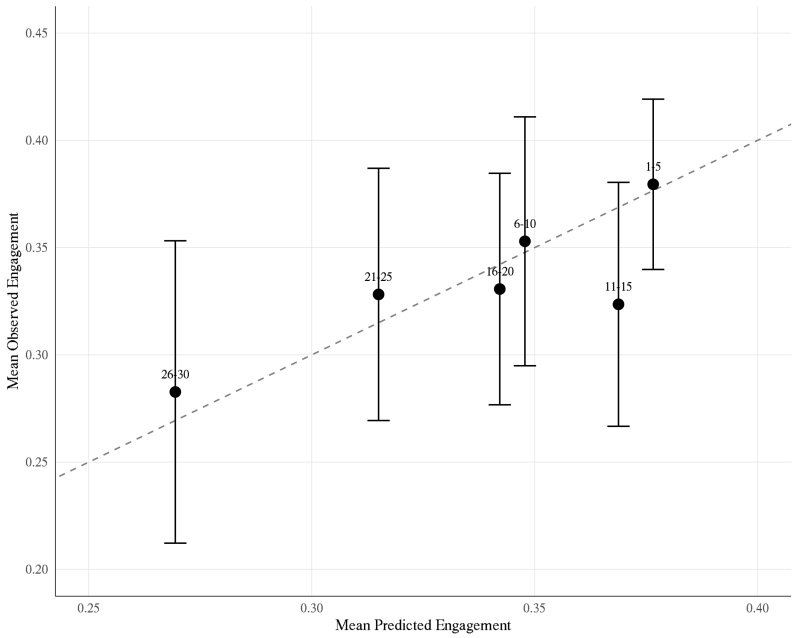
Note: Local average treatment effects of crossing the 60-interaction threshold on Total Story Engagement, estimated separately for each week using local randomization RDD with bandwidth of 35. Black circles: Recommended Story section (personalized recommendations deployed). Red triangles: Top Story section (editorial recommendations maintained; placebo). Orange shading indicates Diwali period; green shading indicates Christmas/New Year period.

users who receive satisfying personalized recommendations in the *Recommended Story* section subsequently consume more content in other sections. If such spillovers were substantial, we would observe positive treatment effects in the *Top Story* section as well. The null results suggest that any cross-section spillovers are at most second-order in magnitude, consistent with our RCT finding that app-wide engagement gains were concentrated in the treated section.

A6 Model Calibration Analysis

A key assumption underlying our analysis of content inventory constraints is that the recommendation algorithm's predicted engagement accurately reflects true engagement differences across story ranks. If the model systematically over- or under-predicts engagement for lower-ranked stories, our conclusions about quality decay would be biased. We assess this by examining model calibration across the distribution of within-user story ranks. We use the RCT period to evaluate the calibration. For each user, we rank all exposed stories by

Figure A4: Calibration plot by within-user story rank

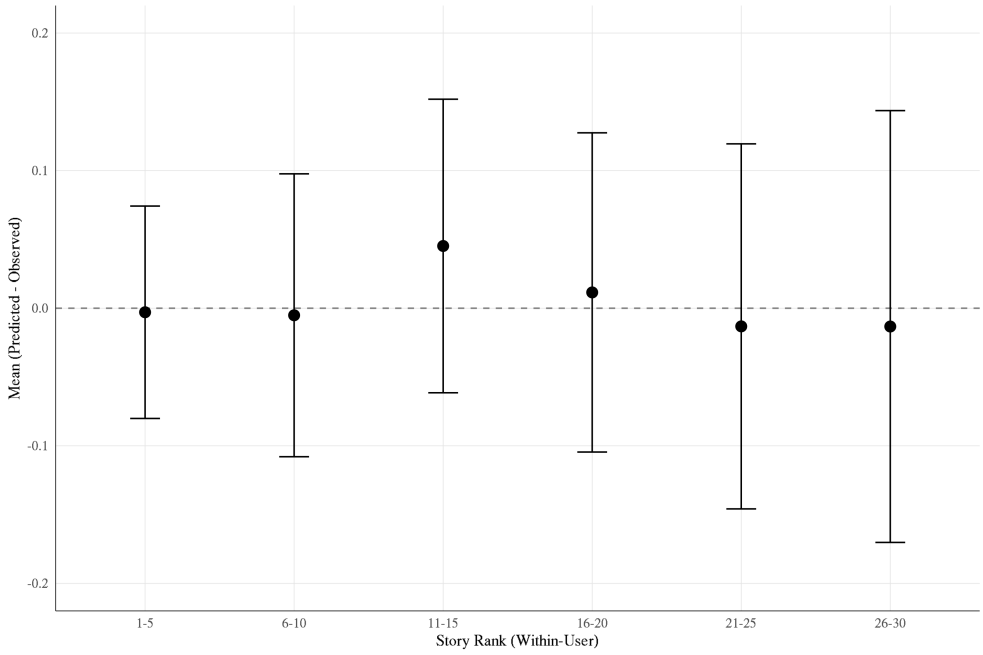


Note: Each point represents the mean predicted and mean observed engagement for stories in that rank bin, aggregated across users. Error bars show 95% confidence intervals. The dashed line at zero indicates perfect calibration.

predicted engagement and group them into bins of five ranks. Within each bin, we compute the mean predicted engagement and mean observed engagement at the user level, then ag-

gregate across users. Figure A4 plots both engagement metrics, for each rank bin, with 95% confidence intervals based on the variance of observed engagement.

Figure A5: Calibration error by within-user story rank.



Note: Each point represents the difference between mean predicted and mean observed engagement for stories in that rank bin, aggregated across users. Error bars show 95% confidence intervals. The dashed line at zero indicates perfect calibration.

The results indicate that the model is well-calibrated across the considered range of story ranks. Calibration errors are centered near 45 degrees line for all rank bins, and the 95% confidence intervals include it throughout. Importantly, we find no evidence of systematic bias at higher ranks: the model does not underpredict engagement for lower-quality stories (ranks 26–30) relative to top-ranked stories (ranks 1–5). Note, that sample changes as we move between ranks. The higher the rank of the story, the fewer users interacted with it, with stories ranked 20 and more, having only interactions by heavy users. One way this can be seen in the plot, is that predicted engagement in stories in ranks 11–15 is higher than of stories in ranks 6–10. This is because users who have interacted with stories at ranks 11–15 have on average higher predicted engagement than users who interacted with stories at ranks 6–10.

Additionally, we plot the differences in predicted and observed engagement. To do that, we first rank all stories within each user by predicted engagement. Then stories are grouped into rank bins of five (1–5, 6–10, etc.), and for each user we compute the mean predicted and mean observed engagement within each bin. Finally, we aggregate across users to obtain the

average calibration error (predicted minus observed) for each rank bin, with standard errors computed from the user-level variation in calibration errors.

We find that the calibration errors are similar across the considered ranked, and in particular we do not find evidence that we are systematically underestimating engagement at lower-ranked stories.

A7 Pre-Analysis Plan

A pre-analysis plan was originally written and posted on GitHub in 2021. The original version can be accessed through Athey et al. (2021), and a permanent archival version has been deposited on OSF (2025). The preregistration outlines the experimental design, primary outcomes, and planned analyses for evaluating the personalized recommendation system.

We note two deviations from the original pre-analysis plan. First, the randomized experiment ran for 12 days rather than the one week as originally proposed. Second, we deployed personalized recommendations only in the *Recommended Story* section rather than in both *Recommended Story* and *Today for You* as initially planned. Technical and engineering limitations on the partner's platform prevented simultaneous deployment across multiple sections. Consequently, we cannot assess whether personalization effectiveness varies by section placement, though our placebo analysis confirms that untreated sections exhibit no discontinuity at the 60-interaction threshold, validating our identification strategy.