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

Keshav Agrawal[*], Susan Athey[*], Ayush Kanodia[*], Shanjukta Nath[†], Emil Palikot[‡]

September 10, 2026

Abstract

We designed and deployed a personalized recommendation algorithm on an app for Indian children learning to read. Children became eligible after accumulating sufficient history. We decompose changes in the algorithm's treatment effect as it scales, using a randomized experiment at launch and, over five months, a rolling regression discontinuity design at the eligibility threshold. The treatment effect on engagement in the first week of scaled deployment doubles relative to the launch experiment, because accumulated data improved subsequent accuracy. The effect falls within an individual as her best matches are exhausted. The share of users receiving best-match content falls over time.

*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, Marshall Van Alstyne, Myongjin Kim, Andrew Ching, Jeffrey Smith and conference participants at DevSouth Conference (Duke University), the Midwest International Economic Development Conference (University of Wisconsin-Madison), the Amazon Decision Sciences Seminar, the BU Platform Strategy Research Symposium and International Industrial Organization Conference for helpful comments and discussions. A pre-analysis plan was prepared before data collection and is available here. An archival version has been deposited in 2025, Athey et al. (2021).

[*]Stanford University; keshavag@stanford.edu, athey@stanford.edu, akanodia@stanford.edu

[†]University of Georgia; shanjukta.nath@uga.edu

[‡]Northeastern University; e.palikot@northeastern.edu, Corresponding Author.

1 Introduction

The treatment effect of a promising pilot often attenuates in scaled deployment because the implementation, population, or environment differs from the original experiment (Banerjee et al., 2017; List, 2024). This "voltage drop" is among the best-documented regularities in policy evaluation (List, 2022; Vivalt, 2020). An increasingly important class of interventions, however, is built to do the opposite. Algorithmic interventions that learn from data generated during deployment improve with scale since every user the algorithm treats makes it better at treating future users (Halevy et al., 2009; Peukert et al., 2024). Yet for these interventions there is little evidence on whether an experimental estimate is a ceiling or a floor for scaled impact. We argue it is neither, as the treatment effect of an algorithmic intervention evolves as a result of opposing forces. More data improve the algorithm's predictions, but every successful match consumes one: the better the algorithm serves a user, the fewer high-quality matches the catalog holds for that user. Furthermore, the share of users at different stages of the content depletion changes over time. In this paper, we decompose and quantify these effects in the context of a specific intervention.

Personalized content recommendation algorithms provide a natural laboratory as both data accumulation and content depletion are visible in the data they generate. They are also among the most valuable algorithmic interventions in the modern economy, and studies on major online platforms show that personalized recommendations themselves raise user engagement (Zielnicki et al., 2026; Gomez-Uribe and Hunt, 2015; Holtz et al., 2020).

We designed and deployed a personalized recommendation system for Freadom, a mobile app on which roughly 6,400 daily active users aged 3–12 in India learn to read in English by engaging with a catalog of about 3,200 illustrated stories.[1] Two deployment features allowed us to track the treatment effect over time. First, in a randomized controlled trial with 7,750 users conducted at the launch of the new recommendation system, we measure

[1] As described below, our collaboration came at early stages of the application's overall scaling, and the parent company now reaches more than a million users as it continues to scale.

the average treatment effect of the personalized recommendation system over the baseline system in which app's editors select the same content for all users. Personalization raises total story engagement by 60% (SE 18%) in the personalized section and 14% (SE 7%) app-wide, showing that the system creates new engagement rather than shifting it from other sections of the app. Second, the deployed algorithm requires 60 past story exposures before a user becomes eligible. This requirement was determined based on empirical analysis of how much individual history was needed to improve upon non-personalized recommendations. Because users cross the threshold at different times, the same threshold enables continuous evaluation using a local regression discontinuity design (Cattaneo et al., 2015, 2018). We use this design to estimate the recommendation system's treatment effect on newly-eligible users over five months of scaled deployment.

Using the local regression discontinuity design (local-RDD), we compare the treatment in the launch period and in the scaled-deployment period. We validate the design's local-randomization assumption and choose the optimal bandwidth around the threshold using the RCT data. We apply the same threshold, estimator, and bandwidth in both periods to users at the eligibility margin of 60 story exposures receiving personalized recommendations for the first time. In the launch period, we estimate a treatment effect of 0.207. Pooling the five months of scaled deployment, we estimate an effect of 0.424 - twice as large ($p < 0.001$). In our setting, the effect of the algorithmic intervention grows from experiment to scaled deployment.

In the first month of scaled deployment, gains are comparable to the launch-period estimate. They rise in subsequent weeks after model retraining, consistent with the algorithm learning from deployment data. But within an individual user, gains decline over time. We measure this path by following each weekly cohort of newly eligible users, together with the comparison group defined in the same week, for six weeks. The treated group consists of users who cross the eligibility threshold in the cohort-formation week, while the comparison group consists of users who remain just below it in that week. Identification beyond the

first week requires the stronger assumption that the exposure count determining treatment in the cohort-formation week is as-if random with respect to potential outcomes over the subsequent follow-up period. We estimate that the gains are largest in the first week of eligibility and decline steadily thereafter, becoming indistinguishable from zero by the sixth week.

The algorithm's design supplies a candidate mechanism for this decline. Once a user has engaged with a story, the app no longer shows it anywhere. In the Recommended Story section this rule has a clear consequence: the section ranks stories by predicted engagement and serves a user's best matches first, so each week it must serve worse matches than the week before. By the model's own prediction (which we validate empirically out of sample), the depletion is steep. Predicted engagement falls 20% within a user's top 20 stories, about three weeks of typical use. Catalog depletion thus adds a new mechanism for voltage drop, one specific to algorithmic interventions: the treatment effect falls not because the implementation, population, or environment changes at scale, but because the intervention uses up the high-quality matches that produce its effect.[2]

We document that the algorithm improves during scaled deployment. First, we explore the hypothesis that the rising gains could reflect a change over time in the characteristics of user cohorts reaching eligibility. We test for compositional change directly and find no evidence of changing user cohort composition. Second, we carry out an analysis that compares engagement for stories that the recommendation system ranked high to engagement for stories the recommendation system ranked low, in settings where (mechanically) the rankings have no impact on user exposure. Because a story can simultaneously appear in several sections of the app, we construct a dataset of stories recommended by the personalized system but exposed to a user elsewhere in the app, for users who never go to the personalized section. This allows us to separate the effect of the story match quality from the effect of additional

[2]Note that outside of our specific application, in digital applications more broadly, content catalogs may be replenished with fresh content, and platforms often optimize this replenishment through direct investments in content acquisition or, for applications that have user-generated content, through platform policies and incentives. Thus, an understanding of the impact of content depletion can guide investments.

exposure due to the inclusion in the personalized section. Outside of the recommended story section, the ranking is not personalized beyond removing stories that had prior engagement. Story selection is not randomized across users or across exposures within a user, but users arriving at different times or going to different sections of the app will see different content. Identification of a measure of prediction accuracy relies on comparing engagement across story exposures for users with similar observed pre-exposure history of browsing, exposure and engagement. We use the Augmented Inverse Propensity Weighted (AIPW) estimator to show that exposure to higher-ranked stories improves engagement conditional on exposure. We then compare consecutive weekly rankings for users who were not active in the first of the two weeks; thus, we can compare two rankings in a setting where historical individual user activity is held constant for both model training and exposure; all that changes is the size of the training data derived from other users. During scaled deployment, a week-to-week model retraining results in more accurate rankings.

The model's improvement does not translate proportionally into gains for the existing stock of eligible users. Following cohorts over tenure, we find that personalization is most valuable when users first become eligible and that this advantage erodes quickly thereafter. The decomposition separates three forces shaping the system's value at scale. First, better model vintages increase the gain for newly eligible users. Second, greater content depth slows the decline as users exhaust high-quality matches; and finally, growth in the inflow of eligible users determines how much of the population remains at short tenure. From September onward, users in their first week account for only 3 to 11 percent of the active eligible population, so most users are already at later tenures where the remaining gain is small. Consequently, realized engagement gains fall well below what the same user base could generate without depletion. By the end of December, depletion reduces the no-depletion gain by about 85 percent. The value of personalization at scale therefore depends jointly on model improvement, content depth, and the flow of new eligible users.

The timing of our research created a rare measurement opportunity. We began collab-

orating with Stones2Milestones (S2M), a for-profit social enterprise behind Freadom, when the app was an early-stage product showing every user the same editor-selected stories. Because we built the platform's first personalized algorithm, our experiment compares personalization with a true no-personalization baseline, which wouldn't be possible on mature platforms, where some variant of personalization is already deployed, and experiments can compare only one algorithm with another. S2M has expanded substantially since our study period in 2021 and has currently acquired 1.13 million users across 8403 schools across all its products in India. For a platform at this stage, our estimates speak directly to unit economics: the engagement gains from personalization raise the value of each user and can fund further user acquisition and complementary investments - where our research highlights the importance of a larger story catalog.

An important limitation of this paper is that we measure engagement, not learning. Related work in EdTech has shown that engagement gains can translate into learning gains in several settings (Oreopoulos et al., 2026; Muralidharan et al., 2019). While we do find that treated users spend more time reading and completing more difficult content, we do not observe learning outcomes such as standardized test scores, so we cannot conclude whether the behavioral changes we estimate in the app usage translated into real learning gains. Instead, our findings can be used to justify the investments required to further improve the application and carry out larger scale evaluations, following the innovation and evaluation journey common to socially impactful digital interventions as they scale (Athey, 2025).

2 Related Literature

This paper is at the intersection of two strands of literature. First, the science of scaling studies how treatment effects change as interventions expand (List, 2022; Al-Ubaydli et al., 2017, 2020; Banerjee et al., 2017; Vivalt, 2020; Muralidharan et al., 2019; Rasul, 2026; List, 2026a). The interventions in this literature are fixed: the program delivered at scale is

the program that was piloted, and effects change because the implementation, population, or environment changes. The second literature evaluates algorithms experimentally (Halevy et al., 2009; Holtz et al., 2020). These papers treat data as an input to a model's performance, but they estimate the effect of a given decision rule at a point in time and rarely follow that effect as the algorithm retrains in deployment. Our contribution is to apply the logic developed by the the science of scaling literature to an algorithmic intervention.

First, we compare the treatment effect of algorithmic intervention at launch to its scaled deployment. Using the same intervention, at the same eligibility margin, and the same estimator we find that the effect doubles. For an intervention that learns from data, the scaled deployment effect might be higher than the experimental estimate, the "high-voltage scaling" of List (2022). Other papers documenting the high-voltage scaling show how organizational improvements can lead to higher treatment effects (Angrist et al., 2023; Banerjee et al., 2017; Muralidharan et al., 2021). We contribute a new mechanism: the intervention itself learns from data its deployment generates.

Second we identify a novel mechanism for the voltage drop relevant for algorithmic interventions. Each week the algorithm selects best matches for each user from a fixed content catalog. Over time the algorithm depletes the library from high quality matches and as a consequence the gains from personalization decline. The mechanisms the scaling literature has documented generally arise from human delivery, such as intervention sites being favorably selected (Allcott, 2015), programs hiring from further down the input supply curve as they expand (Davis et al., 2017), implementation fidelity degrading (Bold et al., 2018), or market equilibrium pushing back as non-treated subjects respond (Crépon et al., 2013). Rasul (2026) catalogs delivery constraints: labor supply, staff retention, and sustainability. Catalog depletion in our setting differs in two ways. First, it operates within each treated user as they receive personalized recommendations over multiple weeks, and second, the intervention's own optimization produces it, so the better the algorithm, the faster it spends the best matches for each user in the catalog. This channel belongs in the "scaling dimen-

sion" of the SANS conditions (List, 2026b); yet it is not captured by the constraints arising in human-delivered programs, and because it is internal to the intervention, it can be engineered in advance rather than discovered at scale, in line with the "Option C thinking" of List (2026a, 2024, 2022).

Our third contribution is to the literature on evaluating adaptive systems in live environments. Long-run holdouts and persistent experiments can be highly informative, but they are often costly to maintain and operationally difficult in deployed products (Brynjolfsson et al., 2025; Johnson et al., 2017). We show how an eligibility threshold can support continuous evaluation during deployment without maintaining a costly permanent holdout group. As successive cohorts cross the threshold, users just above and below it provide a continuous local comparison. We validate this approach using data from an RCT. In doing so, we also contribute to the literature that combines evidence from randomized experiments and regression-discontinuity to validate quasi-experimental designs (Green et al., 2009; Chaplin et al., 2018). Our setting offers an opportunity to study the same intervention first through an RCT and then through an RDD on the same platform. We use the RCT data not only to benchmark the intervention's impact but also to assess user comparability around the eligibility threshold and select the optimal RDD bandwidth, which is a key decision for the researcher in the RDD design.

Finally, this paper contributes to the literature on personalization in educational technology. Prior work has often focused on adaptive learning systems that tailor content to student ability (Muralidharan et al., 2019; Kumar and Mehra, 2024; Oreopoulos et al., 2026). Evaluations of preference-based personalized systems often bundled the change in the algorithm with changes in interface or pedagogy (Drachsler et al., 2009; Ruiz-Iniesta et al., 2018). We show that holding the content catalog and interface fixed, and changing content recommendations based on preferences can raise engagement with educational content.

3 Empirical Setting

3.1 The Freadom Application and User Base

Stones2Milestones (S2M) is an education technology company operating in India. Its flagship product, Freadom, is a mobile and web application for children aged 3-12 learning to read in English. Age-appropriate English reading materials are scarce in Indian homes (Treffers-Daller et al., 2022), and the returns to English proficiency are high (Roy, 2004).

During our study period (July to December 2021), Freadom had approximately 6,400 daily active users. Both the app and the company has grown substantially since. S2M's ecosystem now spans multiple platforms, including Freadom, the CBSE Reading App, CMS Reading App and HELLO; these additional custom platforms were not part of the ecosystem in 2021. In 2021, the ecosystem consisted of one app instance serving 71 schools and around 30,000 users. Today, the same backend supports six app instances serving 8,403 schools and 1.13 million registered students.

Freadom acquires users through two channels: partnerships with schools, in which teachers recommend the app as supplementary learning material, and direct downloads from app stores by parents; all essential features of the app are free. 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 category (e.g., animals, sports, adventure, science), and difficulty level.

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. Before

our intervention, all sections used non-personalized algorithms such as editorial curation, popularity-based ranking, or recency-based selection. We introduce personalized recommendations in the *Recommended Story* section, one of the app's largest sections, accounting for roughly 8 percent of all user-story exposures. Previously, the app's editors refreshed this section weekly, selecting the same stories for every user and removing from each user's slate only the stories that user had already engaged with. Appendix A1 provides screenshots of the interface.

3.2 User Engagement Patterns

When a story is assigned to a slate, users may or may not tap on the story. If a story is assigned and the user decides to tap on it, we say the user was exposed to that story. Once a user is exposed to a story, they progress through several decision stages. Users can skip it, click to view the story's description page, start reading after viewing the description, and decide whether to complete the story after starting. User-story exposures follow a characteristic funnel pattern. Approximately 30% of story exposures result in users clicking to view the description page. Of these, approximately 50% proceed to start reading the story. Of these, 53% complete it. Overall, only about 8% of story exposures result in completion, while most exposures involve exploration and information-gathering as users browse multiple stories before selecting which to read fully.

Our primary outcome is *Story Engagement*, which assigns credit of 0.3 for viewing a story, 0.5 for starting it, and 1.0 for completing it. The partial credit captures reading that a completion count would miss: children often read most of a story without reaching the final page, and they view several stories while deciding what to read. The metric follows S2M's practice: in our discussions, this weighting reflected how the company evaluated engagement, and we use it consistently throughout the analysis. The recommendation engine is trained to optimize it, the pre-analysis plan for the RCT designates it as the primary outcome, and the RDD analysis uses the same outcome. As secondary outcomes, we report total

story completions, total reading time, and number of difficult stories completed. We use the term engagement to refer to engagement conditional on exposure, and when we aggregate engagement across exposures within a time period we refer to various categories of "total engagement."

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 users who launched the app between July 2020 and July 2021. This period precedes our RCT and corresponds to the dataset used to build the recommendation system.

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 exposures per week	7.20	1.50	17.30	21.90
Weeks active on app	3.80	2.00	8.00	5.10
Cumulative story exposures	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 engagements include skipping, viewing, starting, or completing a story.

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 is exposed to only 1.5 stories per week, whereas the 90th-percentile user is exposed to 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 (Palikot et al., 2026). Specifically, an average user is active for 3.8 weeks (a median of 2 weeks). Second, app usage is highly skewed: some users are mostly inactive, while others engage with and complete many stories (the standard deviation of story exposures per week is 22). This heterogeneity in exposure and engagement 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: Research Design Pipeline and Empirical Strategy

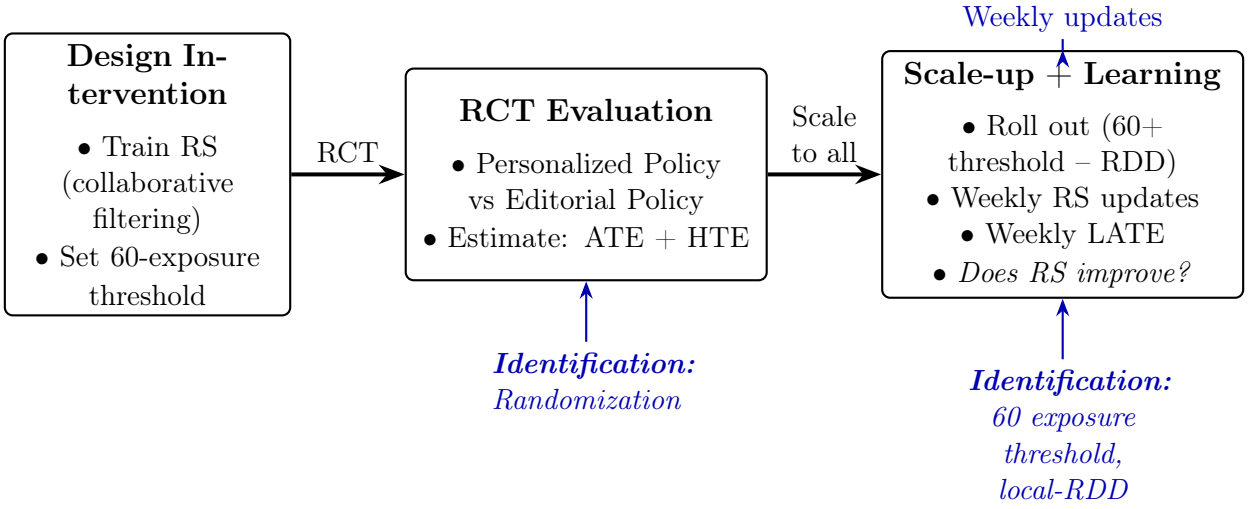


Figure 1 summarizes our research design. We proceed in three stages. First, we develop a collaborative filtering recommendation system and establish a 60-exposure eligibility threshold based on prediction accuracy. Second, we conduct a randomized controlled trial comparing personalized recommendations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

4 Conceptual Framework

4.1 Users, Histories, and Recommendation Policies

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, and let $d \in \{1, \ldots, 7\}$ index days within a week. Let $\mathcal{J} = \{1, \ldots, J\}$ denote the set of available stories. Let $\mathcal{M} = \{1, \ldots, M\}$

denote the set of app sections, with RS $\in \mathcal{M}$ denoting the Recommended Story section. Let $e_{ijdtm} \in \{0, 1\}$ indicate whether user i was exposed to story j in section m on day d of week t.

For user i, let the realized history available at the start of day d in week t collect all prior story exposures and their realized engagement:

$$h_{idt} = \left(\left(e_{ijd't'm} \right)_{j \in \mathcal{J},\, m \in \mathcal{M},\, (t',d') \prec (t,d)},\ \left(u_{ij} \right)_{j \in \mathcal{J}} \right),$$

where $(t', d') \prec (t, d)$ denotes a day preceding day d of week t. Thus, h_{idt} contains the exposure indicators from all app sections through the preceding day, together with realized engagement for previously exposed stories.

The realized total engagement level is $u_{ij} \in \{0, 0.3, 0.5, 1\}$, the realization of the random variable U_{ij} defined in Section 3.2, or NA. The column records u_{ij} if an exposure indicator in the row equals one, and NA if every exposure indicator equals zero. A skipped story has an exposure indicator equal to one and $u_{ij} = 0$, so zero engagement does not indicate the absence of exposure. Because a user can engage with a given story at most once, at most one of the exposure indicators in each row equals one, and u_{ij} does not require a time subscript.[3]

The history h_{idt} is a realization of the random variable H_{idt} and contains the logged exposure and engagement information available to the platform and to us at the start of day d in week t. Let $\mathcal{H}$ denote the set of all exposure histories of this form. The cumulative number of story exposures before week t is $X_{it}(h_{i1t}) = \sum_{t'=1}^{t-1} \sum_{d'=1}^{7} \sum_{j=1}^{J} \sum_{m \in \mathcal{M}} e_{ijd't'm}$. Because $e_{ijd't'm}$ is binary, only observed exposures contribute to this sum.

Outcome Definition. Let $\mathcal{S}_{itm}$ denote the set of stories to which user i is exposed in section m during week t, and let $\mathcal{S}_{it} = \bigcup_{m \in \mathcal{M}} \mathcal{S}_{itm}$ denote the set of stories to which the user is exposed across all app sections during that week.

[3]Since the stories on the app are refreshed daily and not in real time, there are a few cases where a user engaged with the same story more than once within a day. We use the maximum realized engagement in these cases.

We consider two primary weekly engagement outcomes: total engagement in the Recommended Story section, $Y_{it}^{\text{RS}} = \sum_{j \in \mathcal{S}_{it,\text{RS}}} U_{ij}$, and total engagement across all app sections, $Y_{it}^{\text{All}} = \sum_{j \in \mathcal{S}_{it}} U_{ij}$. Additionally, for the ranking analysis in Section A9, we use exposures outside the Recommended Story section. Let $\mathcal{S}_{it}^{-\text{RS}} = \bigcup_{m \in \mathcal{M} \setminus \{\text{RS}\}} \mathcal{S}_{itm}$ denote the corresponding set of stories.

Recommendation Policy. Let $\mathcal{R}$ denote the set of possible ranked lists of stories in the Recommended Story section. For user i on day d of week t, let $r_{idt} = \big(r_{idt}(1), \ldots, r_{idt}(15)\big) \in \mathcal{R}$ denote the daily ranking, where $r_{idt}(\rho)$ is the story placed at rank ρ. A policy is a function $\pi_t : \mathcal{H} \to \mathcal{R}$ that maps history available at the start of the day into a ranking. On the first day of week t, h_{i1t} includes all exposures through the end of week $t-1$. On each subsequent day $d > 1$, h_{idt} also incorporates exposures occurring in all app sections during earlier days of week t. The policy therefore generates the refreshed daily ranking $r_{idt} = \pi_t(h_{idt})$.

We consider two policies: editorial and personalized. Both present stories in ranked order, both omit stories to which the user has already been exposed, and both refresh daily as history h_{idt} updates. They differ in how the ranking is produced. The editorial policy π_t^{Ed} applies a common ordering curated by content specialists, so that every user sees the same ranked list (except the previously omitted stories which are dropped): $r_{idt}^{Ed} = \pi_t^{Ed}(h_{idt})$. The personalized policy π_t^{P} uses collaborative filtering to generate user-specific rankings based on histories, so that $r_{idt}^{P} = \pi_t^{P}(h_{idt})$. The personalized policy is retrained weekly. Thus, the subscript t identifies the policy version deployed in week t, while d identifies the day on which the ranking is generated. We informally describe the date the ranking was generated as the "vintage" of the associated policy version.

Potential Outcomes. Let $\mathbf{R}_{it}$ denote the random weekly ranking sequence for user i in week t, and let $\mathbf{r} = \big(r_1, \ldots, r_7\big) \in \mathcal{R}^7$ denote a possible realization of that sequence. For any fixed ranking sequence $\mathbf{r}$, let $Y_{it}(\mathbf{r})$ denote the total engagement user i would exhibit during week t if presented with $\mathbf{r}$.

Under the editorial policy, the daily ranking is $R_{idt}^{Ed} = \pi_t^{Ed}(H_{idt})$. Under the personalized policy, the daily ranking is $R_{idt}^{P} = \pi_t^{P}(H_{idt})$. These daily rankings form the weekly sequences $\mathbf{R}_{it}^{Ed} = \left(R_{i1t}^{Ed}, \ldots, R_{i7t}^{Ed}\right)$ and $\mathbf{R}_{it}^{P} = \left(R_{i1t}^{P}, \ldots, R_{i7t}^{P}\right)$. The assigned policy determines the random ranking sequence $\mathbf{R}_{it}$. For a realized ranking sequence $\mathbf{r}_{it}$, the observed outcome satisfies $Y_{it} = Y_{it}(\mathbf{r}_{it})$.

4.2 Key Estimands and Identification

Randomized Controlled Trial. During the experimental period $\mathcal{T}_0$, eligible users were randomly assigned to the personalized policy ($D_i = 1$) or the editorial policy ($D_i = 0$). On each day d of week $t \in \mathcal{T}_0$ of the experiment, treatment users received the ranking $R_{idt}^{P} = \pi_t^{P}(H_{idt})$, while control users received the ranking $R_{idt}^{Ed} = \pi_t^{Ed}(H_{idt})$. Let $\mathbf{R}_{i,\mathcal{T}_0}^{P}$ and $\mathbf{R}_{i,\mathcal{T}_0}^{Ed}$ denote the corresponding sequences of daily rankings over the experimental period. Under random assignment, the experimental average treatment effect among eligible users is identified as $\tau^{\text{RCT}} = \mathbb{E}\left[Y_{i,\mathcal{T}_0}(\mathbf{R}_{i,\mathcal{T}_0}^{P}) - Y_{i,\mathcal{T}_0}(\mathbf{R}_{i,\mathcal{T}_0}^{Ed})\right] = \mathbb{E}[Y_{i,\mathcal{T}_0} \mid D_i = 1] - \mathbb{E}[Y_{i,\mathcal{T}_0} \mid D_i = 0]$.

Regression Discontinuity Design. Following the experiment, personalized recommendations became available to all users satisfying the eligibility criterion of 60 past story exposures, i.e., $X_{it} \equiv X_{it}(h_{i1t}) \geq 60$. Users were unaware of this threshold. Treatment assignment is deterministic, $D_{it} = \mathbf{1}\{X_{it} \geq 60\}$. We adopt the local randomization framework provided in Cattaneo et al. (2015, 2018).

The potential outcomes $Y_{it}^{P} \equiv Y_{it}(\mathbf{R}_{it}^{P})$ and $Y_{it}^{Ed} \equiv Y_{it}(\mathbf{R}_{it}^{Ed})$, are the engagement outcomes user i would realize in week t under the personalized and editorial policies, respectively. The observed outcome satisfies $Y_{it} = D_{it}Y_{it}^{P} + (1 - D_{it})Y_{it}^{Ed}$. Eligibility is absorbing: because exposures are cumulative X_{it} never decreases, a user who becomes eligible remains eligible in every later week. Hence $D_{i,t-1} = 0$ implies that user i was not eligible in any week before t, and the pair of conditions $D_{i,t-1} = 0$ and $D_{it} = 1$ identifies users in their first week of eligibility.

Assumption 4.1 (Local Randomization)**.** Within the window $\mathcal{W}_h = [60-h, 60+h)$, and among users not previously eligible, the number of past story exposures is as-if randomly assigned with respect to the potential outcomes: $\left(Y_{it}^{P},\, Y_{it}^{Ed}\right) \perp\!\!\!\perp X_{it} \,\Big|\, D_{i,t-1}=0,\ X_{it}\in\mathcal{W}_h$.

Then, for a given week t, the threshold-local effect is

$$\begin{aligned}\tau_t^{RDD}(h) &= \mathbb{E}\big[\, Y_{it}^{P} - Y_{it}^{Ed} \;\big|\; D_{i,t-1}=0,\ X_{it}\in\mathcal{W}_h\big] \\ &= \mathbb{E}[Y_{it} \mid D_{it}=1,\ D_{i,t-1}=0,\ X_{it}\in\mathcal{W}_h] - \mathbb{E}[Y_{it} \mid D_{it}=0,\ X_{it}\in\mathcal{W}_h]\,. \end{aligned} \tag{1}$$

The restriction $D_{i,t-1}=0$ on the treated side is important for the identification. Without it, the conditioning event $\{D_{it}=1,\ X_{it}\in\mathcal{W}_h\}$ also contains users in their second and later weeks of eligibility whose exposure counts have not yet passed $60+h$. The control-group conditioning set does not include $D_{it-1}=0$, since $D_{it}=0$ implies that these control users were not eligible in the previous week.

In the empirical implementation, we pool all weekly cohort observations across the deployment period and estimate the same threshold-local contrast on the stacked sample. Importantly, this pooled deployment estimand does not follow the same users for five months. Rather, each observation corresponds to a user in the first week after threshold crossing (or, on the control side the first week of falling just short of crossing). Let $\mathcal{T}$ denote the set of deployment weeks. The pooled deployment estimand is the average of the weekly threshold-local effects,

$$\tau^{pool}(h) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \tau_t^{RDD}(h). \tag{2}$$

We adopt the local randomization framework rather than continuity-based RDD for three reasons. First, the 60-exposure threshold is mechanical and unknown to users, making the local randomization assumption plausible. Second, because relatively few users cross the threshold each week, randomization inference provides exact finite-sample p-values (Abadie et al., 2014; Imbens and Rubin, 2015) without the 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.

5 Developing the Recommendation System

We personalize recommendations using collaborative filtering (Koren et al., 2009; Rendle, 2010), the workhorse method underlying many deployed recommendation systems (e.g., Gomez-Uribe and Hunt, 2015). Collaborative filtering predicts a user's engagement with a story from the engagement patterns of similar users and requires only story exposure data. This suits the structure of our data. Freadom tracks user-story engagement precisely, yielding behavioral records on millions of exposures, whereas user characteristics are sparse. We only observe grade level and acquisition channel, but prior reading experience, English proficiency, and ability are not recorded. At the time that the recommendation system was developed, natural language processing technology was not sufficiently advanced, and content-based approaches did not perform well.

5.1 Modeling Development and Design Choices

The model predicts engagement for user-story pairs not yet observed:

$$\mathbb{E}[U_{ij}] = \sigma(\theta_i'\lambda_j + \beta_0 + \psi_i + \gamma_j), \tag{3}$$

where θ_i is a latent preference embedding per user i and λ_j is a latent embedding per story j, and the interaction term $\theta_i'\lambda_j$ 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. To estimate these parameters, we minimize the following objective:

$$\mathcal{L} = \sum_{(i,j)\in\Omega} (U_{ij} - \hat{U}_{ij})^2 + \lambda_\theta \sum_i \|\theta_i\|^2 + \lambda_\gamma \sum_j \|\lambda_j\|^2 + \lambda_\psi \sum_i \psi_i^2 + \lambda_\beta \sum_j \gamma_j^2, \tag{4}$$

where Ω denotes the set of observed user-story exposures in the training data. Unobserved pairs do not enter the loss function. The remaining terms shrink the embeddings and baselines toward zero, mitigating overfitting for users and stories with few exposures. Estimation proceeds via stochastic gradient descent using the Adam optimizer (Kingma and Ba, 2015).

The estimated model predicts engagement for any user-story pair, including pairs never observed in the training data. A prediction for an unseen pair pools information across users. If users with embeddings similar to θ_i engaged heavily with story j, then estimates of $\theta_i'\lambda_j$ will be high. In effect, the model recommends to each user the stories that users with similar revealed preferences enjoyed.

Model Selection. We evaluate model performance using historical data from July 2020 to July 2021. The dataset contains all logged user-story exposures during this period, comprising almost 9 million observations across approximately 170,000 users. We split the data temporally. The exposures from the first eleven months form the training set, while the final month serves as the held-out evaluation set.

We compare three specifications: (i) a constant-only model that predicts the training set mean of outcome variable for all observations, providing a lower bound on performance; (ii) the two way fixed effect (TWFE) model which models the expected user story engagement as a function of only user and story fixed effects (and thus ranks stories in the same way for all users, by aggregate popularity); and (iii) the collaborative filtering approach from Equation (3). Appendix A2 provides details of the model estimation.

The TWFE model has a 22 percent lower mean squared error (MSE) relative to the constant-only baseline, indicating that user and story fixed effects capture substantial variation in engagement (Table A1). 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.

Threshold Design. The performance of collaborative filtering depends on the richness of exposure histories. A user with few exposures gives the model little from which to estimate a preference embedding, and a story with few exposures gives it little from which to estimate a story embedding. We therefore restrict personalized recommendations to users and stories above minimum exposure counts. To set the counts, we train separate models on data restricted to users and stories with more than 20, 60, and 100 prior interactions, and we evaluate every model on the same held-out sample of users and stories with more than 100 interactions. Table 2 reports the mean squared error of each model. Excluding users with 60 or fewer interactions from training lowers the error from 0.097 to 0.093; excluding users with 100 or fewer lowers it no further. The story threshold makes almost no difference.

We set the user threshold at 60 prior story exposures, the point at which the model stops improving. At the time of the experiment, about 15 percent of active users met this criterion; they were the platform's most engaged segment. We use the same threshold for stories. Over 90 percent of the library exceeded the threshold.

Table 2: 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 of predicted Story Engagement. Each cell is a separate model trained on child-story pairs restricted to users (columns) and stories (rows) with more than the indicated number of interactions, across all sections of the app, in the data snapshot of February 22, 2021. All models are evaluated on the same randomly held-out pairs, 15 percent of the data, for users and stories with more than 100 interactions. Standard errors are about 0.0005.

Section Selection. We deploy personalized recommendations in the *Recommended Story* section, one of the app's largest, accounting for roughly 8 percent of all user-story exposures.

All other sections retain their existing algorithms, and the interface is otherwise unchanged: the intervention alters only the ordering of stories within a single section. The unchanged sections allow us to assess whether personalization substitutes engagement away from the rest of the app.[4]

Slate Construction. The *Recommended Story* section contains a slate of 15 stories: those with the highest predicted engagement for the user, excluding stories the user has already completed. The slate updates daily, dropping stories with which the user has already engaged. Appendix A2.4 describes the day-to-day reranking. The algorithm is retrained weekly, and each retraining generates a new ranked list for every user.

6 Randomized Controlled Trial

We evaluate the effect of the personalized recommendation system on user engagement through a randomized experiment.

6.1 Experimental Design

We randomized 7,750 users with at least 60 prior story exposures into treatment and control groups. 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.[5]

During the experimental period, 3,023 users launched the app at least once, of whom

[4]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.

[5]Appendix A3 provides a summary of the pre-analysis plan for this experiment.

525 viewed at least one story in the Recommended Story section.[6] These users generated 77,942 story exposures. Our analysis includes all users who launched the app, regardless of whether they were exposed to stories in 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 A4.1 confirms covariate balance between treatment and control groups.

6.2 Engagement Outcomes

We examine outcomes at two levels: within the Recommended Story and across all sections of the app. Because engagement in one section may substitute for or complement engagement in another, the app-wide outcomes capture the net effect.

Our primary outcome is Total Story Engagement (as defined in Section 4). Our secondary outcome is Total Stories Completed, the count of stories read to completion. We also measure additional learning outcomes: 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 3 presents summary statistics. Panel A reports outcomes for the Recommended Story section (Y_{it}^{RS}); Panel B reports app-wide outcomes (Y_{it}^{All}).

6.3 The Impact of Personalization on Engagement

Table 4 presents our main results: average treatment effects of the personalized recommendation system on four outcomes, estimated via difference-in-means. 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 showed 60 percent higher Total Story Engagement (S.E.

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

Table 3: 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: Summary statistics for outcome variables by treatment status. Panel A presents outcomes measured within the Recommended Story section. Panel B presents outcomes measured across all sections of the app. The sample includes all users who launched the app during the experiment period (N = 3,023). The time period for this summary table corresponds to the RCT period that ran from July 22 to August 4 2021. The unit for reading time is minutes.

Table 4: 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. The time period corresponds to the RCT period that ran from July 22 to August 4 2021. *** p<0.01, ** p<0.05, * p<0.10.*

18%) and completed 77 percent more stories (S.E. 24%) than 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 any substitution away from the non-personalized sections was at most partial. On net, personalization induced users to consume more content overall.

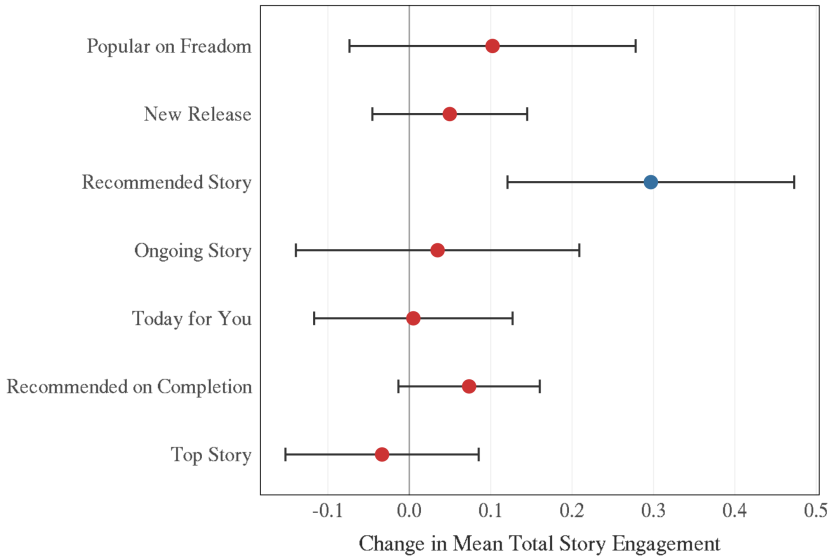
Estimates by section point the same way. Figure 2 plots treatment-control differences in mean Total Story Engagement for the app's major sections. The Recommended Story section, where personalization was implemented, shows a large, statistically significant increase in engagement at the 5% level of significance. No other section shows a statistically significant change in either direction. This pattern is consistent with the app-wide result.

The gains also extend across the distribution of engagement. Personalization lowered the share of users with zero engagement in the Recommended Story section from 87 to 82 percent, and the treated distribution first-order stochastically dominates the control distribution at every positive level of engagement (Appendix Figure A2). A Wilcoxon rank-sum test rejects the hypothesis of no location shift for section-specific engagement ($p < 0.001$) and for app-wide engagement ($p = 0.05$).

In addition, we report the average treatment effects on two learning outcomes, reading time and number of difficult stories completed, in Appendix A4.2. We find that treated users, on average, spend more time reading and complete more difficult stories relative to the control group, alleviating concerns that the recommendation system steered users toward easier, shorter stories.

Heterogeneous Treatment Effects. The experiment enrolled only users with at least 60 prior story exposures, which excluded a large share of the user base. Whether the platform can lower this threshold depends on how the treatment effect varies with the length of a user's history. Collaborative filtering should improve as it accumulates data on a user's preferences, but choosing a threshold requires knowing how steep that gradient is. We therefore estimate treatment effects by prior engagement.

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. The time period for the data used in this figure corresponds to the RCT period.

We split users at the sample median of two pre-experiment measures. *Heavy Engagement* equals one if a user's Total Story Engagement before the experiment exceeds the median, and *Heavy Completion* equals one if the user's number of completed stories before the experiment exceeds the median.

Table 5 reports conditional average treatment effects (CATEs) on Total Story Engagement (Panel A) and Stories Completed (Panel B) in the Recommended Story section. For each indicator we report the control-group mean (Baseline) and the CATE for users with the indicator equal to one and to zero. The last column tests whether the two CATEs differ.

Heavy users gain more than light users. By the Heavy Engagement indicator, the CATE on Total Story Engagement is 0.298 (SE 0.089) for heavy users and 0.052 (SE 0.052) for light users; on Stories Completed it is 0.199 (SE 0.066) and 0.044 (SE 0.034). Both differences are significant at the 5 percent level. The Heavy Completion indicator orders the groups the same way, but its differences are not significant at conventional levels. The pattern is consistent with collaborative filtering performing better when it has more data on a user's

Table 5: Treatment Effects by Prior Engagement

	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)
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)

Note: Conditional average treatment effects (CATEs) of personalized recommendations by pre-experiment engagement. Panel A: Total Story Engagement in the Recommended Story section; Panel B: Stories Completed in the same section. Baseline is the control-group mean. The last column is the difference between the two CATEs. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

preferences. The effect of lowering the eligibility threshold would therefore be smaller than the average effects reported above, and the estimates for light users are too imprecise to say how much smaller.

7 Scaled Deployment and Regression Discontinuity

After the experiment, we scaled the personalized recommendations system to all users with more than 60 story exposures beginning August 5, 2021. This section evaluates whether the treatment effects estimated in the experiment persist during the scaled deployment. To do that, we use a regression discontinuity design that exploits the 60-exposure eligibility threshold. Our main analysis compares users in their first week of personalized recommendations to active users who remain just below the eligibility threshold in that same week. Next we explore how these treatment effects on the newly eligible change across model vintages as training data accumulates. Then, we analyze how treatment effects change as users experience additional weeks with personalization, when the users deplete their best-match content and are exposed to less suitable stories. Finally, we consider the aggregate effect across users over time, including changes over time in the share of eligible users who are newly eligible.

7.1 The Value of Personalization During Scaled Deployment

Table 6 compares the gains from personalization at the eligibility margin during the experiment and during the scaled deployment, estimated with the same local randomization estimator, the same 60-exposure threshold, and the same bandwidths (see Equations 1, 2). The left panel covers the experimental period (July 22 - August 4, 2021). There, the treated group consists of users whose pre-experiment exposure counts fall within the bandwidth above the threshold, $[60, 60 + h)$ and were randomized into the treatment group in the RCT; the control group, labeled *Control RDD*, consists of users outside the experiment whose counts fall just below the threshold, $[60 - h, 60)$. This is the same comparison the deployment-period RDD makes, personalized versus editorial recommendations at the eligibility margin, executed on experimental-period data. The right panel covers the scaled deployment period (August 5 – December 31, 2021) and pools the weekly RDD samples across every week of that period. The outcome in both panels is Total Story Engagement in the Recommended Story section.

The experimental period admits a built-in placebo: the RCT's control group was also above the eligibility threshold, but received editorial recommendations, so the difference in engagement in *Control RCT* and *Control RDD* should not be statistically significant if local randomization holds. The placebo passes for narrow windows: the differences between the two control groups are small and statistically indistinguishable from zero for bandwidths up to 35 (p-values between 0.198 and 0.471), whereas wider windows begin to show statistically significant differences (Appendix A5 for details). We accordingly choose a bandwidth of 35 for our preferred specification, the largest for which the placebo supports local randomization.[7]

[7] In Appendix A6, we provide further robustness of the RDD design. First, treated and control users exhibit statistically indistinguishable *Total Story Engagement* in the week before the treatment group crosses the threshold, week by week from August through December. Second, we carry out the weekly placebo test by re-estimating the RDD with engagement in the *Top Story* section, which never received personalized recommendations, as the outcome: placebo estimates cluster around zero (mean 0.04, against 0.58 in the *Recommended Story* section, where 13 of 20 weeks are significant). The placebo null also indicates that gains from personalization do not spill over into untreated sections, consistent with the experimental finding that app-wide gains concentrate in the treated section.

Table 6: RDD Treatment Effects: Experimental Period versus Scaled Deployment

	Experimental period				Scaled deployment				
BW	N(C)	N(T)	Est.	p-val	N(C)	N(T)	Est.	p-val	Ratio
20	1507	360	0.127	0.034	6885	1940	0.401	< 0.001	3.2
25	2062	417	0.144	0.100	8879	2103	0.407	< 0.001	2.8
30	2820	493	0.202	0.013	10952	2210	0.427	< 0.001	2.1
35	3736	541	0.207	0.006	13062	2295	0.424	< 0.001	2.0
40	4858	587	0.224	0.001	15302	2376	0.430	< 0.001	1.9
45	6408	629	0.253	0.000	17493	2461	0.426	< 0.001	1.7
50	8363	675	0.250	0.000	19487	2551	0.426	< 0.001	1.7
55	11130	707	0.258	0.000	21054	2610	0.428	< 0.001	1.7

Note: Local average treatment effects of personalized recommendations on Total Story Engagement in the Recommended Story section, estimated via local randomization RDD at the 60-exposure eligibility threshold. For each bandwidth h, the estimation window is $[60-h, 60+h)$. In the experimental period, the treated group consists of RCT treatment-arm users with pre-experiment exposure counts in $[60, 60+h)$ and the control group (Control RDD) of non-experimental users with counts in $[60-h, 60)$; the outcome is measured over the experimental window (July 22 to August 4 2021). In the scaled deployment period, the sample pools all user-week observations from August 5 to December 31, 2021. Estimates are differences in means; p-values come from randomization inference under Fisher's sharp null hypothesis of no treatment effect. The final column reports the ratio of the scaled-deployment estimate to the experimental-period estimate.

Table 6 reports three results. First, personalization raised engagement at the margin during the experimental period. The RDD estimates in the left panel range from 0.127 to 0.258, rising as the window widens, and are statistically significant at the 5 percent level at every bandwidth except 25 ($p = 0.100$). At the preferred bandwidth of 35 the estimate is 0.207 ($p = 0.006$).

Second, these estimates line up with the experiment. The experimental ATE of 0.171 (SE 0.052) is closest to the bandwidth-35 estimate. The two estimands differ, however, in several ways. The experimental ATE averages over the full eligible population. In the heterogeneity analysis (Table 5), users with high past engagement gain much more in terms of Total Story engagement than users with low past engagement. The experimental population includes many such high-engagement users far to the right of the 60-exposure threshold. Restricting both experimental arms (treated and control in RCT) to users whose pre-launch exposure counts fall in the RDD window $[60, 95)$ yields an ATE of 0.114 (SE=0.071), below the full-

sample estimate, which aligns with the heterogeneity analysis. This within-window estimate is noisy, because only 476 treated and 485 control users fall in the window.[8]

Third, the gains from personalization are larger during scaled deployment. The right panel pools all users across the scaled period into a single estimation: every user who crosses the threshold between August and December enters one RDD, so cohorts that met the September system and cohorts that met the December system contribute to the same estimate, each in proportion to its size. Pooled estimates range from 0.401 to 0.430, are stable as the window expands, and are statistically significant throughout ($p < 0.001$). At every bandwidth, the deployment-period estimate exceeds the experimental-period RDD estimate. At a bandwidth of 35, the pooled LATE of 0.424 is double the experimental-period LATE of 0.207, and across bandwidths the ratio runs from 1.7 to 3.2. Both panels estimate the same object, the effect of personalization for users at the eligibility margin, with the same estimator and bandwidth.

Our results suggest that the estimate from the experiment conducted prior to launch is a lower bound on the deployed system's treatment effect. We find the effect of personalization roughly doubled between the experiment and deployment. Each week of deployment brings two changes, a model retrained on accumulated data and a new cohort of users crossing the threshold, so the larger effects could reflect an improving algorithm or a changing population of newly eligible users. Section 7.2 traces the week-by-week evolution of the effect and Section A9 provides evidence supporting algorithmic improvement.

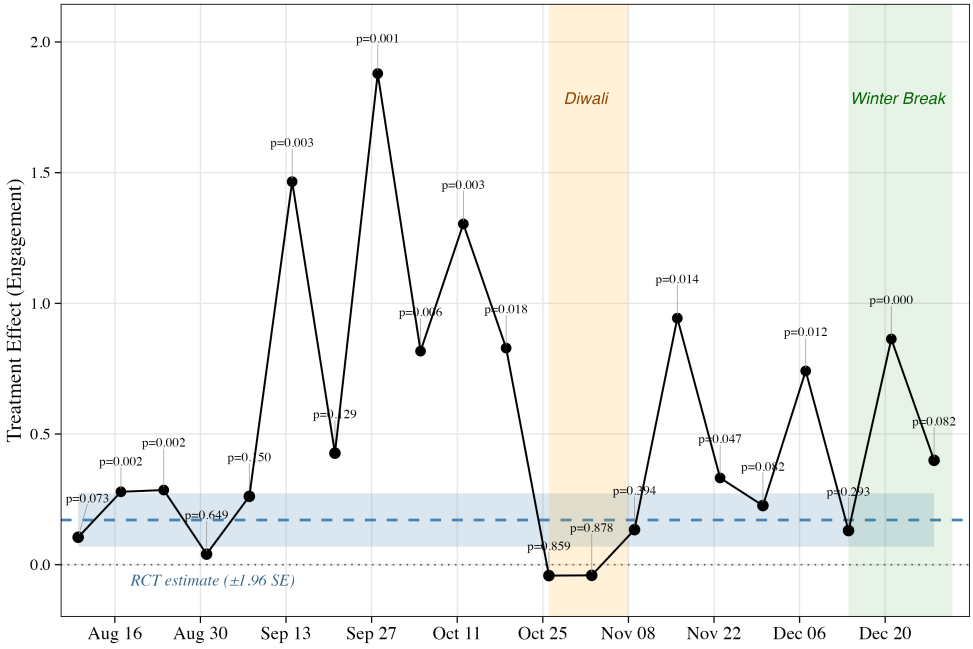
7.2 Weekly RDD Estimates from Scaled Deployment

The pooled estimates in Table 6 aggregate the gains from personalization over five months during which both the algorithm and the platform's context evolved. Figure 3 unpacks this average. Each point is the local randomization estimate for one week, with a bandwidth of 35

[8]The experiment can also be split into its two weeks, each served by a different model vintage, and estimated with the scaled-period discontinuity procedure. Appendix A7 does this. At the bandwidth of 35, the discontinuity in total story engagement is 0.106 in the first week, 0.090 in the second, and 0.098 pooled, each significant at the 1 percent level, and the estimates are positive and significant at every bandwidth.

exposures around the eligibility threshold. It compares two groups of users in the outcome week: those who crossed the 60-exposure threshold in the preceding week and ended it with 60-94 interactions, who receive personalized recommendations for the first time, and those who ended the same week with 25-59 exposures, who continue to receive editorial recommendations. The outcome is mean Total Story Engagement in the Recommended Story section. The dashed line marks the estimate from the RCT of 0.171, with its 95 percent confidence interval shaded in blue.

Figure 3: 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 band (blue). Orange shading indicates the Diwali period; green shading indicates the Winter Break. Values above the dots show the p-values.

The weekly estimates fall into three periods. In the five weeks after the scaled deployment, from August 10 to September 7, the gains from personalization are of the same order as in the experiment. The estimates range from 0.04 to 0.29, and pooled they amount to 0.19 (SE 0.05), inside the 95% confidence interval of the experimental estimate of 0.171 (SE 0.052). The system therefore maintained effectiveness as it scaled to full deployment.

From September 14 to October 19 the gains are far larger. Five of the six weekly estimates lie between 0.8 and 1.9, each with $p < 0.02$, and the six weeks pooled give 1.11 (SE 0.28). From late October, the estimates become volatile. In the two weeks around the time of Diwali (orange shading), they are -0.04, with p-values above 0.85; the number of children launching the app declined by 22 percent below its average in non-festival weeks, and engagement outside the personalized section fell 36 percent, the lowest in the deployment period. After Diwali, the estimates recover but swing from week to week: 0.94, 0.33, and 0.74 in the weeks of November 16, 23, and December 7 (each significant at 5% level), against 0.13 and 0.23 in the weeks between. The winter break (green shading) repeats the Diwali pattern in milder form. We see that the launches decline, reading dips, and the estimates are mixed. Even this volatile period, pooled from October 26 to December 28, yields an estimate of 0.371 ($p < 0.001$), roughly twice the experimental estimate and well above the launch-period estimate of 0.207.

Change in the Composition of Users Crossing the Threshold. A rising weekly estimate could reflect either an improving algorithm or changes in the users reaching the eligibility threshold. Appendix A8 examines the latter. Treated and control users remain closely balanced over the 21 deployment weeks on six characteristics measured before the crossing week, with pooled normalized differences between -0.12 and 0.13 and no systematic trend. The groups separate in activity only in the crossing week, as required for treatment assignment, but the size of this difference is also stable across the three periods of Figure 3: the normalized difference in Total Story Engagement is 0.86, 0.82, and 0.80, with no statistically significant differences across periods. Together with the pre-treatment balance documented in Appendix A6.1, we find no evidence that the rise in weekly treatment effects is driven by changes in the observed composition of users around the threshold.

Story Ranking and Model Improvement. Appendix A9 provides direct evidence that the recommendation model improves during deployment. To separate user-story match qual-

ity from the effect of displaying a story in the Recommended Story section, we study exposures to ranked stories that occur elsewhere in the app among RS-eligible users with no recorded RS exposure, comparing top-half, bottom-half, and off-ranking stories conditional on observed pre-exposure history. AIPW estimates show that top-half stories generate 0.059 more engagement per exposure than off-ranking stories (SE = 0.009), bottom-half stories 0.031 more (SE = 0.008), and top-half stories 0.024 more than bottom-half stories (SE = 0.013). We then compare stories added to versus dropped from consecutive weekly rankings for the same users. Stories added after retraining generate 0.072 more engagement per exposure than dropped stories (SE = 0.032). These results indicate that the rankings contain meaningful user-story match information and that weekly retraining moves them toward stories users engage with more.

7.3 Effects of Sustained Personalization

The weekly estimates of Section 7.2 measure the gain from personalization in a user's first week of eligibility. This section asks whether the gain persist in the subsequent weeks in which the user is receiving personalized recommendations. We follow each weekly cohort of newly eligible users, together with its comparison group (fixed RDD control group defined in the crossing week), over six weeks. Pooled over the deployment, the gain falls from 0.27 units of Total Story Engagement in the first week of eligibility to 0.01 in the sixth, about 0.05 per week. The gain a user receives from the system is largest in the first few weeks of eligibility and gone within six.

Design. The design for the analysis extends the weekly comparison forward in time. For each week g of the deployment, the treated group consists of users in their first week of eligibility, $D_{ig} = 1$, $D_{i,g-1} = 0$, and $X_{ig} \in W_h$, and the comparison group of users with $X_{ig} \in [60 - h, 60)$ who were active in week $g - 1$. These are the two groups of Section 7.2, with the chosen bandwidth $h = 35$. We then observe both groups in weeks $g, g+1, \ldots, g+5$.

Let $c_{it} = t - G_i + 1$ denote the tenure of user i in week t, the number of weeks of eligibility including the current one; for the comparison users of cohort g we use the same index, $c = t - g + 1$, to count the weeks since the cohort was formed. The object of interest is the effect of personalization at tenure c for the cohort formed in week g,

$$\tau_g^{RDD}(c; h) = \mathbb{E}\big[\tilde{Y}_{i,g+c-1}(g) - \tilde{Y}_{i,g+c-1}(\infty) \,\big|\, D_{i,g-1} = 0,\ X_{ig} \in W_h\big], \qquad c = 1, \ldots, 6, \quad (5)$$

where $\tilde{Y}_{it}(g)$ extends the potential outcomes of Section 4 to a user first eligible in week g: it is the engagement user i would realize in week t under personalized recommendations from week g on, and $\tilde{Y}_{it}(\infty)$ the engagement under editorial recommendations in every week. At $c = 1$, $\tau_g^{RDD}(1; h)$ is the weekly estimand $\tau_g^{RDD}(h)$ of Equation (1), and under Assumption 4.1 it is identified from the difference in means between the two groups in week g. For $c \geq 2$ we need to additionally assume that the exposure count in week g must be as-if random with respect to outcomes in the following weeks as well. We therefore consider a stronger version of Assumption 4.1 below.

Assumption 7.1 (Local Randomization over Tenure)**.** Within the window W_h, and among users not previously eligible, the number of past story exposures in week g is as-if randomly assigned with respect to the potential outcomes of week g and the five weeks that follow:

$$\big\{\tilde{Y}_{i,g+c-1}(g),\ \tilde{Y}_{i,g+c-1}(\infty)\big\}_{c=1}^{6} \perp\!\!\!\perp X_{ig} \,\big|\, D_{i,g-1} = 0,\ X_{ig} \in W_h.$$

The case where $c = 1$ is Assumption 4.1. The extension asks that a user's exposure count in week g carry no information about how the user would engage up to five weeks later, under either policy, beyond the restrictions on activity imposed by the window. Under Assumption 7.1, the difference in means between the crossers of cohort g and their comparison users in week $g + c - 1$ identifies $\tau_g^{RDD}(c; h)$ for every c.[9]

[9] Appendix A10 proposes an alternative approach in which we do not adjust for covariates and the comparison group consists only of users who are one week behind the crossers at every tenure. The two approaches lead to the same conclusion.

In Section 7.2, we show that the treatment effects of personalization differ across time and we characterize three periods: *(i.)* the initial period following the scaled deployment (August 10 to September 7), *(ii.)* the high treatment effects period (September 14 to October 19), and *(iii.)* the final period of noisy effects coinciding with many disruptions in the school calendar (October 26 to December 28). In the analysis here we compare the within user treatment effects across these three periods.

Estimating equation. Each cohort and tenure gives one difference in means, in total we have 111 differences. We pool them in a single regression that adjusts for differences between crossers and comparison users in who they are and in how active they were before eligibility. The unit of observation is user, cohort, and tenure, which results in 81,452 total observations, on which we estimate

$$Y_{it}^{RS} = \alpha_{gc} + D_{ig}\left[\beta_{0p(g)} + \beta_{1p(g)}\,(c-1)\right] + \kappa\, D_{ig}\left(\bar{Y}_{gc}^{Ed} - \bar{Y}^{Ed}\right) + \mathbf{Z}_i^{u\prime}\boldsymbol{\gamma} + \tilde{\mathbf{B}}_{ig}^{\prime}\boldsymbol{\eta} + \varepsilon_{it}, \qquad t = g + c - 1. \tag{6}$$

Here Y_{it}^{RS} is Total Story Engagement in the Recommended Story section in week t; D_{ig} equals one for crossers and zero for comparison users; $p(g)$ assigns cohort g to one of the three periods: August 10 to September 7, September 14 to October 19, and October 26 to December 28; and $\bar{Y}_{gc}^{Ed}$ is the mean outcome of the comparison users in cell (g, c), with $\bar{Y}^{Ed}$ its average over cells. As in Section 7.2, a user with no engagement in the section in week t has $Y_{it}^{RS} = 0$, so the estimates are per entering user. Standard errors are clustered by user.

The cell fixed effects α_{gc} absorb the comparison group's mean in every cohort–tenure cell. The treatment coefficients are therefore identified only from differences between crossers and comparison users observed in the same calendar week, the contrast of the weekly design; the calendar, including the rise in engagement in September, Diwali, and the winter break, enters through α_{gc} and not through the treatment terms, except through κ. The intercepts and slopes β_{0p} and β_{1p} let the first-week effect and its change per week of eligibility differ across periods; they are the parameters of interest. The coefficient κ lets the effect scale

with the level of engagement in the comparison group, so that a period observed in high-engagement weeks is evaluated at its own level of untreated engagement. The vector $\mathbf{Z}_i^u$ includes the user-level covariates: age, the number of story exposures before July 21, 2021, and grade level. The vector $\tilde{\mathbf{B}}_{ig}$ includes five measures of behavior before eligibility: log engagement, log completions, and log engagement in the Recommended Story section over the four weeks before week g; log tenure on the app in weeks; and an indicator for no exposure in the Recommended Story section in the two weeks before week g. All five are fixed before personalization begins.[10] A pooled version of Equation (6) replaces $\beta_{0p(g)}$ and $\beta_{1p(g)}$ by a single intercept and slope.

Results. Table 7 reports the estimates. Pooled over the deployment, the effect of personalization is 0.27 (SE =0.04) in the first week and falls by 0.051 (SE=0.013) per week, to 0.01 (SE=0.04) in the sixth. The decline is present in every period. The first-week effect is 0.18 (SE=0.07) for cohorts formed between August 10 and September 7, 0.50 (SE=0.11) for September 14 to October 19, and 0.15 (SE=0.05) from October 26 onwards; the change per week is −0.03 (SE=0.02), −0.09 (SE=0.03), and −0.05 (SE=0.02). By the sixth week the fitted effect is 0.05, 0.03, and −0.08, none distinguishable from zero. The period with the largest first-week effect has the steepest decline, so the three profiles, far apart in the first week, lie within 0.13 of each other by the sixth. The profiles reach the level of the experiment's first-week estimate for the same window and outcome, 0.049 (SE=0.042), by the fourth to sixth week.

Potential mechanism: content inventory constraints. Why does the gain fade within a month? We hypothesize that it reflects the limits of the content library. The algorithm ranks every available story by predicted match quality and shows the highest-ranked ones first, and a user can engage with a story only once. As a user reads through recommended stories, the best matches are used up and the algorithm draws from further down the ranking.

[10] We use the same observed and behavioral characteristics as in Appendix A8.

Table 7: Effect of Personalization by Week of Eligibility

Cohorts formed	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Change per week
Aug 10 – Dec 28 (pooled)	0.27 (0.04)	0.21 (0.03)	0.16 (0.02)	0.11 (0.02)	0.06 (0.03)	0.01 (0.04)	−0.051 (0.013)
Aug 10 – Sep 7	0.18 (0.07)	0.16 (0.05)	0.13 (0.04)	0.10 (0.04)	0.07 (0.06)	0.05 (0.07)	−0.027 (0.023)
Sep 14 – Oct 19	0.50 (0.11)	0.41 (0.08)	0.32 (0.06)	0.22 (0.05)	0.13 (0.05)	0.03 (0.08)	−0.095 (0.031)
Oct 26 – Dec 28	0.15 (0.05)	0.10 (0.04)	0.06 (0.03)	0.01 (0.03)	−0.03 (0.04)	−0.08 (0.05)	−0.046 (0.017)

Note: Fitted values of $\beta_{0p} + \beta_{1p}(c-1) + \kappa\,(\bar{Y}_p^{Ed} - \bar{Y}^{Ed})$ from Equation (6), where c is the week of eligibility and $\bar{Y}_p^{Ed}$ the period's average comparison-group mean; the first row is from the pooled version. Standard errors clustered by user in parentheses. Bandwidth 35; 21 weekly cohorts from August 10 to December 28, 2021, observed through December 28, so that in the last period the later weeks rest on fewer cohorts. The last column is β_{1p}.

Let z_{it} denote the number of stories user i has been exposed to in the Recommended Story section under personalization before week t, the user's position in their ordering, and let $\mu_t(\rho)$ denote expected engagement with the story at rank ρ among the stories not yet shown, under the vintage of week t; let μ_t^{Ed} be expected engagement per exposure under the editorial ordering, which does not depend on the user's position. A user who receives n exposures in the section in week t gains

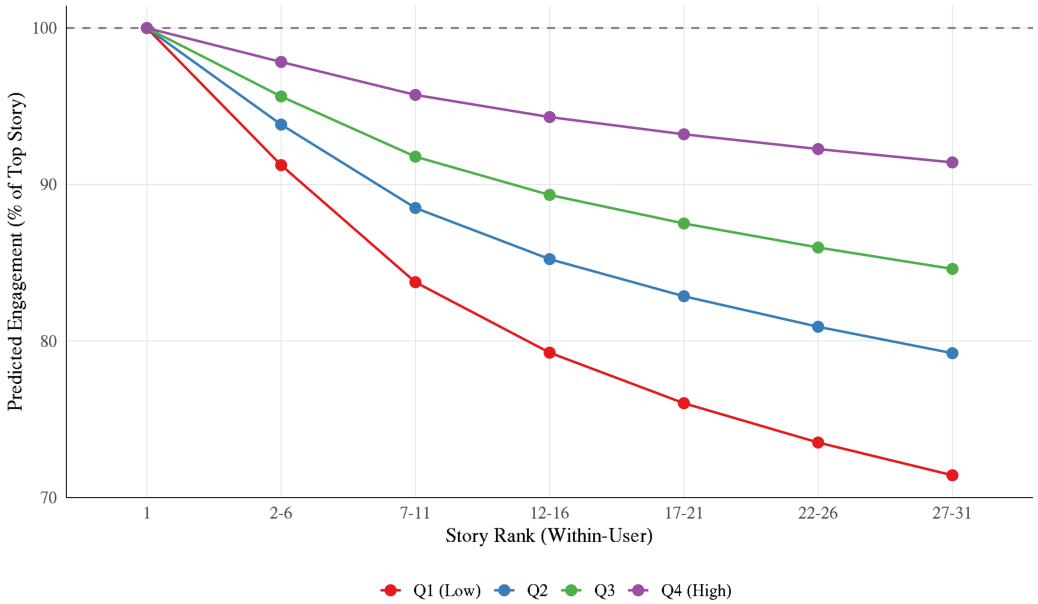
$$\tau(z, n, t) = \sum_{q=1}^{n} \left[\mu_t(z+q) - \mu_t^{Ed}\right]. \tag{7}$$

If $\mu_t(\rho)$ declines in ρ, the gain declines with position even when the model is unchanged.

Figure 4 provides evidence on the shape of $\mu_t(\rho)$. It plots the algorithm's predicted engagement by rank within each user's ordering, relative to the top-ranked story, for users grouped into quartiles of overall predicted engagement. Predicted engagement falls steeply and convexly. A story at rank 20 is predicted to generate 20 percent less engagement than the top story, and most of the decline occurs over the first several ranks. With about seven exposures per week in the section a user reaches rank 20 in the third week of eligibility, by which point the pooled profile has lost about 40 percent of its first-week value.

The library is the likely constraint. It contains about 3,200 stories, partitioned across the eight grade levels the app serves, and producing a story requires pedagogical expertise,

Figure 4: 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 percent predicted engagement.

vocabulary calibrated to age, curriculum alignment, and professional illustration. One alternative explanation is that the model underpredicts the quality of lower-ranked stories. Appendix A11 examines calibration across within-user rank bins and finds the model well calibrated up to rank 30, with calibration errors centered near zero in every bin; predicted and observed engagement track each other throughout the range.[11]

7.4 Decomposition of the Value of Personalization

Sections 7.2 and 7.3 describe two different effects that change over the course of the deployment period. As calendar time passes and the model is retrained, the gain to a user in their first week of eligibility rises, from near the experimental estimate in August to about twice

[11]Predictions for lower-ranked stories could be biased for two reasons: users reach these stories later in their time on the app, when fatigue or fading novelty may lower engagement independently of match quality, and users who persist long enough to reach them may differ from those who stop earlier. Appendix A11 finds no evidence that either matters for the calibration.

that pooled over the deployment. As a user's tenure rises and they work down their ranked list, that gain fades to zero within six weeks. This section combines the two effects. We write the average gain per active eligible user as the sum of three terms: the first-week effect under the current vintage; the loss from depletion as users work down their ranked list; and a reweighting because activity varies with tenure, so the users active in a week have different mix of tenures than the eligible stock as a whole. We estimate each term week by week, then aggregate to the platform total and to the gain that would have been available without depletion.

The Gain per User and Its Elements. We keep the notation of Section 4 and add the following. Let $\mathcal{A}_t$ be the set of users active in week t, with at least one story exposure in any section, and $\mathcal{D}_t = \{i \in \mathcal{A}_t : D_{it} = 1\}$ the set of active eligible users. Tenure $c_{it} = t - G_i + 1$ is as in Section 7.3. The effect of personalization on user i in week t is $\tau_{it} = Y_{it}^P - Y_{it}^{Ed}$, zero for ineligible users, and its average among active eligible users at tenure c is

$$\tau_t(c) = E\big[\tau_{it} \,\big|\, i \in \mathcal{D}_t,\ c_{it} = c\big], \qquad \Delta_t(c) = \tau_t(c+1) - \tau_t(c),$$

with $\Delta_t(c)$ the difference between adjacent tenures in the same week, under the same vintage. Let N_{ct} be the number of active eligible users at tenure c in week t and $\omega_{ct} = N_{ct}/|\mathcal{D}_t|$ their share. Let N_g^{coh} be the number of users whose first week of eligibility is g, and let $\omega_{ct}^0 = N_{t-c+1}^{coh} / \sum_{g \le t} N_g^{coh}$ be the be the share of all users eligible by week t who are at tenure c, including inactive users.[12]

[12]The two sets of shares differ only through activity. Suppose cohorts of 100, 50, and 50 users enter in weeks 1, 2, and 3, and that 0.60 of users are active at tenure one, 0.50 at tenure two, and 0.45 at tenure three. In week 3 the full stock is 200 users and the active stock is 100:

Tenure c	Entry week	N^{coh}	ω_{c3}^0	Active share	N_{c3}	ω_{c3}
1	3	50	0.25	0.60	30	0.30
2	2	50	0.25	0.50	25	0.25
3	1	100	0.50	0.45	45	0.45

The cohort at tenure three is half of the full stock but 45 percent of the active users, because its members are the least likely to open the app. The weight difference $\omega_{ct} - \omega_{ct}^0$ is $+0.05$ at tenure one and -0.05 at tenure three, where the accumulated decline is largest, so term (C) is positive.

Throughout we assume that the number of users who are active and the number of their exposures does not depend on which policy produced the ranking (however, users' level of engagement differs by policy). Under this assumption $\mathcal{A}_t$, $\mathcal{D}_t$, and N_{ct} are the same under both policies; personalization only affects engagement through the selection of stories that occupy a fixed number of exposure slots.[13] The contribution of the system to engagement in week t, relative to a counterfactual in which every user receives the editorial ordering, is

$$\Gamma_t = \sum_{i \in \mathcal{A}_t} \left(Y_{it}^{P} - Y_{it}^{Ed} \right) = \sum_{c \geq 1} N_{ct}\, \tau_t(c), \tag{8}$$

where the second equality uses that the effect is zero for ineligible users and groups the eligible ones by tenure. The average gain per active eligible user is $\bar{\tau}_t^D = \Gamma_t / |\mathcal{D}_t| = \sum_c \omega_{ct}\, \tau_t(c)$. The experiment estimates $\bar{\tau}_t^D$ for a population in which every eligible user is at tenure equal to one; all users are at the beginning of the process of depleting their best-matched content.

Within a week, the effect at any tenure equals the first-week effect under the vintage in force plus the accumulated differences between adjacent tenures under that same vintage,

$$\tau_t(c) = \tau_t(1) + \sum_{c'=1}^{c-1} \Delta_t(c'). \tag{9}$$

The reference point is the current vintage's first-week effect, not the first-week effect that prevailed when the cohort entered. Substituting Equation (9) into the average and using $\sum_c \omega_{ct} = 1$,

$$\bar{\tau}_t^D = \underbrace{\tau_t(1)}_{\text{(A) first-week effect}} + \underbrace{\sum_{c \geq 2} \omega_{ct}^0 \sum_{c'=1}^{c-1} \Delta_t(c')}_{\text{(B) depletion}} + \underbrace{\sum_{c \geq 2} \left(\omega_{ct} - \omega_{ct}^0 \right) \sum_{c'=1}^{c-1} \Delta_t(c')}_{\text{(C) departures}}. \tag{10}$$

[13] The experiment supports the assumption on this margin. The share of users active in each week, and the share who have not permanently left the app, differ between the two arms by at most 0.04 in the full sample and 0.05 within the window, with standard errors of 0.02 to 0.03, in each of the 21 weeks after the launch. Personalization raised engagement, concentrated in the personalized section (Section 6); it did not change the share of users on the app.

Term (A) is the reference level for every eligible user in the week, not only for entrants; its rise over the deployment is the improvement in the model documented in Section 7.2. Term (B) is the accumulated decline for users at tenure two and above, weighted by the tenure distribution that would obtain if every eligible user were active. It is never positive and grows in magnitude as the eligible population ages; the accumulated decline at any tenure is bounded below by $-\tau_t(1)$, since the ranking curve flattens toward zero gain. Term (C) corrects the weights from the full-activity shares ω^0_{ct} to the shares ω_{ct} of users actually active. If users at high tenure are more likely to be inactive in the week, the term is positive. We later show that those absent are disproportionately those for whom personalization had the least remaining value, so the average among those present is higher than it would be with everyone present. The term describes who is in the average. Under our assumption of fixed activity, the same users would have been inactive under the editorial ordering.

7.4.1 Estimating the Elements.

We estimate $\tau_t(1)$ on the cohort-by-tenure sample of Section 7.3, with the treatment effect written as a calendar-week term plus a tenure term,

$$Y^{RS}_{it} = \alpha_{gc} + D_{ig}\big[\lambda_t + \theta\,(c-1)\big] + \kappa\, D_{ig}\big(\bar{Y}^{Ed}_{gc} - \bar{Y}^{Ed}\big) + \mathbf{X}'_i\boldsymbol{\gamma} + \mathbf{B}'_{ig}\boldsymbol{\eta} + \varepsilon_{it}, \qquad t = g + c - 1, \quad (11)$$

where everything other than the treatment terms is as in Equation (6). The two sets of treatment terms are separately identified because each calendar week contains cohorts at every tenure from one to six: the cohort formed in week t at tenure one, the cohort formed in $t-1$ at tenure two, and so on. The coefficient $\hat{\lambda}_t$ is the first-week effect in week t adjusted for the covariates. It tracks the weekly series of Figure 3 with a smaller amplitude: 0.03 (SE=0.09) on August 10, 0.25 (SE=0.08) on September 7, between 0.48 and 0.62 from September 14 to October 19, 0.43 on October 26, and between 0.09 and 0.27 from November on. For the experiment period we use the first week's experimental estimate, 0.049.

Table 8: Elements of the Tenure Profile

Tenure c	1	2	3	4	5	6
Retained fraction $r(c)$ per entering user	1.00	0.81	0.62	0.43	0.24	0.05
Share of crossers active at tenure c	0.65	0.60	0.54	0.52	0.48	0.47
Retained fraction $r(c)$ per active user	1.00	0.88	0.74	0.54	0.32	0.06

Note: First row: the fitted pooled profile of Table 7 divided by its first-week value. Second row: the share with at least one story exposure in the week, at each tenure, pooled over cohorts and weeks. Third row: the first row times the first-week activity rate divided by the activity rate at tenure c, the profile used in Equation (12).

To estimate the contrasts across weeks, we take the shape of the profile from the pooled specification of Section 7.3 and impose that the effect at tenure c is a fixed fraction of the first-week effect under the contemporaneous vintage,

$$\tau_t(c) = r(c)\,\tau_t(1), \qquad r(c) = \frac{\hat{\beta}_0 + \hat{\beta}_1\,(c-1)}{\hat{\beta}_0}, \qquad \text{so that} \qquad \sum_{c'=1}^{c-1} \Delta_t(c') = \tau_t(1)\,\big[r(c) - 1\big], \tag{12}$$

with $\hat{\beta}_0$ and $\hat{\beta}_1$ the pooled intercept and slope. The fractions are 1, 0.81, 0.62, 0.43, 0.24, and 0.05 for tenures one to six (Table 8).[14]

The decomposition week by week. Figure 5 plots the three terms of Equation (10) for every week from the launch of the experiment to December 28. The blue bar is term (A), the first-week effect under the vintage of the week. The orange bar is term (B), the loss from depletion: the accumulated decline at each tenure, weighted by the shares that would prevail had nobody stopped using the app. The green bar is term (C), the correction for

[14]The profile and the weekly effects $\hat{\lambda}_t$ average over every user in a cohort, counting a zero for each week in which a user does not open the app. The effect $\tau_t(c)$ and the count N_{ct} refer to users who are active in the week. Under the assumption that activity does not depend on the policy, the effect per entering user at tenure c equals the share of the cohort active at that tenure times the effect per active user. That share falls from 0.65 in the first week of eligibility to 0.47 in the sixth (Table 8), so part of the decline in the profile reflects users who stop opening the app rather than a smaller effect on those who continue. Dividing by the activity rates removes this part. For example, the first-week effect becomes $\hat{\lambda}_t/0.65$. The retained fraction at tenure c becomes $r(c)$ times the ratio of the first-week activity rate to the rate at tenure c: 1, 0.88, 0.74, 0.54, 0.32, and 0.06, held at the tenure-six value beyond. The experiment's two cohorts use their own activity rates, 0.60 in the first week and 0.41 in the sixth. All figures and tables below use these converted values. Without the conversion, depletion is overstated: the realized share of the no-depletion gain over the deployment is 0.32 rather than 0.36, and 0.12 rather than 0.15 in the last week of December.

departures, small and positive throughout. The black line is their sum, the gain per active eligible user $\bar{\tau}_t^D$.

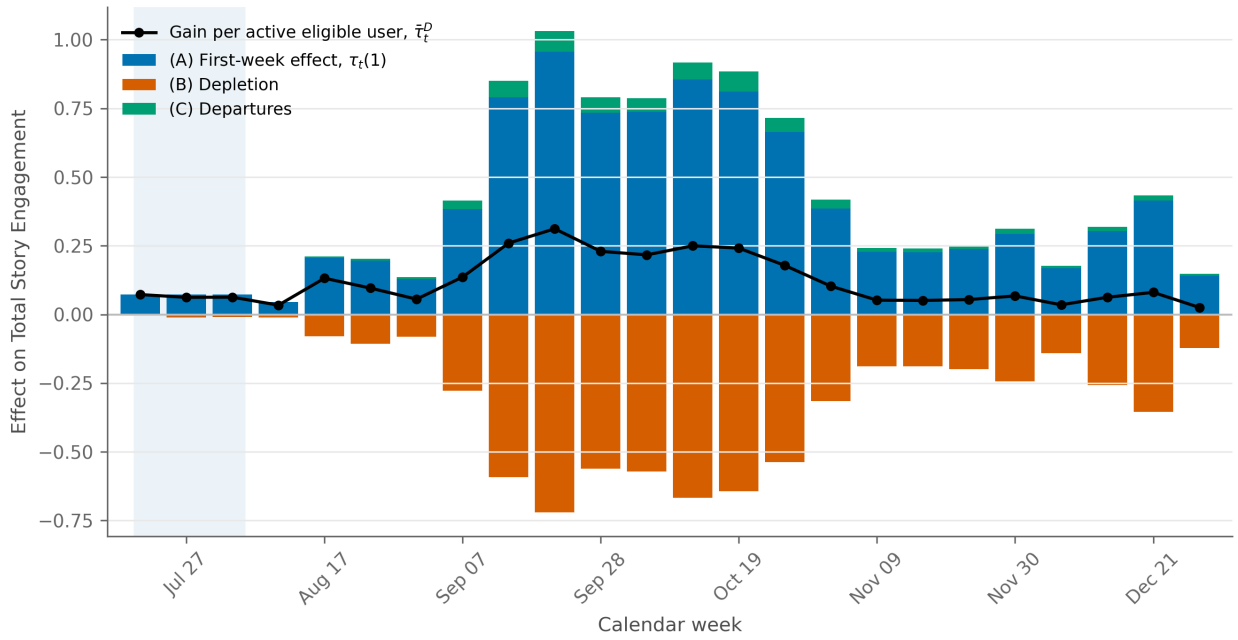
In the first week of the experiment every eligible user is at tenure one, so terms (B) and (C) are zero and the gain per active user equals the per week RCT estimate. From mid-August the orange bar grows as the experimental cohorts age: by September 7 depletion removes 0.28 of a first-week effect of 0.38. In the surge of September and October the first-week effect rises to between 0.66 and 0.96, but the depletion term rises with it, to between -0.54 and -0.72, because by then most active eligible users are at tenures where the retained fraction is small. The gain per active user peaks at 0.31 on September 21 and stays between 0.22 and 0.26 through mid-October. From November the first-week effect falls to between 0.14 and 0.41 and the gain per user to between 0.03 and 0.10. The retained share, $\bar{\tau}_t^D/\tau_t(1)$, falls from 0.75 on August 10 to 0.36 on September 7, about 0.3 in October, 0.22 to 0.27 in November, and 0.18 to 0.21 in December. Term (C) adds at most 0.08, on September 21, and 0.01 to 0.02 in December: users absent in a week are disproportionately at high tenure, so the average among those present is slightly higher than it would be with everyone present.

The figure puts the two findings of Sections 7.2 and 7.3 into one accounting. The rising first-week series is the level of the ranking curve across vintages; the fading profile is the movement of a cohort along it. Under the proportional profile,an improvement in the model raises the gain to entrants and the loss to incumbents in the same proportion, so the retained share is a function of the tenure composition of the active eligible population alone.

7.4.2 Platform-Level Aggregation

Tenure composition. Figure 6 shows the shares ω_{ct}, the composition of active eligible users by week of eligibility, for every week from the launch of the experiment. Darker shades mark longer tenure. During the experiment every active eligible user is in the first week by construction. On August 10, the first week after scale-up, 14 percent of active eligible users are at tenure one, 41 percent at tenure two (the control arm, personalized since August 5),

Figure 5: Decomposition of the Gain per Active Eligible User

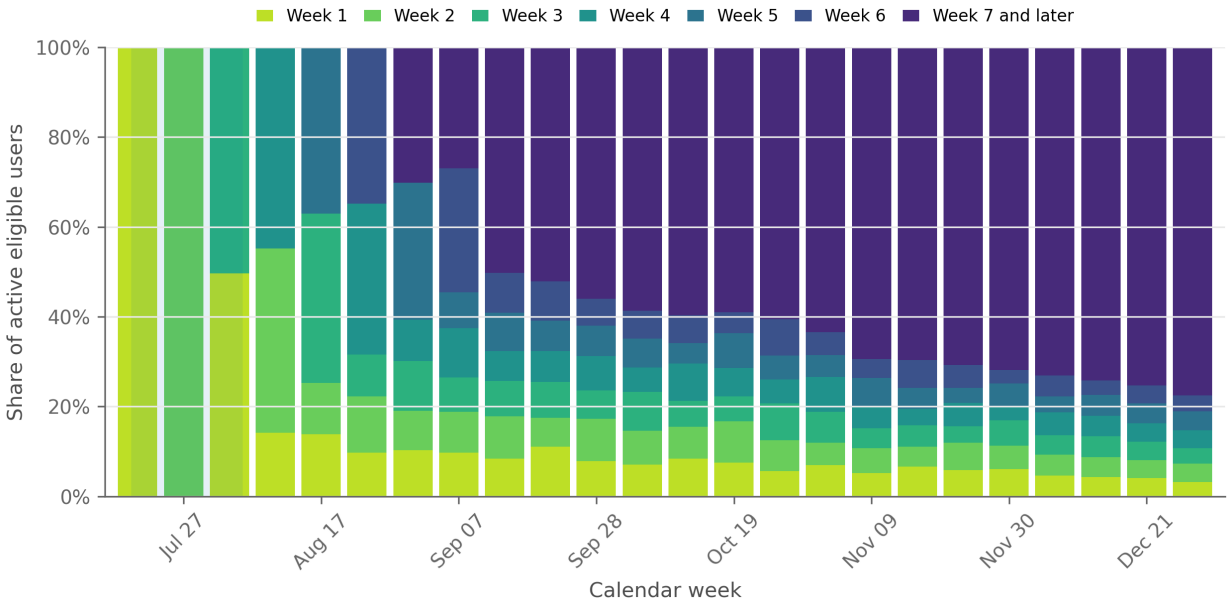


Note: The three terms of Equation (10) by calendar week, in units of Total Story Engagement in the Recommended Story section. Term (A) is $\hat{\lambda}_t$ from Equation (11) from August 10 and the experiment's estimate for users with 60–94 past exposures, 0.049, in the three weeks before, each divided by the first-week activity rate (0.65 for crossers, 0.68 for the personalized arm) to give the effect per active user; terms (B) and (C) use the per-active-user profile of Table 8 and the tenure counts of the threshold-local population, the experiment's arms included. The black line is the gain per active eligible user, the sum of the three terms. Bars stack: term (A) rises from zero, term (C) sits on top of it, and term (B) hangs below zero. Shading marks the experimental period, July 22 to August 4.

and 45 percent at tenure four (the personalized arm). From then on the stock ages. By September 14 half of active eligible users are at tenure seven or more; by October 5, 59 percent; by December 28, 78 percent, at which point the retained fraction is 0.06. The share at tenure one, the users the weekly estimates of Section 7.2 describe, is 8 to 11 percent in September and 3 to 5 percent in December. Within two months of scale-up, most active eligible users are at tenures at which the gain from personalization is close to zero.

Total weekly gain. Figure 7 aggregates the total weekly gain. It plots three series. First, the realized gain, the black line, is the sum over cohorts of the number of active users of the cohort in week t, times the cohort's treatment effect. Second, the no-depletion gain, the blue dashed line, keeps the same active users but assume the treatment effect are the

Figure 6: Composition of Active Eligible Users by Week of Eligibility



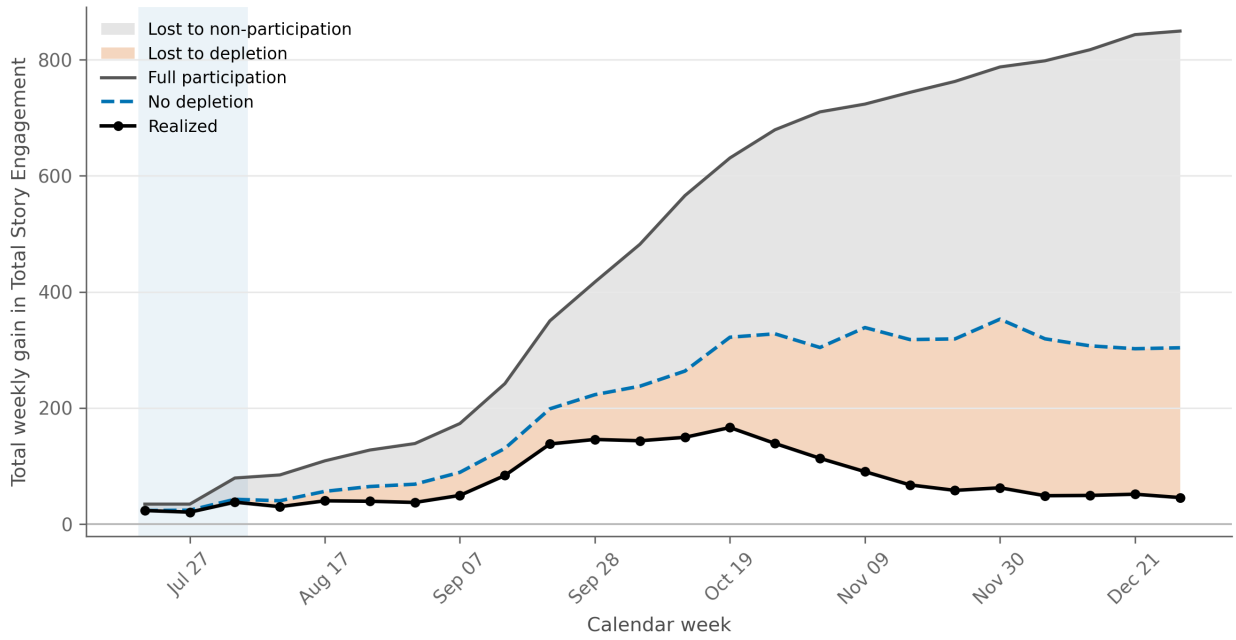
Note: Shares of active eligible users by tenure, ω_{ct}, in each calendar week from July 20 to December 28, 2021, for the threshold-local population: users with 60–94 past exposures when they gained access, including the experiment's two arms, personalized from July 22 and August 5. A user is active in a week with at least one story exposure in any section and is counted once, at one tenure.

users' first week treatment effect. The orange band between the two is the loss to depletion, with participation held at its observed level. Third, the full-participation gain, the grey line, counts every user who has entered by week t, active or not, at their own first-week effect; the grey band between it and the no-depletion gain is the loss to non-participation, weeks in which users who have entered do not open the app. [15]

The figure shows that the potential gain from personalized recommendations (with an unlimited content library) is about three times larger than the actual gain, even with participation held fixed. The no-depletion series keeps rising as data accumulates, while the realized series falls. The gap between the two is the depletion band, and it widens as a larger share of users are at tenures where the gain is near zero. In the week of the peak, depletion

[15]The platform series value a cohort at the vintage under which it entered, whereas Equation (10) values every user at the current vintage. We use the entry vintage here because it holds each cohort at the gain it was measured to receive, so that the depletion band is the loss those same users incurred as they moved down their orderings.

Figure 7: Total Weekly Gain from Personalization: Realized and Potential



Note: Total weekly gain in Total Story Engagement in the Recommended Story section by calendar week. Each cohort is valued at the first-week effect per active user of its own entry week: $\hat{\lambda}_g$ from Equation (11) divided by the crossers' first-week activity rate. Realized (black): active users of each cohort times that effect times the per-active-user retained fraction of Table 8. No depletion (dashed): the same active users at the first-week effect. Full participation (grey line): every user who has entered by the week, at the first-week effect. The orange band is the loss to depletion and the grey band the loss to non-participation. Shading marks the experimental period.

removes 48 percent of the no-depletion gain; by the end of December it removes 85 percent. Under full participation the potential gain rises even more.

Recently eligible users contribute most of the gain from personalization. Because the gain at tenure six and beyond is 6 percent of the first-week effect, the total weekly gain in Equation (8) is close to the inflow of newly eligible users times the first-week effect, plus the contributions of users in their second to fifth weeks. In the long run, the value of the system is proportional to the flow of new eligible users, not to the stock of users receiving personalized recommendations. Equation (8) then identifies three margins on which a platform can act: the model, which raises $\tau_t(1)$; the content library, which flattens $\mu_t(\rho)$ in ρ and extends the number of weeks over which a user gains; and growth in the eligible user base, which raises

the inflow. The improvement in the model over the deployment is substantial, and it is measured on first-week users, where the tenure composition cannot move it. But it raised the gain to entrants, who are only 3 to 11 percent of the active eligible stock from September on, and it did not extend the number of weeks over which a user gains: the profile is steepest in the period with the largest first-week effect (Table 7). In this setting the constraint on sustained gains is the content library rather than the model.

There are two main caveats to consider when interpreting these results. First, the calculation holds activity fixed. If personalization raises the number of sessions or exposures early in tenure, or keeps users on the app who would otherwise have left, the realized series omits that gain; the experiment's evidence that the two arms have the same activity and exit over 21 weeks suggests the omission is small, but the experiment's power to detect small differences in exit is limited. Second, the estimates are built from noisy inputs, both the weekly first-week effects and the rate of decay over time.

8 Conclusion

We designed, tested, and scaled a personalized recommendation system on an educational app for Indian children. We evaluate the system first in a randomized experiment over a short interval and then with a local regression discontinuity design over five months of deployment.

We have four main findings. First, RCT evaluation shows the recommendation system increases engagement by 60% in the personalized section and 14% app-wide. Second, the effects increase during scaled deployment. Pooled RDD estimates for first week gains from personalization reach twice the experimental benchmark, and weekly retraining improves the algorithm. Third, the first week of exposure to personalization exhibits substantial temporal variation, with larger gains mid-deployment and volatility during holidays, when baseline usage drops. Fourth, the within-user benefits of personalization are front-loaded. Users

experience the largest gains during initial exposure; the gains fade high-quality matches are depleted. Our decomposition shows that as the eligible user base ages, content depletion offsets much of the gain from model improvement, so model quality, content depth, and the inflow of newly eligible users together determine the gains from personalization at scale.

Together, these imply that algorithmic sophistication and content production capacity are complements, not substitutes. Gains from personalization last only as long as the catalog offers the matches that the algorithm has not yet surfaced, so they depend on how fast content is replenished relative to how fast recommended content is consumed. Large platforms with cheap content can refresh faster than their users consume; a moderate-scale education platform, where a high-quality story requires pedagogical expertise and professional illustration, may have difficulty doing so in a cost-effective manner.

Our study yields two design principles for practitioners. First, strategic threshold selection can simultaneously ensure algorithmic performance and enable causal identification. Our 60-exposure cutoff provided sufficient training data while creating a sharp discontinuity for local-RDD analysis. This enables continuous monitoring without costly experimental holdouts. Second, the weekly series also carries a lesson for evaluation: single evaluation windows are not representative. An assessment in the third week of deployment would have found effects matching the experiment, one in late September would have found effects several times larger, and one during Diwali would have found none. Point-in-time evaluations of learning algorithms therefore risk mistaking a snapshot for the system. Because eligibility thresholds make quasi-experimental estimates available every week at no additional cost, the natural standard for deployed personalization systems is continuous evaluation, or estimates pooled over many weeks, so that conclusions reflect the system's trajectory rather than the week it happened to be measured.

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.

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.

Al-Ubaydli, O., J. A. List, and D. L. Suskind (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review 107*(5), 282–286.

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly journal of economics 130*(3), 1117–1165.

Angrist, N., C. Cullen, M. Ainomugisha, S. P. Bathena, P. Bergman, C. Crossley, T. Letsomo, M. Matsheng, R. M. Panti, S. Sabarwal, and T. Sullivan (2023). Learning curve: progress in the replication crisis. In *AEA Papers and Proceedings*, Volume 113, pp. 482–488. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Athey, S. (2025). Presidential address: The economist as designer in the innovation process for socially impactful digital products. *American Economic Review 115*(4), 1059–1099.

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.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Athey, S. and S. Wager (2019). Estimating treatment effects with causal forests: An application. *Observational Studies 5*(2), 37–51.

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.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2018). Experimental evidence on scaling up education reforms in kenya. *Journal of Public Economics 168*, 1–20.

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.

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.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013). Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The quarterly journal of economics 128*(2), 531–580.

Davis, J. M. V., J. Guryan, K. Hallberg, and J. Ludwig (2017). The economics of scale-up. Technical report, National Bureau of Economic Research.

Drachsler, H., H. G. Hummel, B. Van den Berg, J. Eshuis, W. Waterink, R. Nadolski, A. J. Berlanga, N. Boers, and R. Koper (2009). Effects of the isis recommender system for navigation support in self-organised learning networks. *Journal of Educational Technology & Society 12*(3), 115–126.

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.

Kingma, D. P. and J. L. Ba (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2015.

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*.

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

List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature 626*(7999), 491–499.

List, J. A. (2026a). The economics of scaling early childhood programs: Lessons from the Chicago school. *Journal of Political Economy 134*(1), 1–48.

List, J. A. (2026b). Experimental economics: Theory and practice. *University of Chicago Press*.

Muralidharan, K., P. Niehaus, S. Sukhtankar, and J. Weaver (2021). Improving last-mile service delivery using phone-based monitoring. *American Economic Journal: Applied Economics 13*(2), 52–82.

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., O. Keyes-Krysakowski, and D. Agarwal (2026). How in-school supervised ed-tech support produces massive learning gains: A khan academy field experiment in india. Technical report, National Bureau of Economic Research.

Palikot, E., K. Agrawal, A. Kanodia, and S. Athey (2026). Contests, gamification, and persistent engagement in educational technology. *Quantitative Marketing and Economics 24*(1), 8.

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

Rasul, I. (2026). From field experiments to policy interventions at scale. *Science 391*(6790).

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

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.

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.

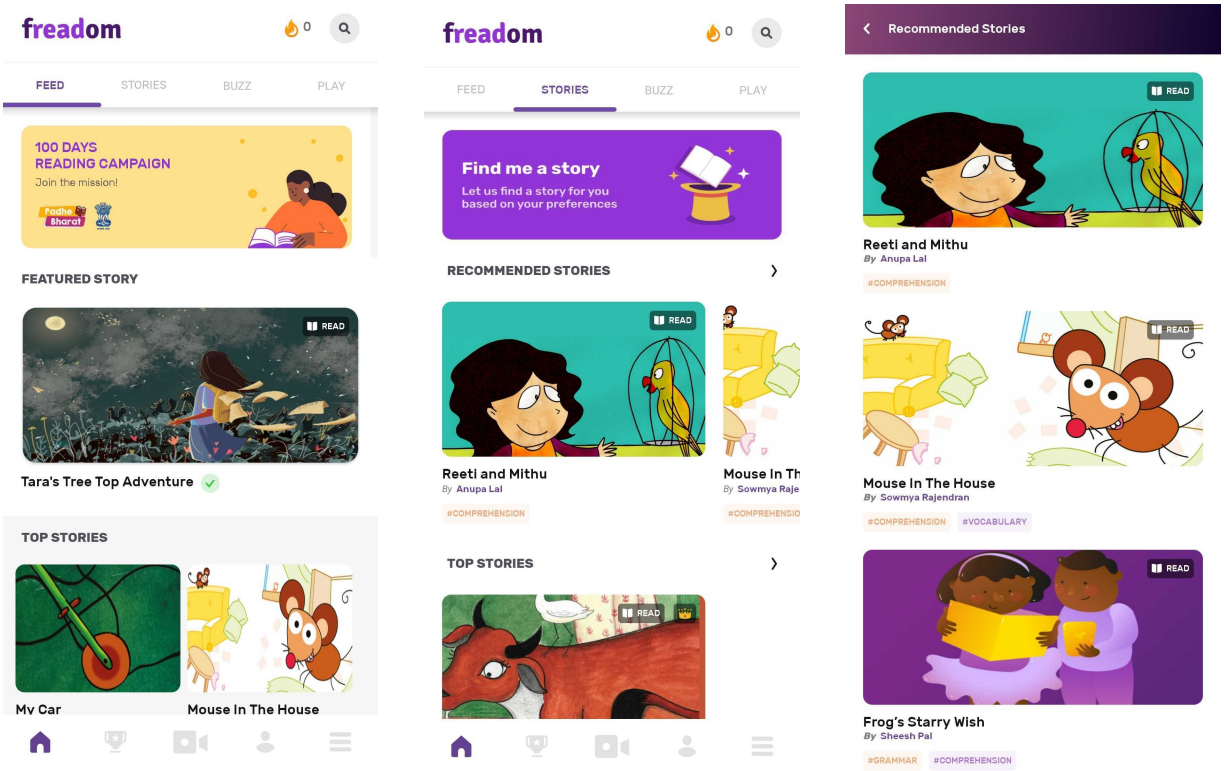
Zielnicki, K., G. Aridor, A. Bibaut, A. Tran, W. Chou, and N. Kallus (2026). The value of personalized recommendations: Evidence from netflix. *International Journal of Industrial Organization*, 103304.

Supplementary Online Appendix

A1 App Interface

Figure A1: Freadom interface

(a) Home Feed page (b) Stories page (c) Recommended Story section



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.

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, where we deployed personalized recommendations that display 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.

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:

$$\hat{U}_{ij} = \sigma(\theta_i'\lambda_j + \psi_i + \gamma_j + \beta_0), \tag{13}$$

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 a preference for humor, an interest in animals, or an affinity for longer narratives. A user's embedding θ_i encodes her preferences along these dimensions. Similarly, 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 (13).
   (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 (Kingma and Ba, 2015).

3. Continue until validation loss stops improving.

We tune the number of latent dimensions p 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 $p = 20$ latent dimensions.

A2.3 Assumptions and Limitations

We assume a low-rank structure: user-story engagement can be well approximated by the interaction of p-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 exposure 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 exposures; similarly for stories. This motivates our threshold requiring users to have at least 60 prior exposures before receiving personalized recommendations. Second, if the editorial 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.

Table A1: 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 $p = 20$ latent dimensions for users and stories.

A2.4 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 engagement; the staleness rule incorporates this implicit negative feedback.

Algorithm 1 Personalized Recommendation System

Require: Beginning-of-week user history h_{i1t} and the trained collaborative filtering model
Ensure: Daily ranking sequence $\mathbf{r}_{it} = (r_{i1t}, \ldots, r_{i7t})$

STEP 1. Ranking Generation
for each story $j \in \mathcal{J}$ **do**
    Compute predicted engagement $\hat{U}_{ij}$ using Equation (13)
end for
Sort stories so that $\hat{U}_{i,j_{(1)}} \geq \hat{U}_{i,j_{(2)}} \geq \cdots \geq \hat{U}_{i,j_{(J)}}$

STEP 2. Daily Ranking Management
Set $r_{i1t} \leftarrow (j_{(1)}, \ldots, j_{(15)})$
for each day $d \in \{2, \ldots, 7\}$ **do**
    Exclude stories recorded as started in h_{idt}
    if user i was active on days $d-1$ and $d-2$ and completed none of the top three stories in r_{i1t} **then**
        Exclude $j_{(1)}, j_{(2)}, j_{(3)}$
    end if
    Set r_{idt} to the 15 remaining stories with the highest predicted engagement
end for
return $\mathbf{r}_{it} = (r_{i1t}, \ldots, r_{i7t})$

Algorithm 1 describes how the personalized policy translates exposure histories into daily recommendations. In Step 1, the personalized policy π_t^P uses the beginning-of-week history h_{i1t} to predict engagement and order stories for user i. In Step 2, the app combines this order-

ing with the updated daily history h_{idt} to generate the 15-story daily ranking r_{idt}. Applying this procedure throughout the week produces the ranking sequence $\mathbf{r}_{it} = (r_{i1t}, \ldots, r_{i7t})$.

The daily management rules are the same under the editorial and personalized policies. These rules include removing previously started stories, limiting each ranking to 15 stories, and removing top stories under the staleness rule. Under the editorial policy π_t^{Ed}, the common editorial ordering replaces the user-specific ordering produced in Step 1, after which the same daily management rules generate r_{idt}^{Ed}.

A3 The Pre-Registered Design and Its Evaluation

We originally wrote and posted a pre-analysis plan on GitHub in 2021. The original version is available in 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.

Table A2: Treatment Effects on the Pre-Registered Seven-Day Horizon

Outcome	N(T)	N(C)	Est.	SE	p
Total Story Engagement	1,095	1,146	0.111	0.038	0.003
Total Completed Stories	1,095	1,146	0.071	0.027	0.008

Note: Effects of personalized recommendations over the pre-registered seven-day horizon, July 22 to July 28, 2021, on outcomes in the Recommended Story *section. The sample consists of children who opened the app during those seven days.*

The pre-analysis plan specified a seven-day experiment with three arms. Two arms were to receive personalized recommendations, one in the *Recommended Story* section and one in the *Today For You* section. The third arm was to receive the editorial ranking across all sections and was the control group. The experiment we ran departs from this plan in two respects: (i) we never launched personalization in the Today For You section, and (ii) the experiment ran for 13 days rather than seven.

We dropped the Today For You arm to reduce deployment complexity. The children assigned to it received the editorial ranking throughout, exactly as the control group did, and the group that we randomized into that arm is completely removed from the analysis. Stones2Milestones extended the experiment beyond seven days, so the main-text estimates cover July 22 to August 3, 2021.

The unused third arm provides a check on the randomization. These children received the same recommendations as the control group, so the two groups should differ only by sampling error. Over the full period, engagement across the app is 3.89 in the third arm and 3.79 in the control group, a difference of 0.098 (standard error 0.265). Completed stories are

1.90 and 1.89, a difference of 0.003 (standard error 0.153).

Table A2 evaluates the experiment on the pre-registered seven-day horizon, July 22 to July 28. The outcome, the estimator, and the sample construction follow the main text. Personalization raises *Total Story Engagement* by 0.111 (SE 0.038) and *Total Completed Stories* by 0.071 (SE 0.027).

A4 Randomized Controlled Trial Details

A4.1 Covariate Balance Check

Table A3 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 A3: 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.

A4.2 Effects on Learning Inputs

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 A4: 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. The time period corresponds to the RCT period that ran from July 22 to August 4 2021. Moreover, the unit for reading time is minutes. *** p<0.01, ** p<0.05, * p<0.10.*

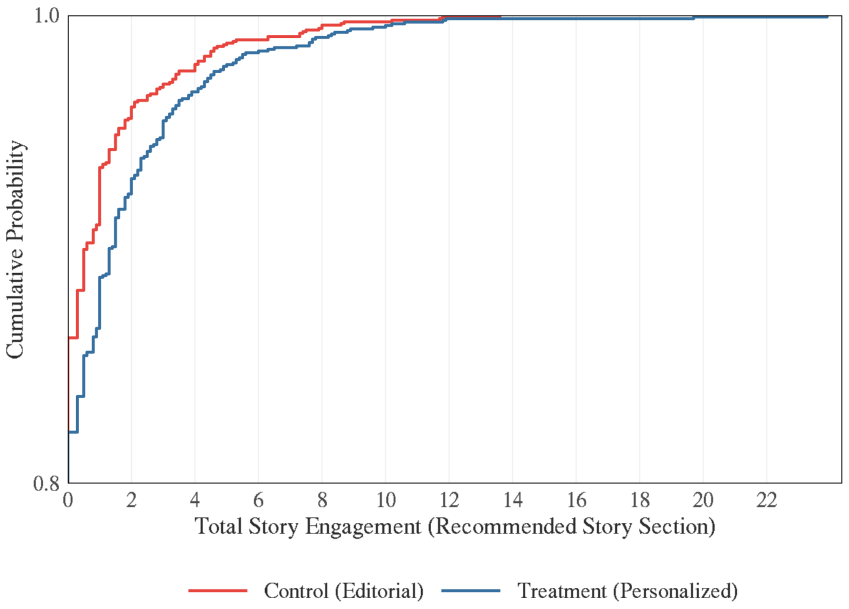
Table A4 presents average treatment effects on these learning-related outcomes measured on the platform. Personalization significantly increased reading time: treated users spent 87 percent more time reading in the *Recommended Story* section (S.E. 26%) and 20 percent more time reading app-wide (S.E. 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.

A4.3 Distribution of Engagement in the Recommended Story Section

This appendix compares the full distributions of Total Story Engagement in the Recommended Story section between the treatment and control groups of the RCT, which ran from July 22 to August 4, 2021. Figure A2 plots the cumulative distribution function for each

group. The treatment distribution first-order stochastically dominates the control distribution: at every level of engagement, a smaller fraction of treated users falls below it. The gap is widest at the bottom of the distribution. Among control users, 87 percent had no engagement with the section, compared with 82 percent of treated users. This five-percentage-point difference on the extensive margin accounts for a large share of the treatment effect. The distributions also diverge at higher levels of engagement, so personalization raised engagement on the intensive margin as well. A Wilcoxon rank-sum test rejects the hypothesis of no location shift for section-specific engagement ($p < 0.001$) and for app-wide engagement ($p = 0.05$).

Figure A2: Distribution of Engagement in the Recommended Story Section



Note: Cumulative distribution functions of Total Story Engagement in the Recommended Story section by treatment status during the RCT, July 22 to August 4, 2021. The y-axis begins at 0.8 to show the upper part of the distribution.

A5 Validation of the RDD Design Using Experimental Data

This appendix reports the placebo tests underlying the bandwidth choice in Section 7.1. Using data from the RCT period, we compare two control groups. The Control RCT group is the experiment's randomized control arm: users who had at least 60 story exposures before the experiment but received editorial recommendations. The Control RDD group consists of users with fewer than 60 exposures 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 the equivalent of the control group in the RDD design. Neither group received personalized recommendations, so if local randomization holds, the difference between their mean outcomes should be small. Table A5 presents this comparison for bandwidths between 20 and 55 in increments of 5, estimated using the local randomization RDD estimator of Cattaneo et al. (2015).

Table A5: Placebo Tests: Control RCT versus Control RDD, Total Engagement

BW	N(L)	N(R)	Est.	P-val
20	1507	345	0.079	0.198
25	2062	408	0.049	0.471
30	2820	470	0.046	0.362
35	3736	534	0.047	0.360
40	4858	599	0.083	0.077
45	6408	646	0.089	0.044
50	8363	703	0.106	0.021
55	11130	749	0.142	0.001

Note: This table reports placebo tests comparing the Control RCT group (the experiment's randomized control arm, above the threshold) to the Control RDD group (users outside the RCT who fall below the threshold), using local randomization RDD with varying bandwidths around the threshold of 60 story exposures. For each bandwidth w, the window is $[60 - w, 60 + w)$. Estimates represent differences 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 (Control RDD) and $N(R)$ corresponds to the number of observations above the threshold (Control RCT). The time period corresponds to the RCT period, July 22 to August 4th.

For bandwidths ≤ 35, the tests comparing *Control RCT* vs *Control RDD* yield p-values

ranging from 0.198 to 0.471. Under the null hypothesis of no difference between these groups, we would observe effect estimates as large as or larger than these in 19.8%-47.1% 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), p-values fall below 0.05, indicating that such large differences would be unlikely under random assignment alone and suggesting potential violations of the local randomization assumption. The comparison of the two control groups thus shows that users just below the threshold are similar to users just above who received editorial recommendations, at least within a bandwidth of 35. The differences between treated users and Control RDD in the left panel of Table 6 can therefore be attributed to personalization rather than pre-existing differences, and we select a bandwidth of 35 for the main analysis, the largest for which we have evidence supporting local randomization.

A6 Validation of the Regression Discontinuity Design during Deployment

In this section, we first compare engagement of treatment and control users in the week before the week in which the treatment group crossed the 60-exposure eligibility threshold. Second, we run a placebo test by considering a section of the app with no personalized recommendations; thus, recommendations do not change at the 60-story exposure threshold.

A6.1 Pre-Treatment Balance Check

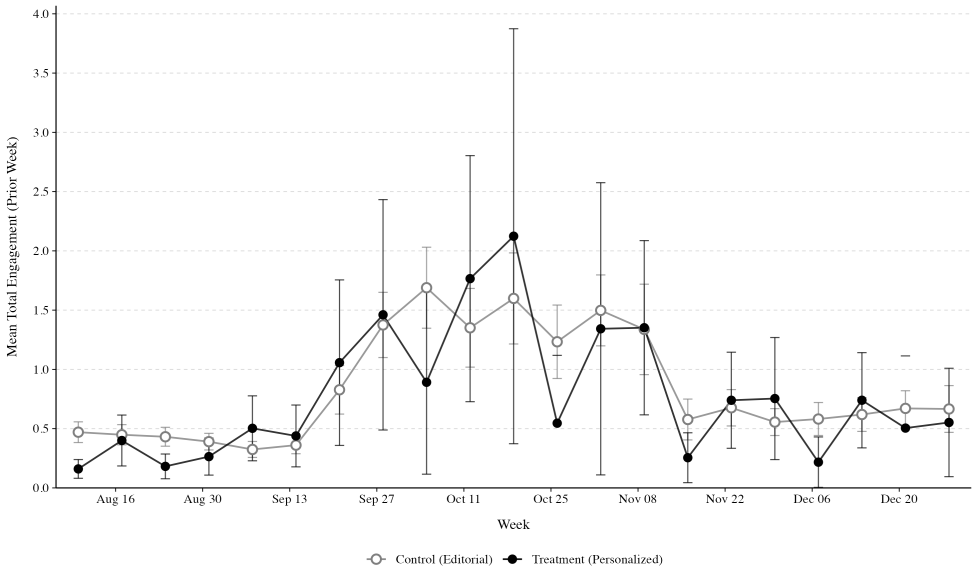
This analysis examines whether users just above versus just below the 60-exposure threshold exhibit comparable engagement levels prior to treatment assignment (crossing the 60-story threshold). We compare mean Total Story Engagement in the week before threshold crossing between treated users (60–94 interactions) and control users (25–59 interactions) throughout the deployment period. Figure A3 presents weekly comparisons of prior engagement by treatment status from August through December 2021.

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 not statistically significant at the 5% level. The pattern holds across both high-engagement periods (September–October) and low-engagement periods (November–December during holidays).

A6.2 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-exposure threshold provide valid counterfactuals during the RCT period. A natural question is whether this identification assumption holds throughout the

Figure A3: Pre-Treatment Balance: Prior Week Engagement by Treatment Status



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

scaled deployment period. We monitor this by examining whether discontinuities emerge at the 60-exposure threshold in sections of the app where we never deployed personalized recommendations. 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 exposures (treated) to users with 25-59 prior exposures (control) within each week. Crucially, the running variable cumulative story exposures (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

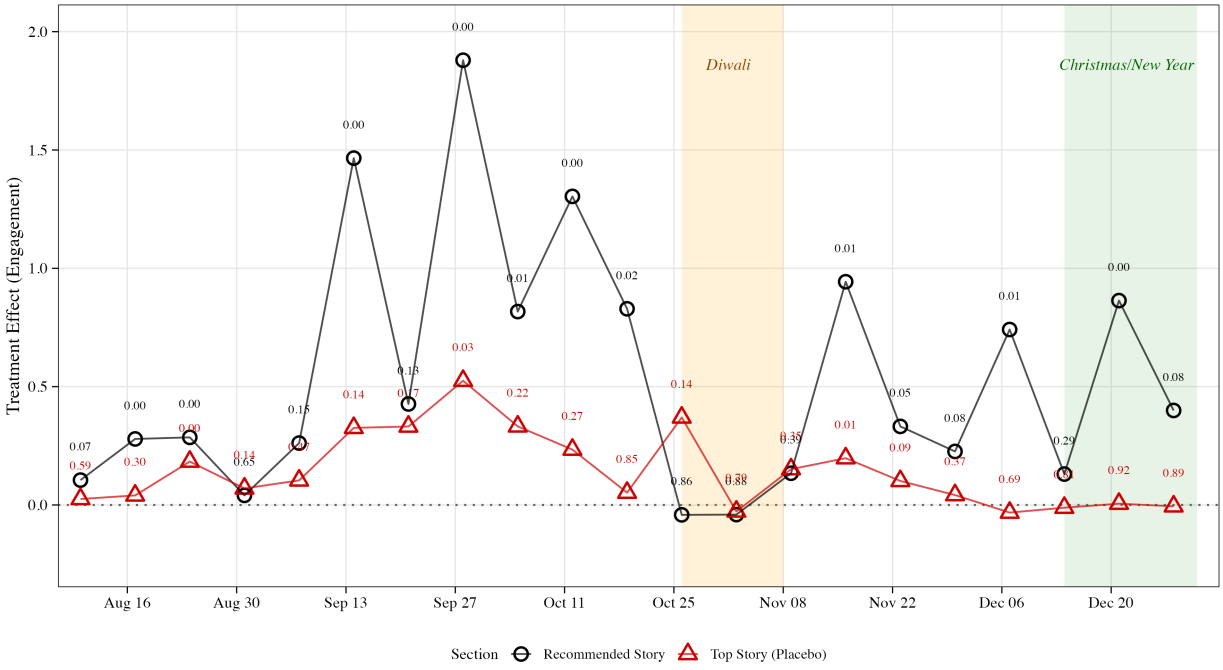
60 exposures reflect the causal effect of personalized recommendations, we should observe significant treatment effects only in the Recommended Story section. If instead the discontinuity captures user maturation, selection effects, or other confounders that operate at the threshold regardless of treatment assignment, we should observe similar patterns in both sections.

Figure A4 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 exposures 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 users who receive satisfying personalized recommendations in the Recommended Story section subsequently consume more content in other sections. If such spillovers were substantial,

Figure A4: Weekly Placebo Test: Recommended Story vs. Top Story Section



Note: Local average treatment effects of crossing the 60-exposure 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.

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 mostly in the treated section.

A7 The discontinuity design inside the experiment

The scaled-period analysis compares children just above and just below the eligibility threshold, week by week, with local-randomization inference. As a check on that design we run the same comparison inside the experiment, where treatment assignment is known. The treated group is the experimental arm that received personalized recommendations, restricted to users whose past exposure count at the start of the week lies within the bandwidth above the threshold. The comparison group is drawn fresh each week, as in the scaled-period analysis: users below the threshold and within the same bandwidth. The outcome is total story engagement in the recommendation tray during the week, with zeros for inactive users. The estimates follow the scaled-period procedure.

Splitting the experiment into two weeks also makes the comparison to the weekly scaled-period estimates more direct: the recommendation system was redeployed between the two weeks, so each week's estimate reflects a distinct model vintage, just as each weekly scaled-period estimate does. Table A6 reports the estimates for the first week of the experiment (July 22-28, 2021), the second week (July 29-August 4), and the two weeks pooled, across bandwidths.

At the bandwidth of 35 used in the main analysis, the discontinuity is 0.106 units of total story engagement in the first week ($p < 0.001$) and 0.090 in the second ($p = 0.008$); pooled, 0.098 ($p < 0.001$). The estimates are positive and significant at every bandwidth in both weeks. They grow with bandwidth as the comparison group moves further below the threshold and the treated window moves further above it.

One caveat qualifies the comparison across the two weeks. In the second week, the treated children are in their second week of exposure to personalization, whereas the scaled-period estimates track children in their first week after crossing the threshold. The second-week estimate therefore mixes the contemporaneous effect of the redeployed system with whatever carries over from the first week of exposure; the first-week estimate is a cleaner comparison of a weekly scaled-period estimate.

Table A6: Discontinuity estimates inside the experiment: personalized arm against children below the eligibility threshold.

	Week 1 (Jul 22-28)			Week 2 (Jul 29-Aug 4)			Pooled		
Bandwidth	Estimate	p	N	Estimate	p	N	Estimate	p	N
20	0.069	0.013	1,511	0.103	0.045	1,513	0.085	0.006	3,024
25	0.075	0.001	2,088	0.090	0.039	2,080	0.082	<0.001	4,168
30	0.097	0.003	2,713	0.084	0.020	2,728	0.091	0.001	5,441
35	0.106	<0.001	3,533	0.090	0.008	3,499	0.098	<0.001	7,032
40	0.131	<0.001	4,556	0.103	0.004	4,512	0.117	<0.001	9,068
45	0.119	<0.001	5,910	0.113	<0.001	5,913	0.116	<0.001	11,823
50	0.123	<0.001	7,818	0.126	<0.001	7,863	0.124	<0.001	15,681
55	0.144	<0.001	10,879	0.143	<0.001	10,904	0.144	<0.001	21,783

Note: Each cell compares users in the personalized arm of the experiment whose past exposure count at the start of the week lies in $[60, 60+bw)$ *against children below the threshold, in* $[60-bw, 60)$*. The outcome is total story engagement in the recommendation tray during the week. Estimates are differences in means within the window; p-values are calculated from local randomization inference, as in the scaled-period analysis.*

A8 Change in the Composition of Users Crossing the Threshold.

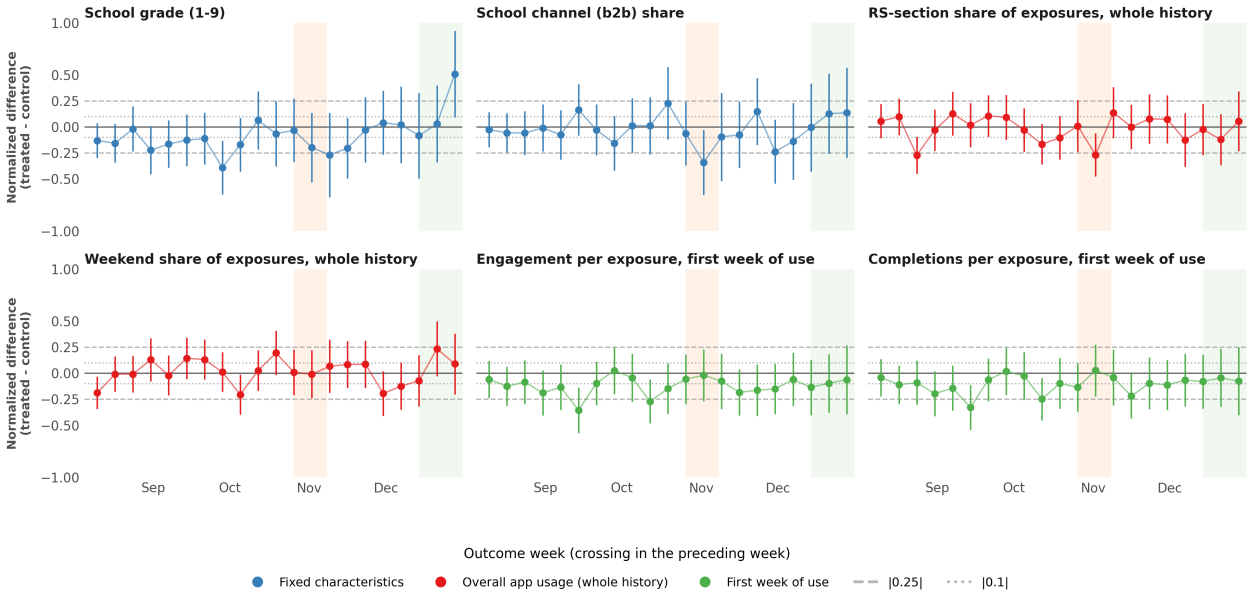
Two explanations can produce a rising weekly estimate. The algorithm may improve as weekly retraining accumulates data, the explanation we take up in Section A9. Or the new groups may drift apart. Each week a new cohort of users crosses the threshold. If later crossers differed from the users just below the cutoff in ways that raise engagement, the estimate would rise with no change in the algorithm.

Figure A5 plots, for each week, the normalized difference between treated and control users in six characteristics fixed before the crossing week. The six fall into three groups. The first is who the users are: school grade, and the share reached through schools rather than parents (the b2b channel). The second is how they use the app, which is the share of their story exposures on the Recommended Story section and the share on weekends, both over the user's whole history before the crossing week. The third is how they engaged when they started: engagement per story exposure, and active days out of seven, in their first week on the app.

The two groups match on these characteristics. Pooled over the period, the normalized differences are -0.11 for grade, -0.03 for the school channel, -0.01 and 0.01 for the two usage shares, and -0.12 and 0.13 for first-week engagement per exposure and active days. There is also no trend over the 21 weeks. Section 7.1 validated the design in the first week of the deployment against the experiment. The experiment's randomized control arm and the users just below the threshold, both on editorial recommendations, are indistinguishable at bandwidths up to 35. The absence of trends is suggestive that the comparison remains valid over time.

The two groups differ only by the crossing week. Figure A6 repeats the exercise for three measures of activity in the week they cross the 60-exposure eligibility threshold for Total Story engagement (our main metric), story exposures per app launch, and the number of

Figure A5: Balance Between Treated and Control Users Who Become Eligible Each Week:

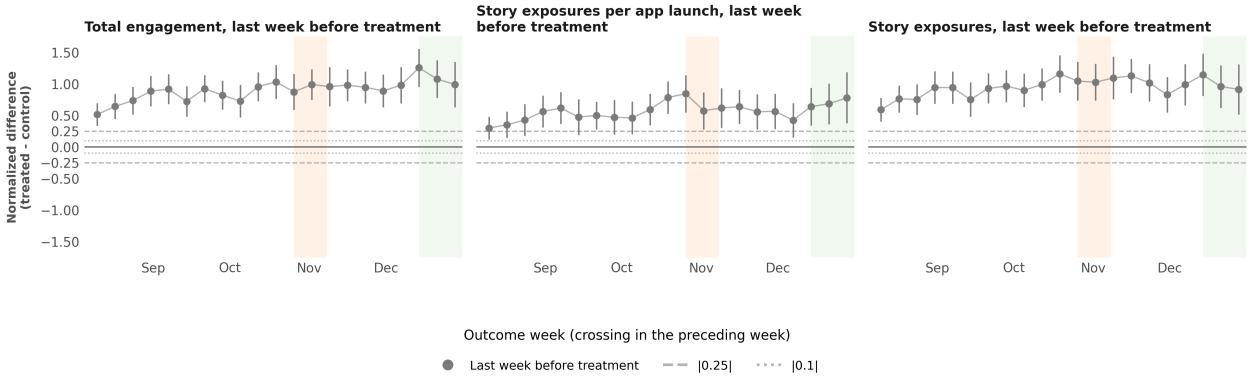


Note: Each point is the normalized difference between treated and control users with a 95% confidence interval; dashed lines at ±0.25, dotted lines at ±0.1. Top row: school grade and the school-channel share from the registration record; the share of a user's story exposures on the Recommended Story section and the share on weekends, over the user's whole history before the crossing week. Bottom row: engagement per story exposure, and active days out of seven, in the first week after first app registration. Orange shading indicates Diwali; green shading indicates the winter break.

story exposures. By construction, treated users differ on all three, since increased activity is what carries a user over the threshold. The question is whether the difference moves across the three periods of Figure 3. For Total Engagement it does not. Treated users accumulate 4.1 to 5.4 units of engagement in the crossing week, against 1.4 to 2.5 for controls, a difference of 0.82 standard deviations pooled over the period; estimated separately for each period, with week fixed effects and standard errors clustered by user, the normalized difference is 0.86, 0.82, and 0.80, and differences in pair-wise comparisons are statistically not significant. The same holds for exposures per app launch. The number of story exposures shows a small upward drift, from 0.75 to 0.97. Our main metric, however, does not trend at any point in the period: the split between the two groups is the same throughout the deployment.

Thus, we compare treated and control users at three moments between registration in the app and the week in which the treated group crosses the eligibility threshold. First, users who

Figure A6: Activity in the Crossing Week for Treated and Control RDD



Note: Each point is the normalized difference between treatment and control, with a 95% confidence interval, for three measures of activity in the crossing week, the seven days before the outcome week: total story engagement, story exposures per app launch, and the number of story exposures, all three computed across all sections. Dashed lines at ± 0.25*, dotted lines at* ± 0.1*. Orange shading indicates Diwali; green shading indicates the winter break.*

will cross the threshold and those who will stay just below it start with similar engagement. Figure A5 shows no difference in engagement per exposure or active days in their first week on the app, in any week of the deployment. Second, they stay alike through the week before crossing; we find that only two of 21 weekly differences are significant at the 5 percent level (Appendix A6.1, Figure A3). Finally, the difference appears in the crossing week, and Figure A6 shows it is stable over time. This pattern is consistent with the explanation that the two groups are similar until the crossing week, when some users receive more exposures for reasons unrelated to their observed characteristics or earlier behavior, plausibly at random, and that this shock remains stable over the entire scaled deployment period.

A9 Story Ranking and Model Improvement

The recommendation model affects user engagement by determining which stories a user sees and in what order. Each week the system builds a ranking for every eligible user and displays its top 15 stories in the Recommended Story (RS) section. We ask whether the stories a user's ranking places higher draw more engagement per exposure from that user. We group stories into three treatments by their position on the user's ranking: the top half, the bottom half, and off the ranking. We compare engagement per exposure across treatments within user and average the contrasts across users.

The main challenge is that, throughout the app, editors and ranking algorithms select the stories they judge best and place them where users are most likely to see them. As a result, in every section a story's prominence and its quality might move together, and the engagement a story draws reflects both. The personalized recommendation system is a particular case of this problem, with the ranking tailored to each user: the model ranks a user's stories by the engagement it predicts for that user, and the app displays the top of the ranking in the RS section. The ranking is therefore not only a prediction of the user's taste but also a decision about what the user is exposed to. When a top-ranked story draws more engagement, we cannot tell whether the model predicted the user's taste or whether the display produced the engagement. Which stories a user meets, and how the user responds to them, also depend on when the user logs in, which sections the user visits, and how far the user has gone in the session. Finally, earlier exposures shape later ones. What a user has already seen changes what the user reads next, whether the session continues, and which stories remain available. Treatment (rank assignment), in short, is not randomized, across users or across exposures of the same user, and a comparison of engagement per exposure across treatments does not by itself reveal whether the ranking sorts stories by the quality of the match.

Our identification strategy has two parts. First, we restrict attention to exposures outside the RS section for eligible users who never open that section during the tracking window. For these users the ranking of the RS section does not affect which stories are displayed in other

sections. Second, within this sample we assume that treatment is unconfounded conditional on observables. We state the conditions formally below. The focus on a target population comprised of users who never open the RS section favors internal over external validity. Users who do open the RS section are heavier users, so our estimates do not necessarily generalize to them.

We focus our analysis on the six weekly ranking cycles from July 21 through September 2, 2021, for which daily ranking logs are available. Formally, we use $\mathcal{T}^{\text{rank}}$ to denote this period. Rankings change weekly with model retraining and daily through slate-management rules that use the user's evolving history (see Appendix A2.4 for details). We therefore restrict attention to eligible users with no recorded RS exposure during $\mathcal{T}^{\text{track}}$, the full deployment-period tracking window. Define $Q_i = \mathbf{1}\{i \text{ is eligible and no RS exposure in } \mathcal{T}^{\text{track}}\}$.

Exposure histories and potential outcomes. Following the conceptual framework in Section 4, i indexes users, t weeks, and d days within a week. Let s index sessions on the app within day (d,t) and k denote the qualifying exposures within a session where $s = 1,\ldots,N^{\text{ses}}_{idt}, k = 1,\ldots,N^{\text{exp}}_{idts}$. Both the number of sessions N^{ses}_{idt} and the number of exposures N^{exp}_{idts} are stochastic outcomes of the user's app-use process and may depend on the user's prior observed history. Their realized values in the data are denoted by n^{ses}_{idt} and n^{exp}_{idts}, respectively. Let C_{idtsk} be the exposure's timestamp and $M_{idtsk} \neq \text{RS}$ its section. We use $\ell = (t,d,s,k)$ to write the indexing in a compact manner. Let $\mathcal{E}_i$ collect exposure indices during $\mathcal{T}^{\text{rank}}$, and set $K_i = |\mathcal{E}_i|$. We treat the indices, study periods, and deployed model vintages as fixed. In particular, $\mathcal{E}^{\text{obs}}_i$ is the recorded exposure set and $k_i = |\mathcal{E}^{\text{obs}}_i|$.

Let H^-_{il} extend the beginning-of-day history H_{idt} to immediately before the current exposure needed for this exposure level ranking analysis. In the estimation section, we describe the variables included in the history. Let $B_{il} \in \mathcal{B} = \{\text{top}, \text{bottom}, \text{off}\}$ denote the current exposure's rank bin in the user's personalized ranking on that day. Ranks 1-7, ranks 8-15, or outside the top 15, respectively. For $b \in \mathcal{B}$, let $U_{il}(b)$ denote the potential engagement per

exposure that would be realized if, at this current exposure, the exposure were filled from rank bin b, holding fixed the pre-exposure history H_{il}^{-}. Thus $U_{il}(\text{top})$ and $U_{il}(\text{bottom})$ are the engagement outcomes that could occur at the same current exposure under the two alternative rank-bin interventions. Only the potential outcome corresponding to the realized rank bin is observed. If $B_{il} = b_{il}$, consistency implies $u_{il} = u_{il}(b_{il})$. The potential-outcome comparison in our setting concerns engagement at the current exposure. Subsequent histories, sessions, and exposure counts may differ across current-bin interventions.

Equal-user estimand. Let $\mathcal{V}_i$ denote user i's stochastic app-use trajectory during $\mathcal{T}^{\text{rank}}$. It includes, for example, when the user logs in, the number and length of sessions, the sections visited, the number, timing, and order of exposures, the evolving pre-exposure histories H_{il}^{-}, and the potential engagement outcomes $U_{il}(b)$ at reached exposures. Let F_Q denote the population distribution of a user and their trajectory conditional on $Q_i = 1, K_i > 0$. Here we draw a user with a trajectory, then select an exposure uniformly from that trajectory $(I, \mathcal{V}_I) \sim F_Q, E \mid I = i, \mathcal{V}_i = v_i \sim \text{Unif}\big(\mathcal{E}(v_i)\big)$. Here $\mathcal{E}(v_i)$ is the exposure set of the drawn trajectory. For bins b, b' our estimand is

$$\tau_{b,b'}^{Q} = \mathbb{E}_{(I,\mathcal{V}_I)\sim F_Q}\left[\mathbb{E}_{E|I,\mathcal{V}_I}\left\{U_{IE}(b) - U_{IE}(b')\right\}\right]. \tag{14}$$

The inner expectation averages across the user's qualifying exposures based on the restrictions we imposed. The outer expectation gives users equal weight and integrates over their random trajectories.

Identification. First, we assume history sufficiency for carryover that is conditional on the observed history, prior rank-bin assignments have no additional effect on either current potential engagement or the subsequent exposure process. Thus, earlier assignments may affect engagement, continuation, and future exposures, but these effects operate through the evolving observed history. Second, we assume sequential conditional exchangeability. For an

exposure,

$$\{U_{IE}(b) : b \in \mathcal{B}\} \perp\!\!\!\perp B_{IE} \mid H^{-}_{IE}. \tag{15}$$

We assume that conditional on the history and current context, the encountered rank bin is as-if random with respect to current potential engagement. Third, each bin entering a contrast must have positive probability at the histories in the target population. Finally, we assume consistency, no anticipation of future assignments, and no contemporaneous cross-user interference conditional on calendar context and deployed model vintages. These assumptions would be weaker if we focused only on a user's first exposure, but we do not have sufficient power when the sample is restricted in that way. Instead, we assume that observable features of history capture those elements of history that affect potential engagement.

Estimation. Our estimator is the augmented inverse-propensity weighting (AIPW) estimator, implemented with causal forests (Athey et al., 2019; Athey and Wager, 2019). Let $\mathcal{Q} = \{i : q_i = 1, k_i > 0\}$ be the observed estimation sample and $n_Q = |\mathcal{Q}|$. We stack all exposures of users in $\mathcal{Q}$ to stories in the catalog from which the rankings were built and construct the history variables from information available before each exposure. These variables represent H^{-} in estimation.

The history enters through three blocks of covariates. User and calendar context: age, grade, pre-period activity, calendar day, day of week, and the section in which the exposure occurred. Session state: the engagement the user accumulated earlier in the same app session, the minutes since the previous event, and the minutes since the session opened, all zero for the first exposure of a session. Story characteristics: reading level and its distance from the user's level, type (read, listen, or watch), reading-time bucket, quiz flag, source, collection, the story's age in days, and its popularity before the tracking window, measured by the catalog's view, read, and story like counts and by exposures, distinct readers, view rate, and completion rate across all children before July 21, 2021.

For each pair of bins (b, b') we fit a causal forest on the exposures in those two bins

(Athey et al., 2019), with observations clustered by user. All subsampling then operates on clusters rather than on individual exposures: each tree is grown on a random subsample of users, and honest splitting divides that subsample in half at the user level. Predictions for user i are out-of-bag, computed from the trees whose subsample did not contain user i. The propensity $\widehat{p}_b$ and the bin-specific conditional means $\widehat{m}_b$ and $\widehat{m}_{b'}$ that enter the scores are out-of-bag in the same sense. Writing $\widehat{p}_b^{(-i)}$ and $\widehat{m}_b^{(-i)}$ for these estimates, define

$$\widehat{\phi}_{i\ell}(b) = \widehat{m}_b^{(-i)}(h_{i\ell}^-) + \frac{\mathbf{1}\{b_{i\ell} = b\}}{\widehat{p}_b^{(-i)}(h_{i\ell}^-)} \left\{ u_{i\ell} - \widehat{m}_b^{(-i)}(h_{i\ell}^-) \right\}, \tag{16}$$

with $\widehat{p}_{b'}^{(-i)} = 1 - \widehat{p}_b^{(-i)}$ within the two-bin sample. Let $\mathcal{E}_i^{\text{obs}}(b, b')$ collect user i's exposures in the two bins, let $k_i(b)$ be their number in bin b, and give exposure ℓ the weight $w_{i\ell} = 1/k_i(b_{i\ell})$. The estimator is

$$\widehat{\tau}_{b,b'}^{\mathcal{Q}} = \frac{\sum_{i \in \mathcal{Q}} \sum_{\ell \in \mathcal{E}_i^{\text{obs}}(b,b')} w_{i\ell} \left\{ \widehat{\phi}_{i\ell}(b) - \widehat{\phi}_{i\ell}(b') \right\}}{\sum_{i \in \mathcal{Q}} \sum_{\ell \in \mathcal{E}_i^{\text{obs}}(b,b')} w_{i\ell}}. \tag{17}$$

It averages exposure-level scores within each user and then across users so every user carries the same weight within a bin. We report this average treatment effect (ATE) and the average effect on the treated (ATT), which averages the effect predictions over the exposures in bin b and applies the correction to bin b' with the odds $\widehat{p}_b/(1 - \widehat{p}_b)$ as weights.

Table A7: Engagement by ranking placement, within-user contrasts

	AIPW ATE, by covariate set			AIPW ATT
	User and calendar	+ Session state	+ Story characteristics	Full set
Top vs. off ranking ($\tau_{\text{top,off}}$)	0.122	0.113	0.059	0.059
	(0.014)	(0.013)	(0.009)	(0.014)
Bottom vs. off ranking ($\tau_{\text{bottom,off}}$)	0.095	0.091	0.031	0.030
	(0.012)	(0.012)	(0.008)	(0.012)
Top vs. bottom of ranking ($\tau_{\text{top,bottom}}$)	0.026	0.022	0.024	0.026
	(0.014)	(0.013)	(0.013)	(0.013)

Notes: The covariate sets are cumulative: the first column uses the user and calendar controls; the second adds session state; the third adds story characteristics. Standard errors are clustered by user.

Table A7 reports the results. Stories on the ranking draw substantially more engagement per exposure than stories off it. Engagement per exposure averages 0.401 for top-half stories, 0.387 for bottom-half stories, and 0.265 for off-ranking stories. With the full history held

fixed, the AIPW estimates are 0.059 (SE= 0.009) for the top half and 0.031 (SE=0.008) for the bottom half, 22 and 12 percent of the off-ranking mean. The ATT estimates are similar to ATE. Within the ranking, the top half outperforms the bottom half by 0.024 (SE=0.013) under the ATE and 0.026 (SE=0.013) under the ATT, about 6 percent of the bottom-half mean.

Retraining improves the ranking week to week. The second question is whether the weekly retraining improves the ranking. The placement analysis compares stories at different positions of one ranking; here we compare two consecutive rankings of the same user, the one the model built in week $t-1$ and the one it built in week t after retraining. The unit of analysis is a transition, the pair of rankings a user holds in consecutive weeks. We observe five transitions and pool them.

Let Q'_{it} indicate that $Q_i = 1$, $K_i > 0$, user i holds personalized rankings in weeks $t-1$ and t, and had no exposure during week $t-1$ to any story on the initial ranking $R_{i,1,t-1}$. Let q'_{it} denote its observed realization. The observed retraining sample for transition $t-1$ to t is $\mathcal{Q}'_t = \{i \in \mathcal{Q} : q'_{it} = 1\}$. The last condition ensures that the old ranking reached the retraining unmodified, so that the two rankings differ only through retraining. Stories on both rankings contribute the same outcome to both and cancel, so the comparison is between the stories the retraining added and the stories it dropped.

The outcome, the exposures, and the logic of the comparison carry over from Table A7. As before, users never see the ranking, so a ranked story reaches them only through other sections of the app, and the outcome is engagement per exposure. The exposures we consider are the user's outside-RS exposures during week t to stories that the transition added or dropped. The treatment is whether the story was added or dropped, in place of its rank bin. Evaluating each ranking over the week it was in force keeps the outcome windows of successive transitions disjoint, so a story that enters the ranking one week and leaves it later is never counted in both arms with the same outcome.

The estimand is the population average difference in potential engagement per exposure when the current exposure is filled from the added rather than the dropped group, among outside-RS exposures satisfying the retraining restrictions. We estimate the population contrasts over the qualifying exposures of users in $\mathcal{Q}'_t$ pooled across the five transitions. We estimate this contrast with the AIPW estimator of Table A7, one causal forest with user-level clustering, and report the ATE and the ATT with the added arm as the treated arm.

Table A8: Engagement with stories added versus dropped by retraining

	Dropped mean	AIPW ATE	AIPW ATT
Added vs. dropped	0.324	0.072	0.072
		(0.032)	(0.033)

Notes: The sample pools the five transitions between consecutive rankings for users in $\mathcal{Q}'_t$ and evaluates each ranking over the week it was in force. Standard errors, in parentheses, are clustered by user.

Table A8 reports the results. Engagement per exposure averages 0.420 for added and 0.324 for dropped stories. Retraining adds 0.072 more engagement per exposure (standard error 0.032), 22 percent of the dropped-story mean, and the ATT gives the same figure. Retraining moves a user's ranking toward stories that user engages with more.

Finally, we provide covariate balance between treated and control arms for the AIPW contrasts below.

Covariate balance. We examine whether the covariates are balanced between the treated and control arms of each AIPW contrast: the three placement contrasts in Table A7 and the added-versus-dropped contrast in Table A8. For each covariate we compute the standardized mean difference, the difference in means between the two arms divided by their pooled standard deviation in the estimation sample. We report three differences. The unadjusted difference uses only the estimator's sample weights, so it describes the sample the forest sees. The two adjusted differences reweight by the propensities that enter the AIPW scores: under ATE weights a treated observation counts $1/\widehat{p}$ and a control observation $1/(1-\widehat{p})$; under ATT weights a treated observation counts one and a control observation the odds $\widehat{p}/(1-\widehat{p})$.

The propensities are the forests' own out-of-bag predictions, so the adjusted differences show the balance the estimator's weights achieve. Figure A7 presents the results, one panel per contrast.

The placement contrasts against off-ranking stories start far from balance, and the imbalance sits in the story block. Ranked stories are more popular and more often completed: the pre-window completion rate differs by 1.2 standard deviations, the view rate by 1.0, and the catalog's like, read, and view counts by 0.3 to 0.5. Ranked stories are also more often the platform's own titles, by 0.5 to 0.6 standard deviations, less often videos, and more often reached from the *Trending Now* section. User characteristics, calendar context, and session state differ by less than 0.1 standard deviations before any adjustment. This is the pattern behind the covariate sets of Table A7: story characteristics, not user or session characteristics, move the on-ranking premium.

The two sets of weights balance these contrasts unequally. ATT weights bring the mean absolute difference from 0.19 to 0.06 for top versus off and from 0.18 to 0.05 for bottom versus off; the largest remaining difference is below 0.3, and fewer than a fifth of the covariates exceed 0.1. ATE weights reduce the means to 0.13 and 0.11 and leave the popularity measures near 0.5 and the publisher indicators at 0.2 to 0.4. The difference comes from the tail of the propensity distribution. ATE weights give off-ranking exposures with propensities near 0.014 weights of up to 70, and a few such exposures dominate the reweighted control arm; ATT weights never exceed the odds of a treated exposure. The AIPW estimator does not rely on the weights alone: the outcome model corrects the remaining imbalance, and the score is consistent if either model is correct. The ATE and ATT estimates in Table A7 agree to the third decimal.

The within-ranking contrast and the retraining contrast are close to balance before adjustment and balanced after it. Top-half and bottom-half stories differ by at most 0.16 standard deviations on any covariate, publisher and collection included, and no covariate exceeds 0.1 after weighting under either target. For added versus dropped slots the only

unadjusted difference above 0.1 is the transition indicator, which the forest removes. The two arms of the retraining comparison come from the same users in the same weeks, and that delivers the balance by design.

Figure A7: Covariate balance before and after AIPW weighting

Note: Absolute standardized mean differences between the treated arm (listed first in each panel title) and the control arm, ordered by the unadjusted difference. Filled circles: unadjusted. Open squares: ATE weights. Open triangles: ATT weights. Dashed lines mark 0.1 and 0.25.

A10 Adjacent-Cohort Comparison

Section 7.3 follows each cohort of crossers against the entire comparison group of the weekly design. This appendix repeats the analysis for the comparison group that becomes eligible one week after the treated group meaning for each cohort g, the users just below the threshold at g who cross it during week g and are therefore first eligible in week $g+1$. In week $g+c-1$ the crossers of cohort g are at tenure c and the comparison users at tenure $c-1$, so the contrast at tenure $c \geq 2$ is the difference in the gain between adjacent tenures under the same vintage. At $c = 1$ the crossers' are in first personalized week and the comparison users' are still in the editorial system; this is the comparison of Section 2.

We use the entire RDD comparison groups in Section 7.3 because it offers a much larger sample size. The comparison users who cross in the following week are 888 of the 13,062 comparison observations of the weekly design, 6.8 percent, in 20 cohorts; the cohort of December 28 is dropped because its following week is outside the panel. Stacking cohorts and tenures as in Equation (6) gives 15,659 observations in 110 cells. We estimate one coefficient per tenure, pooled over cohorts, with the cell fixed effects, with and without the covariates; the baseline interaction is omitted, because the comparison mean here includes the comparison group's own personalization gain. Results are in Table A9.

Table A9: Difference in Engagement Between Adjacent Tenures, Same Week

Tenures compared	2 vs 1	3 vs 2	4 vs 3	5 vs 4	6 vs 5
Comparison group: the cohort that crossed one week later					
Difference in means, randomization inference	−0.34 [0.003]	−0.27 [0.002]	−0.07 [0.43]	−0.15 [0.08]	−0.05 [0.49]
Stacked regression, cell fixed effects	−0.22 (0.10)	−0.25 (0.09)	−0.18 (0.09)	−0.12 (0.09)	−0.20 (0.09)
with covariates	−0.18 (0.10)	−0.21 (0.09)	−0.13 (0.09)	−0.07 (0.09)	−0.15 (0.09)

Note: Total Story Engagement in the Recommended Story section. First block: crossers of cohort g at tenure c against users just below the threshold at g who crossed during week g, observed in the same week at tenure $c-1$. Row 1: difference in means pooled over all cohorts, with randomization-inference p-values in brackets. Rows 2 and 3: one coefficient per tenure from the stacked regression with cell fixed effects, standard errors clustered by user in parentheses; row 3 adds the observed and behavioral covariates of Equation (6).

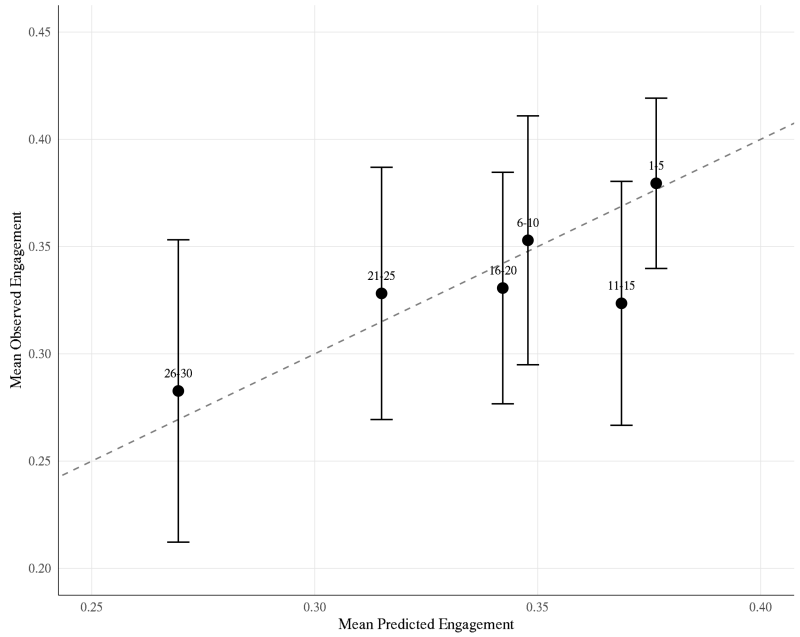
Every contrast between adjacent tenures is negative, which is consistent with our main analysis. Users one week further into eligibility engage less with the personalized section

than users one week behind them, under the same vintage, and the gain from personalization therefore declines with each week of eligibility. The adjacent-cohort differences are larger than the steps of the main profile, by about 0.1 per week, and they do not shrink at later tenures. Part of the difference is the increase in engagement in the crossing week (Figure A6).

A11 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 A8: Calibration plot by within-user story rank

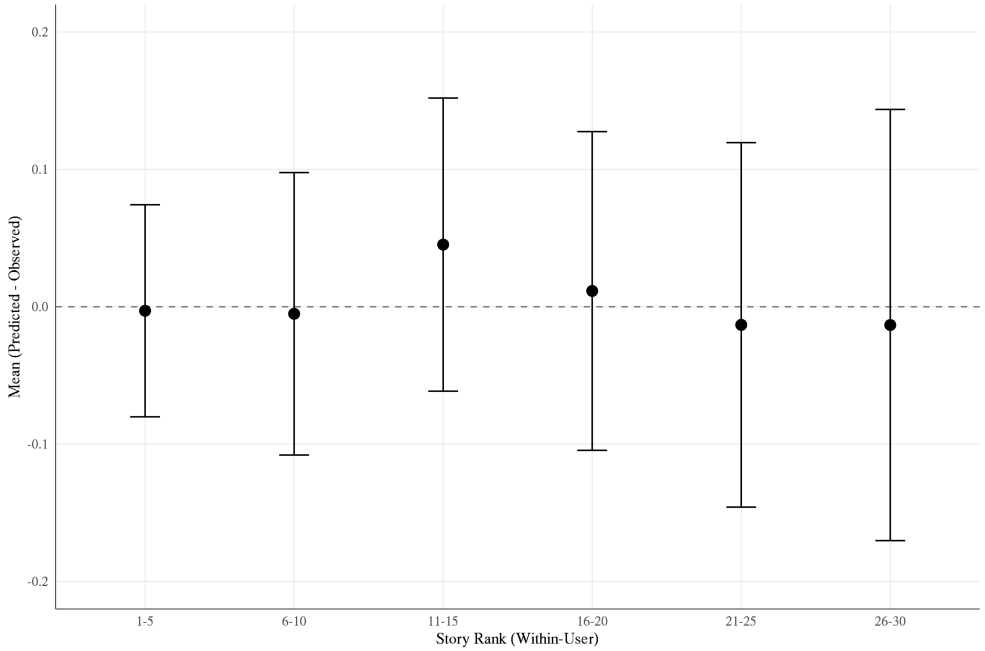


Note: Each point represents the mean predicted and mean observed engagement for stories in that rank bin, aggregated across users. Error bars show 95% confidence intervals. The dashed line at zero indicates perfect calibration. The time period corresponds to the RCT period between July 22, 2021 and August 4, 2021.

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 aggregate across users. Figure A8 plots both engagement metrics, for each rank bin, with 95% confidence intervals based on the variance of observed engagement.

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%

Figure A9: 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 time period corresponds to the RCT period between July 22, 2021 and August 4, 2021.

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 the sample changes as we move between ranks. The higher the rank, the fewer users were exposed to the story, with stories ranked 20 and higher having exposures only from heavy users. One way to see this in the plot is that predicted engagement for stories in ranks 11-15 is higher than for stories in ranks 6-10. This is because users exposed to stories in ranks 11-15 have, on average, higher predicted engagement than users who interacted with stories in 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 calibration errors are similar across the considered ranks, and we find no evidence that we systematically underestimate engagement at lower-ranked stories.