Exploring Revealed Behavior and Stated Profiles in Longitudinal News Recommendation

Thomas Elmar Kolb*
CDL-RecSys, TU Wien
Vienna, Austria
thomas.kolb@tuwien.ac.at

Alain Dominique Starke
MediaFutures, University of Bergen;
Amsterdam School of Communication Research, University of Amsterdam
Amsterdam, Netherlands
a.d.starke@uva.nl

Christoph Trattner
MediaFutures, University of Bergen
Bergen, Norway
christoph.trattner@uib.no

Abstract

News recommenders increasingly need to balance relevance, diversity, and stated preferences over repeated interactions. We report preliminary evidence from a four-wave user study with 262 participants comparing four personalization trajectories in a content-based news feed. Phase-wise comparisons suggest that on-profile selection varied with the assigned trajectories, while stated topic profiles remained comparatively stable during the two-week study (92.6%–98.0% retention). These preliminary findings suggest that short-term selection behavior partly reflects feed composition rather than stable preference change. Ongoing mixed-effects, exposure-normalized, and attrition-sensitive analyses will test this interpretation.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Recommender Systems, Longitudinal Evaluation, User Study, News

ACM Reference Format:
Thomas Elmar Kolb, Alain Dominique Starke, and Christoph Trattner. 2026. Exploring Revealed Behavior and Stated Profiles in Longitudinal News Recommendation. In *20th ACM Conference on Recommender Systems (RecSys '26), September 27–October 02, 2026, Minneapolis, MN, USA*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3773078.3841262

1 Introduction & Related Work

News recommenders must balance exploiting known user interests with exposing readers to diverse and timely content. Similarity-based approaches remain central to recommender systems, where items closely aligned with a user model are commonly treated as exploiting known preferences, while less similar or more diverse items support exploration [4, 5, 11]. This trade-off is particularly pronounced in news recommendation, as articles have a short shelf life while news feeds must account for recency, topic change, and breaking-news dynamics [6, 9, 10]. Recent work has examined user perceptions of recommender objectives, controlled personalization in operational news services, and user-centered recommender evaluation more broadly [1, 3, 8]. However, controlled longitudinal evidence remains limited. Many evaluations are offline or single-session, and recent longitudinal work on similarity-based news recommendation leaves open how changing exposure trajectories affect revealed behavior and stated profiles over time [7]. We address this gap and present our preliminary results in this Research & Practice note with a four-wave user study (N = 262) comparing four content-based news-feed trajectories: *consistently similar (SIM)*, *consistently dissimilar (DIS)*, *increasing personalization (INC)*, and *decreasing personalization (DEC)*. Against this background, we ask how different personalization trajectories affect revealed reading behavior, perceived recommendation experience, and explicitly stated topic-profile stability. We provide a longitudinal study design, early indications of divergence between revealed article selections and stated profiles, and an analysis agenda accounting for repeated measures, exposure, attrition, and user- and topic-level differences.

*This work was conducted as part of an ERASMUS+ traineeship hosted by SFI MediaFutures.

This work is licensed under a Creative Commons Attribution 4.0 International License.
RecSys '26, Minneapolis, MN, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2284-4/26/09
https://doi.org/10.1145/3773078.3841262

2 Study Design and Measures

We conducted a longitudinal between-subjects user study with one calibration wave (Wave 0) and three treatment waves (Waves 1–3). Participants were randomly assigned using count-balanced allocation to SIM, DIS, INC, or DEC. An a priori power analysis in G*Power for a one-way ANOVA with four groups (f = .25, α = .05, $1 - \beta$ = .80) required N = 180. At baseline, we recruited 419 U.S.-based Prolific participants; 262 completed all four waves and formed the preliminary complete-case sample (SIM n = 73, DIS n = 61, INC n = 59, DEC n = 69). Participants were compensated for each completed study phase. Before recruitment, the study's research ethics and data-protection requirements were assessed in accordance with the University of Bergen's institutional procedures. No separate ethics committee approval was required. All participants provided informed consent, and only anonymized data were analyzed. Eligibility required English as a primary language and completion on a desktop or laptop device. The final analytic sample was approximately gender-balanced (51.1% male, 47.7% female, 1.1% undisclosed), spanned adult age groups (largest group: 35–44 years, 30.9%), and was educationally varied (46.6% BA, 34.0% high school, 13.0% MSc, 5.7% doctorate, and 0.8% other or unreported). During calibration, participants selected topic preferences and liked articles they wanted to read; in each treatment wave, they interacted with a content-based news-feed interface (Fig. 1) and then completed repeated survey measures. The recommender followed the similarity-based setup of [7]. Articles were represented with OpenAI text embeddings (*text-embedding-3-small*), and for each participant we constructed a user profile by averaging the embeddings of articles they had liked in previous waves. Similar on-profile candidates were retrieved with FAISS [2] nearest-neighbor search and filtered to topics in the participant's stated profile and then re-ranked with a freshness-aware score that linearly combines similarity and recency (0.6 · FAISS similarity + 0.4 · normalized 1/(1+age_hours)), yielding the final on-profile list; off-profile candidates were sampled from topics outside that profile. Each wave used a fresh NewsCatcher article pool restricted to U.S. outlets covered in the 2025 Reuters Institute Digital News Report[1]. In the calibration wave, before prior article selections were available, recommendations were based on participants' stated topic preferences. The treatment feeds then combined on-profile and off-profile articles according to four personalization trajectories: *SIM* (90–90–90% on-profile), *DIS* (10–10–10%), *INC* (10–50–90%), and *DEC* (90–50–10%). We measured **revealed behavior** from submitted article selections made using the heart/like control. Measures included on- and off-profile selections, total selections, and the on-profile selection ratio. Profile retention was measured at the participant level as $|P_0 \cap P_3|/|P_0|$, where P_0 and P_3 denote the baseline and final topic sets. **Subjective measures** captured *choice satisfaction*, *perceived recommendation quality*, and *perceived diversity* on five-point Likert scales. We examined the item structure using a preliminary confirmatory factor analysis (CFA) of the pooled treatment-wave responses. Two low-loading items were omitted before scale construction; the retained items loaded strongly on their intended constructs (Tab. 2), and all retained constructs showed ordinal alpha values > .80. We report phase-wise ANOVAs and descriptive profile comparisons as **first analyses**; mixed-effects, exposure-normalized, and attrition-sensitive analyses are ongoing.

3 Preliminary Findings

Personalization manipulation. On-profile selection varied with the assigned trajectories (Tab. 1; Fig. 2). Condition differences were significant in all treatment waves, with the strongest separation in Wave 3, $F(3, 258) = 87.77$, $p < .001$, $\eta^2 = .505$. Despite a 10% feed target, DIS retained substantial on-profile selection (48.9%–35.6%), motivating exposure-normalized analyses. Total selection volume did not differ systematically between conditions. **Subjective perceptions** changed less consistently than selection behavior (Fig. 2). Perceived diversity differed by condition in the first and final treatment waves: $F(3, 258) = 5.04$, $p = .002$, $\eta^2 = .055$, and $F(3, 258) = 6.34$, $p < .001$, $\eta^2 = .069$, but not in the middle treatment wave. Perceived recommendation quality differed across all treatment waves, with the largest effect at the start of treatment: $F(3, 258) = 5.97$, $p < .001$, $\eta^2 = .065$. Choice satisfaction showed only an early condition effect: $F(3, 258) = 3.07$, $p = .028$, $\eta^2 = .034$, and did not differ significantly in later waves. Thus, condition differences in selection behavior appeared more persistent than differences in subjective evaluations. **Explicit topic profiles** changed little (Tab. 1): participants retained 92.6%–98.0% of their baseline topic selections. DEC showed the most movement (92.6% retention; 29.0% changed at least one topic). Thus, article selection varied with feed composition while stated profiles remained comparatively stable over two weeks.

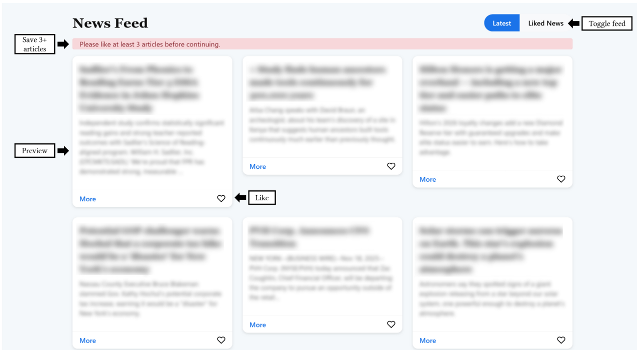
4 Initial Interpretation and Next Steps

These preliminary analyses suggest that article selection may respond more quickly to feed composition than stated profiles. On-profile selection varied with the assigned feed-composition trajectory, while stated topic profiles remained comparatively stable. For news recommender evaluation, this revealed-versus-stated distinction matters: short-term selection shifts should not be interpreted too directly as stable preference change, especially when exposure itself is experimentally or algorithmically constrained. Subjective benefits were more limited: phase-wise comparisons suggested higher perceived recommendation quality and early choice satisfaction under stronger personalization, but satisfaction differences did not persist across later waves. These findings provide initial evidence from a controlled longitudinal study while also defining clear boundary conditions for interpretation. Future work can examine whether the observed divergence between revealed selection behavior and stated topic profiles extends to other countries, mobile news use, and more naturalistic consumption contexts. The article pool was intentionally drawn from selected U.S. outlets and topic categories to support experimental control; as a result, topic availability varied across categories, with sports representing a comparatively large share. These design choices may have influenced both profile construction and perceived diversity. The study covered four waves over a two-week period, making the results most informative about short-term behavioral adaptation. Longer deployments would be valuable for examining sustained preference development, repeated exposure effects, and possible delayed changes in stated profiles. In addition, planned on-profile rates should be understood as feed-composition targets rather than deterministic exposure or selection outcomes: participants could ignore articles, feed lengths varied in practice, and scroll depth likely differed across users. Future analyses can build on these findings by modeling individual differences, user segments, topic-specific trajectories, and finer-grained exposure measures that distinguish articles delivered, articles seen, and articles selected. Next, we will apply mixed-effects trajectory models, exposure-normalized measures distinguishing articles delivered, seen, and selected, attrition-sensitive analyses, user- and topic-level models, and a longitudinal structural equation model linking feed composition, perceived diversity, satisfaction, and downstream informedness. These analyses will test the robustness, heterogeneity, and mechanisms of the preliminary patterns.

Acknowledgments

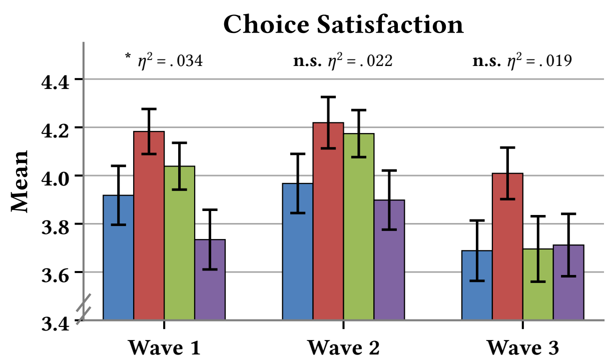
This work was supported by the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centre for Research-based Innovation scheme, project number 309339. The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

[1]https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-06/Digital_News-Report_2025.pdf

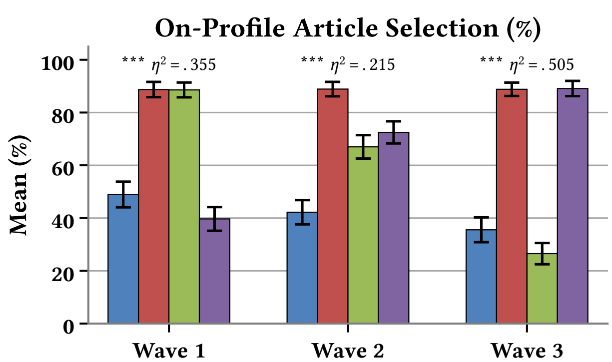


Note. Participants could preview and like articles and switch between the latest and liked-article feeds, making exposure and explicit article choices observable.

Figure 1: Experimental news-feed interface used during the treatment waves.



(a) Choice Satisfaction



(b) On-Profile Article Selection (%)

Note. Subjective and behavioral responses across the three treatment waves. Bars show condition means; error bars indicate ±1 standard error. The upper panel shows that choice satisfaction varies only modestly and mainly early in the study. The lower panel visualizes a preliminary aggregate pattern: on-profile selection varies with the assigned feed-composition trajectories. Annotations report phase-wise ANOVAs ($*p < .05$, $***p < .001$; n.s. denotes $p \geq .05$). η^2 denotes the corresponding effect size; mixed-effects and exposure-normalized analyses are ongoing.

Figure 2: Choice satisfaction and on-profile article selection across treatment waves.

Table 1: Planned personalization trajectories and preliminary observed outcomes.

Cond.	Plan	Selection ratio	Ret.	Chg.
SIM	90–90–90	.887–.889–.888	96.9%	23.3%
DIS	10–10–10	.489–.422–.356	95.9%	24.6%
INC	10–50–90	.397–.725–.891	98.0%	22.0%
DEC	90–50–10	.886–.670–.265	92.6%	29.0%

Note. Plan and selection ratio report treatment Waves 1–3. Ret. is the mean participant-level $|P_0 \cap P_3|/|P_0|$. Chg. is the proportion whose final topic set differed from baseline; additions and removals both count.

Table 2: Preliminary measurement checks for the repeated user-experience measures. Standardized loadings are reported for retained items; two weak-loading items from the initial pool are shown for transparency.

Construct and item	Loading
Choice Satisfaction *(adapted from Knijnenburg and Willemsen [8])*	
I like the articles I have chosen.	.943
I would like to read more articles like the one I have chosen.	.912
The chosen articles fit my general preferences.	.899
Omitted: I would know several articles that are better than the ones I have chosen.	–
Perceived Recommendation Quality *(adapted from Knijnenburg and Willemsen [8])*	
I found the articles in the news feed to be interesting.	.919
The articles in the news feed fitted my preferences.	.956
The articles in the news feed were relevant to me.	.933
Perceived Diversity *(adapted from Kasangu et al. [7])*	
The articles in the news feed differed in terms of their topics.	.832
The diversity of the articles in the news feed was high.	.969
Omitted: The articles in the news feed were similar to each other.	–

Note. The table reports retained loadings from the reduced measurement model. Italicized items show the initial item pool but were omitted due to weak loadings in the initial CFA: one choice satisfaction item ($\lambda = .315$) and one reverse-coded perceived diversity item ($\lambda = .507$). All retained constructs showed ordinal alphas above .80.

References

[1] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2025. How Do Users Perceive Recommender Systems' Objectives?. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 165–176. doi:10.1145/3705328.3748066

[2] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[3] Marlene Holzleitner, Stephan Leitner, Hanna Lind Jorgensen, Christoph Schmitz, Jacob Welander, and Dietmar Jannach. 2026. Controlled Personalization in Legacy Media Online Services: A Case Study in News Recommendation. *ACM Trans. Recomm. Syst.* 4, 4, Article 62 (July 2026), 32 pages. doi:10.1145/3814614

[4] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender systems: an introduction.* Cambridge university press.

[5] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.

[6] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems – Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227. doi:10.1016/j.ipm.2018.04.008

[7] Gloria Anne Babile Kasangu, Alain D. Starke, and Christoph Trattner. 2025. More of the Same? A Longitudinal Evaluation of Two Similarity-Based Approaches in a News Recommender System. In *Proceedings of the 13th International Workshop on News Recommendation and Analytics (INRA 2025) (CEUR Workshop Proceedings, Vol. 4056).* https://ceur-ws.org/Vol-4056/full1.pdf

[8] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. *Evaluating Recommender Systems with User Experiments.* Springer US, Boston, MA, 309–352. doi:10.1007/978-1-4899-7637-6_9

[9] Andreas Lommatzsch, Benjamin Kille, and Sahin Albayrak. 2017. Incorporating context and trends in news recommender systems. In *Proceedings of the international conference on web intelligence.* 1062–1068.

[10] Shaina Raza and Chen Ding. 2022. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review* 55, 1 (2022), 749–800.

[11] Neil Rubens, Dain Kaplan, and Masashi Sugiyama. 2010. Active learning in recommender systems. In *Recommender systems handbook.* Springer, 735–767.