More of the Same? A Longitudinal Evaluation of Two Similarity-based Approaches in a News Recommender System

Gloria A.B. Kasangu[1], Alain D. Starke[1,2,*] and Christoph Trattner[1]

[1]MediaFutures, University of Bergen, Vestland, Norway
[2]ASCoR, University of Amsterdam, Netherlands

Abstract

Similarity-based personalization is generally assumed to boost engagement in recommender systems. However, is this also true beyond a single session in a news recommender? Amid concerns about filter bubbles and preference volatility, we propose an empirical evaluation of both short-term and longer-term effects of a news recommender system with two phases of data collection: Initial preference elicitation and evaluation (Phase 1), a 48-hour interval, and a personalized follow-up (Phase 2). We compared two recommendation strategies in a preliminary longitudinal experiment ($N = 166$): An **'Aligned'** feed that included articles that met a $\geq 70\%$ cosine-similarity threshold, and a **'Disaligned'** feed with only a 30% similarity threshold. We collected behavioral metrics (article clicks, time on feed) and evaluative metrics (self-reported familiarity, perceived recommendation quality, choice satisfaction, topic preferences) in both phases. The Aligned feed was perceived to have more familiar content, while perceived diversity did not differ between recommendation strategies. Users clicked on significantly fewer articles in Phase 2, particularly in the Disaligned condition. We also explored the volatility of topic preferences, but did not observe significant differences across phases. These findings suggest that short-term increases in feed–profile similarity can enhance familiarity and maintain behavioral engagement (i.e., clicks). In contrast, they do not lead to higher levels of perceived quality and choice satisfaction, which raises doubts about the relationship between the similarity of preference-based articles and user satisfaction.

Keywords

News, Recommender Systems, Similarity, News Aggregator, Longitudinal Evaluation

1. Introduction

A large number of news platforms rely on recommender systems to provide digital news [1]. This has fundamentally reshaped the way audiences consume information [2]. By tailoring content based on individual preferences and past behavior, news recommenders aim to enhance user engagement and satisfaction, yet the dominance of similarity-based personalization raises unresolved questions about its implications for both individual users and the broader information ecosystem. While short-term evaluations of recommender systems are common, longitudinal assessments remain rare in the news domain [3].

In this preliminary study, we conduct a field experiment at two time points to compare two recommendation strategies that differ in their degree of personalization. Our goal is not to settle the long-term debate but to offer an initial look at how user satisfaction and engagement evolve when exposed to higher versus lower content similarity over a short interval. We compare two conditions: One condition ("Misaligned") in which users receive more generic than personalized content, and another condition ("Aligned") in which users receive more personalized content than generic content. For the latter, users are presented only articles with at least 70% similar (by cosine similarity) to their past click history, reflecting a typical personalization threshold.

Proceedings of the 13th International Workshop on News Recommendation and Analytics (INRA 2025), co-located with the 19th ACM Conference on Recommender Systems (RecSys 2025), September 22–26, 2025, Prague, Czech Republic.
*Corresponding author.
gloria.kasangu@student.uib.no (G. A.B. Kasangu); alain.starke@uib.no (A. D. Starke); christoph.trattner@uib.no (C. Trattner)
0009-0000-8585-3966 (G. A.B. Kasangu); 0000-0002-9873-8016 (A. D. Starke); 0000-0002-1193-0508 (C. Trattner)

This design is motivated by concerns about filter bubbles and echo chambers [4, 5], in which highly personalized consumption can reinforce ideological entrapment and reduce exposure to diverse viewpoints [6]. Although some work suggests that personalization alone does not always produce these effects [7], the larger impact on public discourse continues to be debated [2]. At the same time, recommender systems face the challenge of preference volatility, as user interests change over time and algorithms often struggle to detect or adapt to these changes [3].

By examining user behavior and perceptions in two phases, our study provides an exploratory window into (1) whether short-term preference shifts occur under different personalization regimes and (2) how alignment with past behavior influences satisfaction with chosen articles. We address the following research questions:

- **RQ1:** To what extent does presenting more aligned news recommender content (i.e., based on user-item similarity) positively affect choice satisfaction over time?
- **RQ2:** To what extent does user-item similarity affect a user's perceived recommendation quality and clicking behavior in a news recommender system?

To answer these questions, we collected behavioral measures (article clicks and self-reported percent familiarity as a proxy for cosine similarity) alongside subjective ratings of choice satisfaction and perceived quality. Although our two-timepoint design does not capture long-term dynamics, it offers a critical first step toward understanding how brief exposures to different levels of personalization shape the user's experience.

Our contributions are threefold: (1) we provide early evidence on how short-term exposure to high versus low similarity news feeds affects satisfaction and engagement; (2) we demonstrate the utility of self-reported familiarity as a practical manipulation check; and (3) we highlight directions for future longitudinal work on adaptive recommendation strategies that balance personalization with informational diversity.

2. Related Work

We discuss literature in the context of news similarity and diversity and its related effects. For example, content-based approaches that strongly optimize for similarity may lead to 'more of the same' content that is less diverse [1, 8]. Therefore, we will discuss filter bubbles and echo chamber effects in the context of news recommenders, as well as research that included longitudinal evaluation components.

2.1. Filter Bubbles and Echo Chamber Effects in Recommender Systems

A lack of recommended diversity over a longer time period can be described as a filter bubble. Although definitions vary [9], filter bubbles and echo chambers can be defined as the tendency of personalization algorithms to enclose users within a narrow band of similar content, potentially compromising exposure to diverse viewpoints. Pariser's influential work introduced the term filter bubble to warn that algorithmic curation can invisibly tailor information streams around a user's past behavior, reinforcing existing beliefs rather than challenging them [10]. Subsequent empirical studies have confirmed that personalization can increase ideological segregation. Flaxman et al. [5] demonstrated that search result personalization led users toward more politically extreme news sources compared to non-personalized search. Nguyen et al. [6] showed that collaborative filtering methods tend to prioritize popular or similar items at the expense of topic diversity, thereby fostering echo chamber effects. In the news domain, such narrowing raises grave concerns for democratic discourse, since access to a plurality of perspectives is essential [2]. Recent scholarship calls for critical reflection on how recommender design choices, including similarity thresholds, feedback loops and diversification strategies, can exacerbate or mitigate enclosure effects. Researchers also urge longitudinal studies to assess the real-world impacts of these design decisions over time [4, 3].

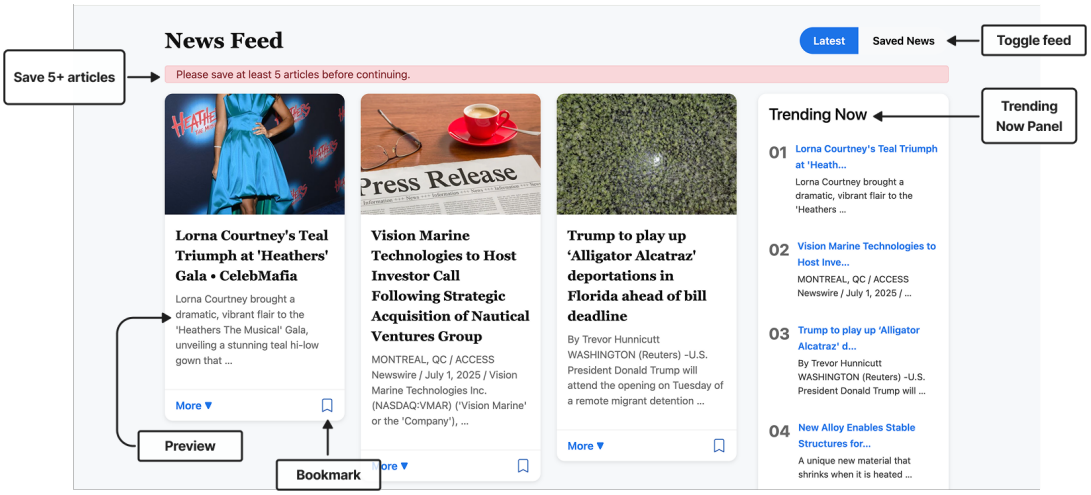

Figure 1: Participant-facing news-feed interface showing the uniform grid, bookmark icon, preview expansion, and "Save 5+" banner.

2.2. News Personalization and Engagement

Personalized news recommender systems aim to increase user engagement by tailoring content presentation to individual users, often leveraging stored user attributes and past user behavior [11]. Previous studies have shown that personalization can significantly improve short-term engagement metrics such as click-through rates and next-item prediction accuracy [12, 13]. By reducing information overload and presenting content that aligns with user preferences, personalized systems often improve perceived relevance and user satisfaction [14].

Nonetheless, the relationship between personalization and long-term engagement is more complicated. Previous research has shown that overly narrow recommendation strategies can lead to a decrease in information diversity [6, 15], as well as user fatigue [16]. Moreover, high similarity between recommended and previously consumed content can boost engagement in the short term, while limiting average individual exposure to diverse viewpoints [17]. In the context of news, the trade-off between personalization and exposure diversity is especially important as democratic deliberation relies on access to varied perspectives [2]. Some studies suggest that moderate personalization can maintain diversification while mitigating negative effects like the filter bubble. For instance, Gao et al [18] found that moderate diversification retains recommendation accuracy while promoting exposure to a more varied set of topics. Additionally, others have proposed approaches that combine personalization with editorial or novel content to maintain reader interest without reinforcing filter bubble effects [19, 20].

2.3. Longitudinal Experiments in Recommender Systems

Recent work has begun to explore the long-term effects of recommender systems through longitudinal field experiments. The use of simulation-based methods, such as agent-based modelling, has been a staple research methodology in many fields such as sociology and managerial science [21, 22]. Recently, Zhang et al. [23] introduced an agent-based simulation framework to analyze the longitudinal effects of recommender systems. Their findings revealed a phenomenon called the performance paradox, in which user interaction with recommendation algorithms can paradoxically degrade overall system performance over time. Their findings emphasize the risk of overpersonalization leading to decreased user satisfaction, a finding that was also explored in a follow-up work using extended modeling techniques. Similarly, Ferraro et al. [24], using a simulation-based framework tailored to session-based recommender systems, found that repeated interactions can reinforce popularity bias and reduce item diversity over time.

To empirically validate the long-term effects of recommender systems on content exposure, some longitudinal field experiments have been conducted. For example, Lee and Hosanger [25] ran a random-

ized field experiment in multiple product categories and demonstrated that personalized recommender systems led to a decrease in overall sales diversity, particularly when using collaborative filtering methods. Furthermore, Fleder and Hosanger [26] found that recommendations, in the long term, can lead to a concentrated consumption of a small group of popular items. As a response to these observed concentration effects, various studies have proposed mitigation strategies such as hybrid recommendation models that blend personalization with popularity-neutral signals [23] and ranking-based diversification techniques [24, 27].

Based on these findings, our study presents a longitudinal user experiment in the domain of news recommendation. Unlike previous work that focuses primarily on e-commerce, we examine how varying degrees of personalization affect user satisfaction, engagement, and content diversity over time. By collecting both behavioral and subjective data at two time points, our work provides empirical insights into how users respond to different recommendation strategies and how recommender systems might better respond to evolving user preferences.

3. Methods

3.1. Research Design

This study employed a between-subjects longitudinal experimental design with two waves of data collection, separated by a mandatory 48 h waiting period. The two experimental conditions differed in the personalization strategy used to generate news recommendations in the second phase of the study. Participants were randomly assigned to one of two conditions:

Condition 1: Alignment. Participants received a feed of articles that were mostly topically aligned with their established preferences.

Condition 2: Disalignment. Participants received a feed of articles that were mostly dissimilar to their preferences, encouraging exploration of new topics.

Outcome measures, including self-reported satisfaction, perceived quality, and behavioral click data were collected at both timepoints to evaluate the effect of each recommendation strategy.

3.2. Participants

Two hundred English-fluent adults (95–100% approval rate) were recruited via Prolific and randomly assigned to Alignment or Disalignment ($n = 100$ each). As 34 participants dropped out between both phases (i.e., attrition), a sample of $N = 166$ participants remained for analysis (Alignment: $n = 81$; Disalignment: $n = 85$). Participant ages ranged from 18 to 65 ($M = 34.8$, $SD = 8.9$). Of the participants, 46.4% identified as female, 52.4% as male, and 1.2% declined to specify (see Table 1). Participants were paid £9.00/hr for Phase 1 and £16.00/hr for Phase 2.

Table 1
Participant demographics (N=166).

	N	%	M (SD)
Age (years)	–	–	34.8 (8.9)
Female	77	46.4	–
Male	87	52.4	–
Undisclosed	2	1.2	–

3.3. Materials & Algorithms

We sourced news articles in real time via the NewsCatcher API[1] , restricted to 15 reputable English-language outlets.[2]

The NewsCatcher API provides real-time access to news articles from a wide range of publishers. We configured it to return JSON-formatted metadata (headline, summary, publication date, and URL) along with thumbnail images when available. Queries were limited to English-language content and filtered to include only articles published within the past 24 hours, ensuring both recency and relevance. Articles were displayed in a uniform grid (title, image, short description). We removed all publisher logos and other branding elements to control for any presentation-based biases. Participants browsed this feed via our web interface (Figure 1), which supported bookmarking, article previews, and enforced a minimum of five saves before proceeding.

In Phase 2, a content-based filtering algorithm generated each user's personalized feed. We built a TF–IDF profile vector from that user's Phase 1 bookmarks, then computed each candidate article's final relevance score:

$$\text{score} = 0.60 \times \cos\big(\text{TF-IDF}_{\text{user}},\ \text{TF-IDF}_{\text{article}}\big) + 0.40 \times \text{freshness\_bonus}(\text{publication\_date}).$$

Articles with score ≥ 0.5 were labeled *familiar*, and those with score ≤ 0.4 *novel*. Under the Alignment condition, feeds comprised 70% familiar and 30% novel articles; under Disalignment, this ratio was reversed (30% familiar, 70% novel).

3.4. Procedure

The study consisted of two phases, separated by a 48 h interval. The study procedure is shown in Figure 2.

Phase 1: Preference Elicitation & Baseline. After completing an informed consent form and a survey on the participant's demographics and media habits, participants indicated "like"/"dislike" for 12 news topics. An initial feed of up to 30 de-duplicated articles was generated and a persistent banner ("Please save at least five articles before continuing") enforced a minimum of five bookmarks before they could advance. These bookmarks formed their profile, and they then completed the evaluation survey as a baseline.

Phase 2: Recommendation & Evaluation. Participants were invited back exactly 48 hours after Phase 1 to again complete the topic-preferences survey. A new pool of 40 candidate articles was fetched; the algorithm scored and selected items per condition to form the Phase 2 feed. After interacting with this feed, they completed the evaluation survey again, concluding their participation.

3.5. Measures

We operationalized a set of behavioral and subjective measures to evaluate user interactions with the news recommendation system.

3.5.1. Behavioral Measures

We logged three objective metrics at each phase:

[1]Reuters, Associated Press, BBC, The New York Times, The Wall Street Journal, The Washington Post, NPR, PBS, The Guardian, The Times (UK), Financial Times, The Independent, Al Jazeera, The Economist, CBS News.

[2]Prototype code and data are available at https://anonymous.4open.science/r/news-diversification-study-disalignment-1722/ and https://anonymous.4open.science/r/news-diversification-study-45FB/.

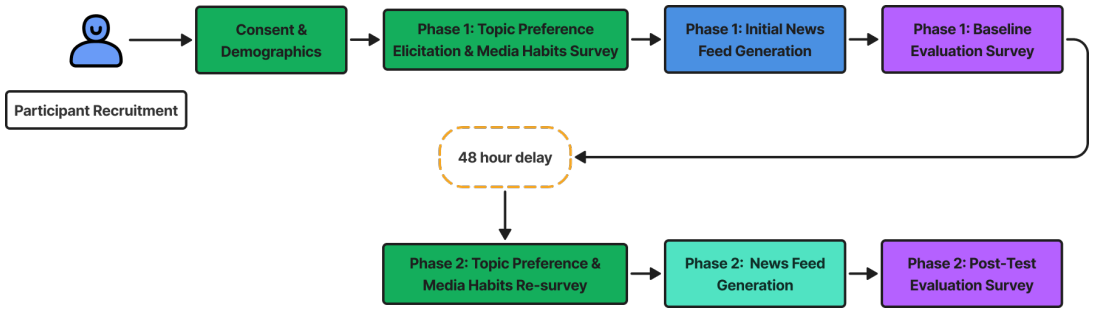

Figure 2: Flowchart of the two-phase experiment procedure: Phase 1 (preference elicitation, feed interaction, baseline survey), 48 h delay, Phase 2 (preference re-survey, feed interaction, post-test survey).

Article Clicks The total number of recommended articles a participant clicked. This metric serves as a direct indicator of engagement: more clicks suggest greater interest in the feed content, whereas fewer clicks may reflect disengagement or irrelevance of the recommendations.

Article Similarity The mean cosine similarity between each recommended article and the set of articles clicked in the previous phase. By quantifying how closely new recommendations match past behavior, this measure operationalizes the degree of "alignment" versus "disalignment" in the feed and allows us to link algorithmic similarity to downstream outcomes.

Total Time on Feed The total time in seconds spent viewing the news feed. Total dwell time captures sustained attention beyond the point of click, reflecting the extent to which participants explored the feed and consumed article content even when they did not click through.

Percent Familiarity The self-reported percentage of recommended articles with which participants felt they were already familiar. Although subjective, this rating serves as a practical proxy for the underlying cosine-similarity between new recommendations and each participant's prior click history. Higher values indicate stronger alignment between the feed and past interests.

3.5.2. Subjective Measures

All subjective items were rated on a 5-point Likert scale (1 = *Strongly disagree*, 5 = *Strongly agree*). We assessed three constructs:

Choice Satisfaction Adapted from Knijnenburg et al. [28], this construct comprised two positively-phrased statements: "I like the articles I've chosen" and "I was/am looking forward to reading the chosen articles." Responses to these items were averaged to form a single Choice Satisfaction score. An exploratory factor analysis (principal-axis factoring with varimax rotation) on Phase 1 responses supported a clean two-factor solution, with the two satisfaction items loading strongly on one factor ($\lambda = 0.94$ and $\lambda = 0.62$) that accounted for 38 % of the variance. Cronbach's $\alpha = 0.87$ indicated good internal consistency.

Perceived Recommendation Quality Based on the advice-solicitation scale of Starke et al. [29], we used three items: "I found the recommended articles to be interesting," "The recommended articles fitted my preferences," and "The recommended articles were relevant to me." Responses were averaged into a single perceived quality score. An exploratory factor analysis (principal-axis factoring with varimax rotation) on Phase 1 data showed that all three items loaded strongly on one factor ($\lambda = 0.74$ for "interesting," $\lambda = 0.88$ for "relevant," $\lambda = 0.51$ for "fit preferences"), which accounted for 40% of the variance. Cronbach's $\alpha = 0.87$ for this Quality factor, indicating good internal consistency.

Perceived Diversity To assess recommendation variety, we developed a three-item scale: "The recommended articles were similar to each other" (reverse-scored), "The recommended articles differed in terms of their topics," and "The diversity in the recommended list of articles was high." An exploratory factor analysis (principal-axis factoring, no rotation) on Phase 1 responses revealed that the "similar to each other" item loaded weakly and negatively ($\lambda = -0.13$), whereas the "differed in topics" and "diversity high" items loaded strongly ($\lambda = 0.96$ and $\lambda = 0.58$, respectively). Initial internal consistency across all three items was poor (Cronbach's $\alpha = 0.21$). After reverse-scoring and removing the "similar to each other" item, reliability for the remaining two items improved to Cronbach's $\alpha = 0.72$, supporting the use of this two-item composite diversity score in subsequent analyses.

4. Results

We confirmed internal consistency with Cronbach's α (quality: $\alpha = .87$; satisfaction: $\alpha = .87$) and assessed temporal stability via intraclass correlations over the 48 h interval (quality: $ICC_{1,2} = .54$; satisfaction: $ICC_{1,2} = .43$). A manipulation check on self-reported familiarity in Phase 2 showed that Alignment participants reported greater familiarity ($M = 71.5\%$, $SD = 29.6$) than Disalignment participants ($M = 36.5\%$, $SD = 34.2$), $t(162.5) = 7.05$, $p < .001$. Behavioral outcomes were evaluated with a repeated-measures ANOVA on total article clicks. Changes in topic preferences between phases were examined using paired-sample t-tests. Finally, to address RQ1, we compared Phase 2 Choice Satisfaction across conditions with an independent-samples t-test.

4.1. Manipulation Check

To verify that our feed alignment manipulation had its intended effect on subjective familiarity without inadvertently altering perceived diversity, we ran two Welch's t-tests on Phase 2 self-reports. First, participants in the Alignment condition reported substantially higher percentage of familiar articles (M = 71.46%, SD = 29.6) than those in the Disalignment condition (M = 36.46%, SD = 34.2): $t(162.5) = 7.05$, $p < .001$; see also Figure 3. This suggested that users facing more similar articles indeed reported they were more familiar. Second, perceived diversity did not differ between the Alignment (M = 3.80, SD = 1.07) and Disalignment (M = 3.82, SD = 1.03) conditions: $t(162.8) = -0.13$, $p = .897$. This showed that while familiarity corresponded to the research design changes in similarity, it was perceived by users in terms of the diversity in the presented content.

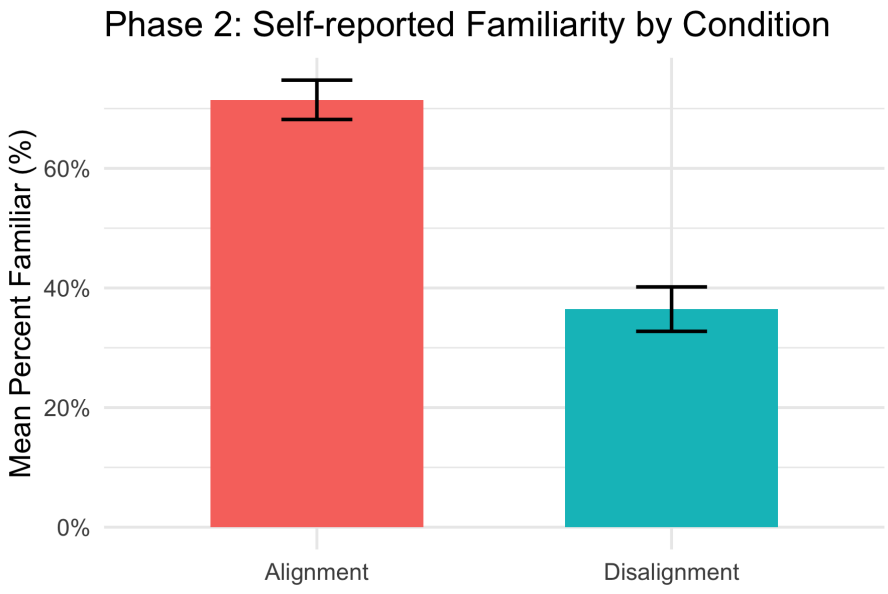

Figure 3: Self-reported familiarity in Phase 2 by condition. Error bars show ±1 S.E. of the mean.

4.2. Behavioral Outcomes

Table 2 reports the mean (and SD) number of article clicks by condition and phase. A 2 (Condition: Alignment vs. Disalignment) x 2 (Phase: 1 vs. 2) repeated-measures ANOVA on total clicks revealed no main effect of Condition, $F(1, 164) = 2.59$, $p = .109$, $\eta_p^2 = .011$, but a significant main effect of Phase, $F(1, 164) = 9.34$, $p = .003$, $\eta_p^2 = .018$, indicating an overall change in click behavior over time. The Condition × Phase interaction was not significant, $F(1, 164) = 4.08$, $p = .045$, $\eta_p^2 = .008$, suggesting similar click-patterns across groups. The time-course of clicks is visualized in Figure 4.

Table 2
Mean (SD) total article clicks by condition and phase.

Condition	Phase	*M*	*SD*
Alignment	Phase 1	5.85	1.41
	Phase 2	5.51	5.78
Disalignment	Phase 1	5.75	2.03
	Phase 2	4.06	4.20

Table 3
Repeated-measures ANOVA on the total number of article clicks by a user.

	df_{num}	df_{den}	F	p	η_p^2
Condition	1	164	2.59	.109	.011
Phase	1	164	9.34	.003	.018
Cond. × Phase	1	164	4.08	.045	.008

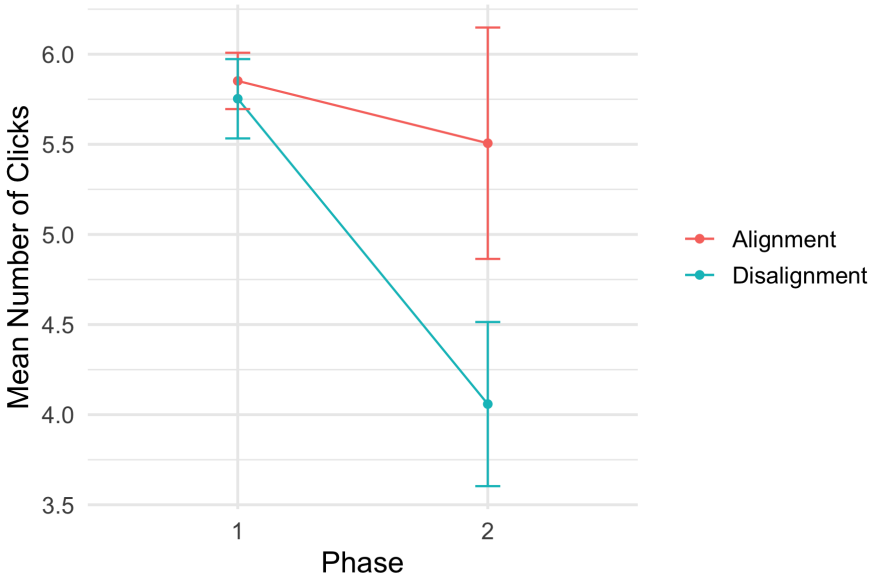

Figure 4: Mean total article clicks in Phase 1 and Phase 2 by recommendation strategy. Error bars denote ±1 SE.

4.3. Topic-Preference Shifts

To explore whether exposure to aligned versus Disaligned feeds induced any shifts in topical interests, we conducted paired-sample *t*-tests on the Phase 2 and Phase 1 difference scores for each of the 12 topics.

As shown in Table 4, all mean changes were small and not significant at the $\alpha = .05$ level. Figure 5 visualizes the distribution of these change scores (Phase 2 minus Phase 1).

Table 4
Mean changes in topic preferences (Phase 2–Phase 1) for each topic, using paired *t*-tests. All *p* values exceeded .05, indicating that topic preferences did not change significantly.

Topic	ΔM	t	df	p
Sport	−0.08	−1.61	165	.109
Politics	−0.05	−0.94	165	.347
Food & Drink	−0.06	−1.15	165	.253
Climate & Environment	−0.01	−0.20	165	.842
Lifestyle & Health	−0.04	−0.73	165	.469
Health & Research	−0.08	−1.22	165	.224
Society & Work	−0.08	−1.22	165	.224
Economy & Business	−0.05	−0.89	165	.373
Technology & Science	−0.04	−0.73	165	.469
Crime & Legal	−0.06	−0.93	165	.355
Entertainment & Celebrities	+0.02	+0.43	165	.671
International & Global Conflicts	−0.02	−0.34	165	.733

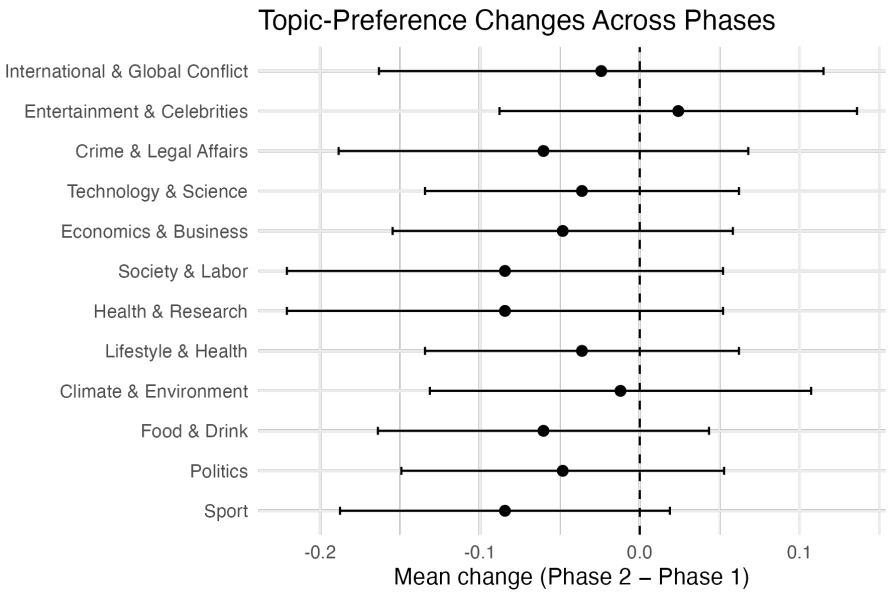

Figure 5: Mean change in topic preference (Phase 2 minus Phase 1) for each topic. Error bars denote ±1 SE; the horizontal dashed line indicates zero change.

4.4. RQ1: Choice Satisfaction by Strategy

An independent-samples *t*-test on Phase 2 Choice Satisfaction revealed no significant differences between Alignment ($M = 4.15$, $SD = 0.89$) and Disalignment ($M = 4.34$, $SD = 0.74$); $t(155.96) = -1.47$, $p = .144$. As shown in Figure 6, the distributions of satisfaction scores largely overlapped, indicating comparable levels of satisfaction between the recommendation strategies.

4.5. RQ2: Familiarity Effects

Next, we examined whether Phase 2 percent-familiarity predicted perceived recommendation quality and engagement (article clicks). As shown in Table 5, the percentage of perceived familiarity did not

Figure 6: Choice Satisfaction by recommendation strategy in Phase 2. Boxplots with overlaid jitter show individual scores, horizontal lines represent medians and boxes span the interquartile range.

significantly predict perceived quality: $b = -0.003$, $SE = 0.002$, $t = -1.65$, $p = .10$ ($R^2 = .016$). Likewise, the familiarity percentage neither predicted article clicks (cf. Table 6): $b = 0.011$, $SE = 0.011$, $t = 0.99$, $p = .32$ ($R^2 = .006$).

Table 5
Phase 2 Regression: Familiarity Predicting Perceived Quality

Parameter	b	SE	p
Intercept	4.198	0.126	$< .001$
Percent Familiar	−0.003	0.002	.100
R^2		.016	

Table 6
Phase 2 Regression: Familiarity Predicting Article Clicks

Parameter	b	SE	p
Intercept	4.193	0.700	$< .001$
Percent Familiar	0.011	0.011	.325
R^2		.006	

In general, both conditions encouraged robust engagement and satisfaction. Participants remained consistently satisfied with their choices, maintained stable topic interests throughout phases, and reported high familiarity without any adverse effects on perceived quality or clicking behavior.

5. Discussion

This study investigated how the algorithmic alignment of a news recommendation feed influences both subjective and behavioral user outcomes. Using a two-phase, between-subjects design, we manipulated whether recommended articles were more or less "aligned" with participants' elicited topic preferences. In addition, we measured (1) self-reported familiarity, (2) perceived recommendation

quality, (3) choice satisfaction, (4) the total number of article clicks, and (5) changes in topic preferences. Our manipulation successfully induced a higher percentage of familiarity in the Alignment condition, compared to the Disalignment condition. However, this has not led to differences in perceived quality, satisfaction, or engagement; Hence, we have not observed any statistically significant differences across conditions in terms of choice satisfaction. Moreover, nor has our repeated-measures ANOVA on click behavior revealed any interaction effects between the feed condition and time phase. Furthermore, exploring participants' topic interests, we have found that they remain stable over the 48h interval, while percent-familiarity did not predict either quality perceptions or clicking behavior.

Regarding RQ1, we hypothesized that presenting articles with high similarity would increase user satisfaction relative to a more diverse feed. Instead, satisfaction scores were statistically equivalent across the Alignment and Disalignment conditions. This suggests that once recommendations reach a certain relevance threshold, further increases in similarity do not lead to additional benefit, mirroring previous studies that reported diminishing personalization returns on engagement [18]. Users may value novelty or variety just as much as pure similarity, and a feed that is too narrowly focused may not improve, and might even reduce, perceived choice satisfaction in the long run.

As for RQ2, we find that neither regression reached significance, familiarity did not predict perceived quality or clicking behavior. This aligns with simulation studies suggesting that similarity alone cannot sustain engagement over time and that moderate diversification may be equally effective, or even necessary, to prevent user fatigue [24, 23]. Our results might imply that in a real-world news context, users do not simply click more or rate higher quality when they recognize content as familiar, which might underscore the need for hybrid strategies that balance relevance with serendipity.

6. Limitations and Future Work

Our study is subject to a few limitations. First, our study only considers two time points separated by 48 hours. This relatively short time frame constrains our ability to detect longer-term changes in topic interests or the cumulative effects of personalization. Future work should include additional follow-up waves over weeks or months to capture preference volatility and longer-term behavioral adaptation.

Second, we relied on self-reported familiarity and topic preference ratings, which are subject to recall biases. Incorporating objective measures, such as feed-level cosine similarity scores or diversity indices calculated on the full recommendation list, might strengthen the robustness of future findings.

Third, our participant pool was drawn from Prolific, a relatively inexpensive crowdsourcing platform. While it provides rapid data collection, it does not necessarily represent the full diversity of news consumers. The demographics and engagement patterns of Prolific users may differ from those of general audiences, potentially limiting external validity. Future studies should sample from multiple platforms and demographic strata.

Fourth, our behavioral engagement metrics were limited to article clicks and time on feed within the experimental interface. These metrics do not capture longer-term news consumption behaviors, such as sharing, commenting, or return visits. Expanding engagement measures to include social interactions might yield a more comprehensive picture of user response.

Finally, our diversity manipulation was implemented using a single "high diversity" indicator. This may not fully capture the multifaceted nature of content variety or ideological breadth. More sophisticated diversification strategies, such as topic-aware re-ranking or hybrid filtering, could produce different patterns of user response and merit exploration in future work.

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.

References

[1] M. Karimi, D. Jannach, M. Jugovac, News recommender systems–survey and roads ahead, Information Processing & Management 54 (2018) 1203–1227.

[2] N. Helberger, On the democratic role of news recommenders, in: Algorithms, automation, and news, Routledge, 2021, pp. 14–33.

[3] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, Artificial Intelligence Review (2022) 1–52.

[4] P. M. Dahlgren, A critical review of filter bubbles and a comparison with selective exposure., Nordicom Review 42 (2021).

[5] S. Flaxman, S. Goel, J. M. Rao, Filter bubbles, echo chambers, and online news consumption, Public opinion quarterly 80 (2016) 298–320.

[6] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, J. A. Konstan, Exploring the filter bubble: the effect of using recommender systems on content diversity, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 677–686.

[7] J. Möller, Filter bubbles and digital echo chambers 1, in: The routledge companion to media disinformation and populism, Routledge, 2021, pp. 92–100.

[8] D. Rosnes, A. D. Starke, C. Trattner, Shaping the future of content-based news recommenders: Insights from evaluating feature-specific similarity metrics, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 201–211.

[9] L. Michiels, J. Leysen, A. Smets, B. Goethals, What are filter bubbles really? a review of the conceptual and empirical work, in: Adjunct proceedings of the 30th ACM conference on user modeling, adaptation and personalization, 2022, pp. 274–279.

[10] E. Pariser, The filter bubble: What the Internet is hiding from you, penguin UK, 2011.

[11] L. Li, W. Chu, J. Langford, R. E. Schapire, A contextual-bandit approach to personalized news article recommendation, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 661–670.

[12] J. Liu, P. Dolan, E. R. Pedersen, Personalized news recommendation based on click behavior, in: Proceedings of the 15th international conference on Intelligent user interfaces, 2010, pp. 31–40.

[13] F. Garcin, C. Dimitrakakis, B. Faltings, Personalized news recommendation with context trees, in: Proceedings of the 7th ACM Conference on Recommender Systems, 2013, pp. 105–112.

[14] X. He, Q. Liu, S. Jung, The impact of recommendation system on user satisfaction: A moderated mediation approach, Journal of Theoretical and Applied Electronic Commerce Research 19 (2024) 448–466. URL: https://www.mdpi.com/0718-1876/19/1/24.

[15] N. Helberger, K. Karppinen, L. D'Acunto, Exposure diversity as a design principle for recommender systems, Information, Communication & Society 21 (2018) 191–207. doi:10.1080/1369118X.2016.1271900.

[16] H. Sagtani, M. G. Jhawar, A. Gupta, R. Mehrotra, Quantifying and leveraging user fatigue for interventions in recommender systems, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2293–2297. doi:10.1145/3539618.3592044.

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

[18] Z. Gao, T. Shen, Z. Mai, M. R. Bouadjenek, I. Waller, A. Anderson, R. Bodkin, S. Sanner, Mitigating the filter bubble while maintaining relevance: Targeted diversification with vae-based recommender systems, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2524–2531.

[19] F. Lu, A. Dumitrache, D. Graus, Beyond optimizing for clicks: Incorporating editorial values in news recommendation, in: Proceedings of the 28th ACM conference on user modeling, adaptation and personalization, 2020, pp. 145–153.

[20] V. W. Anelli, V. Bellini, T. Di Noia, W. La Bruna, P. Tomeo, E. Di Sciascio, An analysis on time- and

session-aware diversification in recommender systems, in: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 270–274. URL: https://doi.org/10.1145/3079628.3079703. doi:10.1145/3079628.3079703.

[21] F. Bianchi, F. Squazzoni, Agent-based models in sociology, Wiley Interdisciplinary Reviews: Computational Statistics 7 (2015) 284–306.

[22] F. Wall, Agent-based modeling in managerial science: an illustrative survey and study, Review of Managerial Science 10 (2016) 135–193.

[23] J. Zhang, G. Adomavicius, A. Gupta, W. Ketter, Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework, Info. Sys. Research 31 (2020) 76–101. URL: https://doi.org/10.1287/isre.2019.0876. doi:10.1287/isre.2019.0876.

[24] A. Ferraro, D. Jannach, X. Serra, Exploring longitudinal effects of session-based recommendations, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 474–479.

[25] D. Lee, K. Hosanagar, How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment, SSRN Electronic Journal (2017). doi:10.2139/ssrn.2603361.

[26] D. Fleder, K. Hosanagar, Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity, Management science 55 (2009) 697–712.

[27] G. Adomavicius, Y. Kwon, Improving aggregate recommendation diversity using ranking-based techniques, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 896–911. doi:10.1109/TKDE.2011.15.

[28] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, in: Proceedings of the 2012 ACM Conference on Recommender Systems (RecSys '12), ACM, New York, NY, USA, 2012, pp. 141–148. doi:10.1145/2365952.2365974.

[29] A. Starke, The effectiveness of advice solicitation and social peers in an energy recommender system, in: 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, IntRS 2019, CEUR-WS. org, 2019, pp. 65–71.