From Previous Plays to Long-Term Tastes
Exploring the Long-term Reliability of Recommender Systems Simulations for Children

Ungruh, Robin; Bellogín, Alejandro; Pera, Maria Soledad

DOI
10.1145/3705328.3759301

Publication date
2025

Document Version
Final published version

Published in
RecSys2025 - Proceedings of the 19th ACM Conference on Recommender Systems

Citation (APA)
Ungruh, R., Bellogín, A., & Pera, M. S. (2025). From Previous Plays to Long-Term Tastes: Exploring the Long-term Reliability of Recommender Systems Simulations for Children. In M. Bielikova (Ed.), *RecSys2025 - Proceedings of the 19th ACM Conference on Recommender Systems: Proceedings of the 19th ACM Conference on Recommender Systems* (pp. 1193 - 1198). (RecSys2025 - Proceedings of the 19th ACM Conference on Recommender Systems). https://doi.org/10.1145/3705328.3759301

From Previous Plays to Long-Term Tastes

Exploring the Long-term Reliability of Recommender Systems Simulations for Children

Robin Ungruh
Delft University of Technology
Delft, Netherlands
R.Ungruh@tudelft.nl

Alejandro Bellogín
Universidad Autónoma de Madrid
Madrid, Spain
alejandro.bellogin@uam.es

Maria Soledad Pera
Delft University of Technology
Delft, Netherlands
M.S.Pera@TUDelft.nl

Abstract

Studying the interplay of children and recommender systems (RS) is ethically and practically challenging, making simulation a promising alternative for exploration. However, recent simulation approaches that aim to model natural user-RS interactions typically rely on behavioral data and assume that user preferences remain consistent over time—an assumption that may not hold for children who undergo continuous developmental changes. With that in mind, we explore the extent to which simulations based on historical data can meaningfully reflect children's long-term consumption patterns. We do this via a simulation study using real-world data in which user behavior is modeled from observed listening preferences. Specifically, we probe whether simulation mirrors user preferences over time by comparing with organic (i.e., real) consumption patterns. Our findings offer a critical reflection on the reliability of simulation-based RS research for children and question the reliability of using behavioral assumptions to model users.

CCS Concepts

• Social and professional topics → Children; **• Information systems → Recommender systems**; **• Computing methodologies →** *Simulation environments*; *Simulation evaluation.*

Keywords

Recommender Systems, Children, Simulations

ACM Reference Format:
Robin Ungruh, Alejandro Bellogín, and Maria Soledad Pera. 2025. From Previous Plays to Long-Term Tastes: Exploring the Long-term Reliability of Recommender Systems Simulations for Children. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3705328.3759301

1 Introduction

Recommender systems (RS) often curate the items children encounter online; yet, they are driven by the needs of the 'majority', i.e., adults, overlooking children's interests and/or interactions with these systems [11, 20, 27]. This is not surprising, as children seldom play a role in (informing) the design and evaluation of RS [14]. The rare studies for which children are the main stakeholders show that recommendation algorithms (RAs) struggle to capture their preferences [43], deeming them "difficult" users [34]. Scarce data availability and limited access to real users [11, 20] make simulations a promising avenue to advance understanding of children's interplay with RS and evaluate how RS shape their experiences.

Simulations allow gauging what users would be exposed to when interacting with RS [7, 12, 13, 25]; they enable examination of RS' effects on long-term preferences [10, 15, 44] and recommendation utility for both users and providers [6]. Simulation outcomes can further reveal broader societal impacts, such as bias amplification [17–19, 31], preference homogenization [8], or the reinforced presentation of misinformation [4, 16]. To uncover such effects, long-term explorations are especially critical since many of these effects do not manifest immediately: feedback loops of RS-user interactions [8, 31] can gradually shift exposure distributions or amplify biases [19, 31]. Overlooking the long-term lens risks to underestimate cumulative effects or overestimate the persistence of short-term suitability of recommendations in the long term.

Still, previously mentioned explorations largely center on system-level effects, often overlooking whether the simulated behavior of users accurately reflects organic consumption patterns [7]. Recently, data-driven approaches have been introduced to better align simulated behavior with organic consumption patterns. They produce more realistic models of user behavior based on observed interactions with real-world systems, using past data to predict future choices [22, 25, 42]. Grounding simulations in empirical evidence offers a way to approximate long-term dynamics in RS interactions.

Data-driven approaches for modeling user behavior aim to mimic users' historical interactions via simulation [9, 45] and thus rely on a crucial assumption: *previously observed interaction patterns and preferences remain consistent over time.* We posit that for children, such an assumption does not hold, due to deviating consumption patterns across age [32, 36, 40, 43], alongside developmental changes throughout childhood [3, 5, 24]. Yet, there is no empirical evidence on whether RS simulations reflect children's preferences and consumption patterns. Establishing this knowledge is essential if simulation studies are to meaningfully capture longitudinal dynamics between children and RS. This prompts our **research question *(RQ)*:** *To what extent can RS simulations reflect the long-term consumption patterns of children?*

To address this RQ, we simulate long-term consumption behavior using an RS and compare the resulting interaction patterns to organic behavior. To facilitate this exploration, we use the LFM-2b dataset as it provides a long history of user interactions along with detailed age information, which allows us to track both short- and long-term changes in preferences focused on users of different ages. We conduct a two-phase exploration: In the *choice model fitting*

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3759301

phase, akin to [43], we explore varied model configurations that replicate users' listening choices to yield a holistic model that best fits the behavior of each user, i.e., tailored to individual patterns. In the *long-term simulation phase*, we evaluate whether this individually fitted model continues to reflect users' organic consumption beyond the phase in which it was fitted. We do this through a long-term simulation with users of varying ages. This allows us to assess how reliably previous user behavior can serve as a foundation for realistically simulating long-term consumption in RS simulations. Ultimately, this comparison enables us to reflect on the capacity of simulation-based studies to capture longitudinal dynamics, particularly for children. We publish our code at https://github.com/rUngruh/2025_RecSys_ChildSimulations.

2 Experimental Framework

Here, we lay out the experimental framework for our exploration.

2.1 Pipeline

As shown in Fig. 1, in the **choice model fitting** phase, RS simulations run for n iterations with different parameter configurations of a homogeneous choice model (CM), one that assumes the same behavior for all users, to identify the configuration that best replicates organic consumption behavior of each user $u \in U$. The best configuration for each u leads to the holistic model CM_{fit} [42], which considers individual differences in behavior. In the **long-term simulation** phase, we probe how well CM_{fit} continues to reflect users' behavior over time. We simulate with CM_{fit} for m iterations ($m \gg n$), spanning both the time used for fitting and subsequent periods, and compare simulated to organic consumption.

We extract organic interactions comparable to the simulation setup. For each u, we identify the set of items consumed across the entire exploration timeframe and extract interactions between timestamps t_{l-1} and t_l for $0 < l \leq m$ ($C_{u,l}$). For the simulation, we follow common approaches [e.g., 8, 23, 30, 31]. The simulation is initialized by creating a rating matrix M_0, including all dataset interactions before t_0 (start of the simulation). We set $l \longleftarrow 1$ and repeat for n (choice model fitting) or m (long-term simulation) interactions:

(i) Train a RA r on M_{l-1} and create for each $u \in U$ recommendations $R_{u,l}(r)$.
(ii) Simulate choices $\hat{C}_{u,l}(r, cm)$ for u assuming behavior based on a CM cm.
(iii) Update the rating matrix with the choices ($M_l \longleftarrow M_{l-1} \cup \hat{C}_{u,l}$) and set $l \longleftarrow l + 1$.

By retraining the RA at each iteration and feeding choices back into training, we model the feedback loops inherent in user–RS interactions [8, 31].

2.2 Algorithms & Choice Model

We use `ItemKNN` and `EASEr`, two well-established RAs [1]. In addition to their generally high performance, both have short training times, a valuable property when repeated training is necessary. At each iteration, the algorithm provides a recommendation list of 10 items to each user.

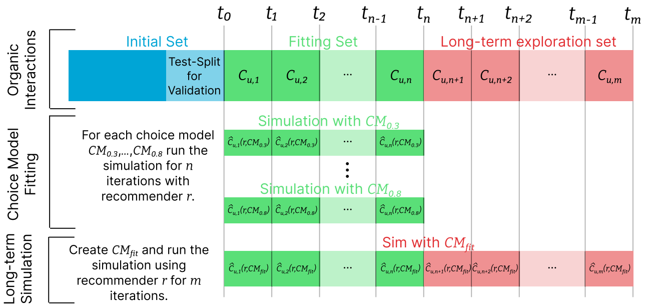

t_0
t_1
t_2
t_{n-1}
t_n
t_{n+1}
t_{n+2}
t_{m-1}
t_m
Organic Interactions
Initial Set
Test-Split for Validation
Fitting Set
$C_{u,1}$
$C_{u,2}$
···
$C_{u,n}$
Long-term exploration set
$C_{u,n+1}$
$C_{u,n+2}$
···
$C_{u,m}$
Choice Model Fitting
For each choice model $CM_{0.3},\ldots,CM_{0.8}$ run the simulation for n iterations with recommender r.
Simulation with $CM_{0.3}$
$\hat{C}_{u,1}(r,CM_{0.3})$
$\hat{C}_{u,2}(r,CM_{0.3})$
···
$\hat{C}_{u,n}(r,CM_{0.3})$
⋮
Simulation with $CM_{0.8}$
$\hat{C}_{u,1}(r,CM_{0.8})$
$\hat{C}_{u,2}(r,CM_{0.8})$
···
$\hat{C}_{u,n}(r,CM_{0.8})$
Long-term Simulation
Create CM_{fit} and run the simulation using recommender r for m iterations.
Sim with CM_{fit}
$\hat{C}_{u,1}(r,CM_{fit})$
$\hat{C}_{u,2}(r,CM_{fit})$
···
$\hat{C}_{u,n}(r,CM_{fit})$
$\hat{C}_{u,n+1}(r,CM_{fit})$
$\hat{C}_{u,n+2}(r,CM_{fit})$
···
$\hat{C}_{u,m}(r,CM_{fit})$

Figure 1: Exploration Pipeline

CMs directly impact simulation outcomes [21, 23]; but typically, they simplify realistic user behavior [7, 22]. To bypass this limitation, we use a cognitive CM based on ACT-R's base-level activation [33], which determines the likelihood of a user consuming an item based on the frequency and time since previous interactions with items and their genres. This model has been shown to capture music preferences [26] and can be adapted to individual users by adjusting its parameters. At each iteration, we compute

$$B(g, u) = \sum_{j=1}^{n} t_{u,g,j}^{-d} \tag{1}$$

for each genre g and u (normalized using softmax), where n is the number of times u has interacted with an item j of genre g and $t_{u,g,j}$ is the time in hours since u listened to j. The probability of u interacting with j is determined by summing the probabilities of genres associated with that item. The CM selects items based on these summed probabilities. This approach allows for the consumption of multiple items from recommendations in a single iteration. If no item is chosen, a random recommendation is picked. To determine parameters representative of user behavior, we explore a range within the fitting phase: $d \in [0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$. We denote the resulting CM as CM_d. A low d indicates that items of genres *frequently* consumed are favored; a high d favors *recently* consumed.

2.3 Data

We use **LFM-2b** [37], which includes detailed user demographics and over 2 billion user-song interactions for an extended time frame. We only explore the period with the most listening events: October 1, 2011 to September 30, 2016. We collect users' observed ages as of October 31, 2013[1]. For consistency in age annotation for each observed user-item interaction over the years, we assume users turned that age on October 1, 2013. This allows us to estimate users' ages for each interaction, with a potential error margin of ±1 year. We consider user-artist interaction, binarized by selecting the first interaction of a user with an artist. By focusing on first interactions, we capture newly discovered music and examine emerging preferences over time. To ensure that interactions with an artist are valid and reflect users' preferences, we only consider user-artist interactions if the user listened to the artist at least 5 times. Further, we only consider artists that have annotated genre information, as

[1] This information was kindly provided by the dataset creators via email.

gathered from LFM-1b UGP [38], which maps 219,022 artists to at least one out of 20 genres from Allmusic[2]. All genres belonging to an artist have a uniform weight distribution. Given our focus on children, we include a representative sample of young adults for context, as they represent a prominent mainstream group assumed not to undergo major developmental changes [43]. To create a balanced set that does not favor users of certain ages, we randomly sample 100 users for each *base age* (as of October 1, 2011) from 12 to 25, selecting only those who interacted with artists in every year of the exploration period[3].

We split the dataset (1,262 users, 120,245 artists, and 1,243,165 interactions) in three parts. The **initial set** (the first 6 months) is used to train RAs. The next 6 months mark the **fitting set**. This time frame enables an exploration of sufficient iterations to study users' behavior to fit the CM to them individually. The remaining four years (from October 1st, 2012) create the **long-term exploration set**. In the initial set, users interact on average with 177.45 artists. In the fitting set and the long-term exploration set, users interact with an average of 14.95 new artists each month.

2.4 Measures

We measure how well simulations match natural consumption with the overlap between genres of artists organically explored and those simulated to be explored during the same time, using genre calibration [29, 41, 43]. We aggregate artists explored throughout 28 days (i.e., 4 simulation rounds) to get a wider picture of consumed artists. **Artist-Genre Similarity** $AGS_u(l)$ measures the alignment by computing the complement of the Jensen-Shannon Divergence (JSD) between the average genre distribution of artists explored between t_l and t_{l-3} and the average genre distribution of artists simulated throughout the same timeframe. This sliding window approach helps smooth out short-term consumption fluctuations and provides a more robust signal of alignment between real and simulated behavior over time. We define the aggregated Artist-Genre Similarity between multiple iterations k and m as $AGS_u^{k,m} = \frac{1}{m-k+1} \sum_{l=k}^{m} AGS_u(l)$; which allows us to compare sustained alignment with real user behavior over an extended period.

Choice Consistency $CC_u^{sim}(l)$ determines the alignment of simulated choices with the initial preferences of u (i.e., the artists captured by the *initial set*). Akin to $AGS_u(l)$, it measures the complement of the JSD between the average genre distribution of artists simulated to be chosen by u between $t_{l-3}s$ and t_l and those organically consumed before t_0. $CC_u^{org}(l)$ measures the similarity of artists organically consumed between t_{l-3} and t_l and the artists interacted with before t_0. These metrics indicate how consistent (organic or simulated) choices are with users' previous listening patterns.

2.5 Setup

We use the Elliot framework [2] and tune the hyperparameters of the RAs, utilizing the last month of the *initial set* as the validation set. We use the entire *initial set* to initialize M_0. We run six *choice model fitting* simulations for each parameter configuration of the CM for $n = 25$ iterations as outlined at the beginning of this section, where the RA suggests 10 items, and we simulate user choices based on the CM. Based on the average alignment across all iterations, $AGS_u^{4,25}$ for each $u \in U$, we construct the CM_{fit}. For the *long-term simulation*, we run the simulation pipeline with CM_{fit} for $m = 235$ iterations, i.e., 6 months reflecting the *fitting set* and additional **four years**. We measure $AGS_u(l)$ for $4 \leq l \leq 235$ and analyze differences across varied *ages* and *over time*.

Table 1: Average $AGS_u^{4,25}$ across all $u \in U$ (AGS); percentage of users for which this parameter configuration resulted in highest $AGS_u^{4,25}$ (%); regression slope based on linear regression on $AGS_u(l)$ for $l \in \{4, 8, 12, 16, 20, 24\}$ (Slope). Bold indicates significant difference from CM_{fit}.

	ItemKNN			EASEr		
	AGS	%	Slope	*AGS*	%	Slope
$CM_{0.3}$	0.603	15.8	**-0.0033**	0.624	24.5	**-0.0027**
$CM_{0.4}$	0.610	17.0	**-0.0025**	0.623	16.0	**-0.0031**
$CM_{0.5}$	0.610	21.5	**-0.0025**	0.620	19.0	**-0.0032**
$CM_{0.6}$	0.605	21.9	**-0.0025**	**0.616**	18.0	**-0.0028**
$CM_{0.7}$	0.599	13.0	**-0.0028**	**0.611**	12.0	**-0.0030**
$CM_{0.8}$	**0.594**	13.8	**-0.0031**	**0.607**	10.5	**-0.0031**
CM_{fit}	0.604	—	-0.0005	0.625	—	-0.0004

3 Results

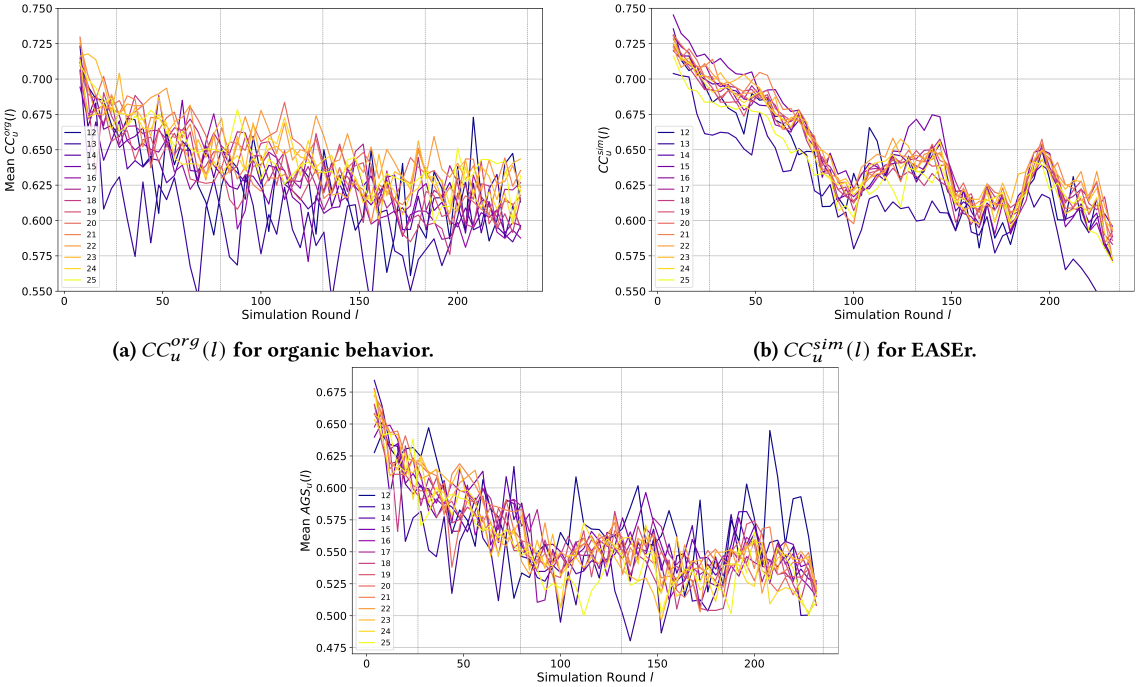
We present and reflect on the results from our 2-phase exploration.

Choice Model Fitting. Table 1 presents the average $AGS_u^{4,25}$ across all users $u \in U$ for each parameter configuration d of CM_d, highlighting how closely different CM configurations align with organic consumption during the *choice model fitting*. It also reports the percentage of users for whom each configuration was most aligned and thus selected for CM_{fit}. Overall, $AGS_u^{4,25}$ is similar across configurations. In turn, each configuration was the best fit for at least 10% of users, indicating diverse choice behaviors. Selecting the best-aligned behavior per user (reflected by CM_{fit}) leads to high short-term alignment, as reflected in higher $AGS_u^{4,25}$ than some of the homogeneous models (paired t-tests, $p < .05$). Importantly, CM_{fit} is never significantly worse than any of the homogeneous baselines. To examine alignment consistency during *choice model fitting*, we run linear regressions on $AGS_u(l)$ for CM_{fit} and each CM_d ($l \in \{4, 8, 12, 16, 20, 24\}$, to avoid overlapping windows). All models show declining alignment over time; CM_{fit} exhibits a significantly less steep decline (Welch's t-tests, $p < .05$) across recommenders and compared to all homogeneous configurations. These results suggest that while CM_{fit} is not always significantly better on average, its performance remains more stable over time.

Long-Term Simulation. To study long-term alignment, we compute $AGS_u(l)$, $CC_u^{org}(l)$, and $CC_u^{sim}(l)$ at every 4th iteration (i.e., $l \in 4, 8, \ldots, 232$), reflecting their definition over a sliding window of the last four iterations and avoiding overlapping measurements. Although we focus on longitudinal trends, we compute metrics separately for each year to capture potential variations over time.

To contextualize how users of different ages change in their preferences over time, we study users' consistency; Fig. 2a shows decreasing $CC_u^{org}(l)$ over time, uncovering that users deviate more

[2] https://www.allmusic.com/genres
[3] Due to filtering, only 23 12-year-olds and 42 13-year-olds, 97 14-year-olds remain.



(a) $CC_u^{org}(l)$ **for organic behavior.**

(b) $CC_u^{sim}(l)$ **for EASEr.**

(c) $AGS_u(l)$ **for EASEr.**

Figure 2: Average metrics by base age over iterations. Vertical lines mark the end of a year.

from their initial preferences as time progresses. For most users, consumption drifts gradually; yet, children's (particularly 13-year-olds) consumption immediately diverges from their initial preferences. Linear mixed models with *age* and *iteration* as fixed effects fitted on each year confirm this (test statistics and model details are provided in the repository); each year, *iteration* has a significant negative effect on $CC_u^{org}(l)$, confirming a temporal decline in consistency, while *age* has a significant positive effect, indicating that younger users tend to deviate more from their initial preferences.

To assess whether the simulation reproduces these organic shifts in consumption behavior, we compare the findings to the simulation-based metric $CC_u^{sim}(l)$ for EASEr (Fig. 2b). Here, choices align more closely with the initial user profile, as per significantly higher average $CC_u^{sim}(l)$ (mean = 0.646) than $CC_u^{org}(l)$ (mean = 0.632). Using ItemKNN yields significantly lower $CC_u^{sim}(l)$ (mean = 0.617).

A mixed linear model finds that the *iteration* has—in line with trends of organic behavior—a significantly negative effect on $CC_u^{sim}(l)$ for simulations with EASEr; alignment with initial preferences diminishes over time, albeit more gradually. A small but significant increase in the third year slightly moderates this trend. *Age* does not affect $CC_u^{sim}(l)$ in any year. To explore whether specific age groups differ, we compare average $CC_u^{sim}(l)$ across all years between base ages using independent t-tests ($p < .05$), revealing that users with base age 13 tend to deviate more from their initial preferences than other groups. Average $CC_u^{sim}(l)$ are significantly lower than for users with base age 15, 17, and 20 to 23. For ItemKNN the mixed linear model also shows a consistent decline in $CC_u^{sim}(l)$ across all years except Year 4. Here, *age* has no overall effect either, and no base age stands out for which users' mean $CC_u^{sim}(l)$ significantly differ from other age groups (independent t-test, $p < .05$). These findings show that while simulations mostly seem to replicate the temporal decline observed in organic consumption behavior, they do not fully capture the stronger age-related effects present in organic data, particularly the greater deviations among children.

To study whether inconsistent trends between simulations and organic consumption reflect long-term alignment of simulated and organic consumption, we analyze $AGS_u(l)$ for EASEr, Fig. 2c shows no notable difference between users of different ages. We run linear mixed models: the predictor *iteration* has a significantly negative effect on $AGS_u(l)$ in all years (except year 3), but *age* has no impact (except a negligible negative trend in year 5), indicating declining alignment between simulated and organic choices, regardless of user age. The drop is most pronounced during the first two years (80 iterations). Results for ItemKNN mirror these findings (additional figures in our repository). Overall, although both organic and simulated consumption patterns evolve, their trajectories diverge; changes in simulated behavior do not replicate the shifts observed in organic user behavior—regardless of age.

4 Discussion

We designed the CM to best align with each user's organic consumption patterns. However, alignment between simulated and organic interactions declines rapidly— noticeable already within the first months. Decreasing alignment between organic and simulated behavior raises concerns about the reliability of simulations to faithfully simulate children over time. Although the rate of deviation slows after 80 iterations, the CM fails to regain alignment with users' organic behavior. Our results indicate that this misalignment is not solely a result of the interaction between the RA and CM. Data-driven approaches try to mimic what a user consumed through observations. If they do that well (e.g., CM_{fit} with EASEr), they drift only slightly from initial preferences and remain relatively consistent in how they simulate user preferences. In contrast, organic consumption patterns captured by the used dataset tend to be less consistent. Users often deviate quickly from their initial preferences—the very signals that the RA and CM rely on for generating recommendations and simulating interactions.

Consider 13-year-olds. Early on, alignment between simulated and organic choices is relatively high (mean $AGS_u(4) = 0.684$). For example, during the first 4 weeks ($1 \leq l \leq 4$), 21.1% of the artists simulated to be consumed belong to the genre Alternative; in organic consumption, 17.1% are Alternative. After a year ($53 \leq l \leq 56$), simulated choices are relatively consistent; still, 23.8% of artists are Alternative, but organic consumption behavior shifts: only 15.7% of artists are Alternative. As a result, alignment drops notably (mean $AGS_u(56) = 0.578$). This trend continues: by age 17, the average alignment falls to $AGS_u(222) = 0.497$, showing that simulations capture even less of what these users actually consumed. In the organic behavior ($219 \leq l \leq 222$), Alternative becomes less prominent (12.9%) while the simulation remains consistent (21.6%).

These trends show that preference changes of individuals may not be addressed by simulation settings. For example, a user who at 13 eagerly listens to Alternative music might initially be closely represented by the simulation's predictions. Over the next few years, however, they could develop a preference for electronic music, whether due to changing personal taste, exploration of new genres, or peer influences [28]. This causes their actual consumption to diverge from the simulation, which continues to predict steady Alternative interest.

Although not the primary focus of this study, our analysis also revealed salient deviations in other consumption patterns. For example, the number of consumed items at each iteration decreases strongly for 13-year-olds. During the first 4 weeks, 13-year-olds consume an average of 13.9 artists. In the same timeframe, 1 year later ($53 \leq l \leq 56$), only 7.8 artists, and by age 17 ($219 \leq l \leq 222$), only 6.8. Simulated behavior remains markedly more consistent: In the same timeframes, the number of consumed artists changes from 18.7 to 18.8 in the first year, and only decreases to 8.9 at age 17. This indicates that the simulation does not directly capture the rapid decline in consumption observed in organic behavior.

These examples clearly illustrate a broader pattern: simulations do **not** mirror real child behavior as preferences develop over time. Surprisingly, this pattern is not unique to children; while our study setup mainly aimed to uncover discrepancies between simulated and organic user consumption of children, misalignment is observable across all considered ages, indicating broader limitations in capturing true user dynamics and evolving preferences.

Our study suggests that simulations can only yield meaningful insights if preferences and behavior are assumed to remain consistent or when alignment with actual behavior is not crucial. This makes alignment with organic consumption a relevant challenge for future RS simulations. Simulations hold great potential for studying the interplay between children and RS; what they are presented with, how systems shape their experiences, but also how effective RAs and interventions targeting children (i.e., safeguarding mechanisms) can be. However, such insights depend on simulations that realistically model long-term interactions. As children's preferences are particularly inconsistent, they warrant specific attention. To make simulations a reliable tool for studying child–RS interactions, data-driven models must move beyond replicating past behavior. Instead, they must incorporate mechanisms that capture the dynamic, evolving, and often inconsistent nature of human preferences.

Limitations & Future Work. The combination of RA and ACT-R-based CM offers a simple and easily applicable framework, but simplifies complex user behavior. Nuanced data-driven approaches [e.g., 9, 25, 45] should be explored to enhance realism and generalizability of findings. Further, consumption is never solely driven by RS; users also engage with items independently of RS. To capture organic behavior, simulations should account for non-recommended interactions; e.g., Hazrati and Ricci [22] introduce an "awareness set" to model items that users might discover and consume outside algorithmic curation. Changes in preferences and consumption patterns are not only impacted by intrinsic developments. Instead, external influences, such as evolving trends, can affect what a user consumes over time—a facet that future simulation settings should consider.

Finally, focusing on artist consumption follows previous research [17, 35, 39] and enables long-term simulations and analysis of the kind of items users consumed. While this simplification does not consider behavioral patterns like repetitions, it serves as a starting point for uncovering children's unique habits and preferences, which can be extended to more realistic interaction patterns. Explorations of the realism of simulation settings should be extended—next to better-suited RAs and CMs—to different domains and a variety of robust datasets, more nuanced user groups, and different timeframes to increase the reliability and generalizability of findings.

5 Concluding Remarks

Simulations are widely used to explore RS behavior, yet—based on a balanced sample of children and young adults in the music domain—our findings highlight that RS simulations fail to reflect organic user consumption patterns. We advocate that to unlock the full potential of RS simulation studies, they must move beyond reproducing past behavior and instead acknowledge the evolving and inconsistent nature of user behavior and preferences. Only then can simulations become a trustworthy lens for understanding how RS influence what children see, choose, and become.

Acknowledgments

Work partially supported by Grant PID2022-139131NB-I00 funded by MCIN/ AEI/10.13039/501100011033 and "ERDF, a way of making Europe." This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-14461.

References

[1] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-n recommendation algorithms: A quest for the

state-of-the-art. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization.* 121–131.

[2] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 2405–2414.

[3] Albert Bandura. 2009. Social cognitive theory of mass communication. In *Media effects*. Routledge, 110–140.

[4] Alejandro Bellogín and Yashar Deldjoo. 2021. Simulations for novel problems in recommendation: analyzing misinformation and data characteristics. http://arxiv.org/abs/2110.04037 arXiv:2110.04037 [cs].

[5] Laura Berk. 2015. *Child development.* Pearson Higher Education AU.

[6] Anas Buhayh, Elizabeth McKinnie, Clement Canel, and Robin Burke. 2025. Simulating the Algorithm Store: Multistakeholder Impacts of Recommender Choice. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization.* 274–279.

[7] Allison JB Chaney. 2021. Recommendation system simulations: A discussion of two key challenges. *arXiv preprint arXiv:2109.02475* (2021).

[8] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems.* 224–232.

[9] Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. 2019. Generative adversarial user model for reinforcement learning based recommendation system. In *International conference on machine learning.* PMLR, 1052–1061.

[10] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 525–534.

[11] Michael Ekstrand. 2017. Challenges in evaluating recommendations for children. In *International Workshop on Children & Recommender Systems. Available at: shorturl. at/osFV9.*

[12] Michael D Ekstrand. 2021. Multiversal Simulacra: Understanding Hypotheticals and Possible Worlds Through Simulation. *SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research, in conjunction with 15th ACM Conference on Recommender Systems (RecSys 2021), Available at https://doi.org/10.48550/arXiv.2110.00811* (2021).

[13] Michael D. Ekstrand, Allison Chaney, Pablo Castells, Robin Burke, David Rohde, and Manel Slokom. 2021. SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 803–805. doi:10.1145/3460231.3470938

[14] Michael D Ekstrand, Afsaneh Razi, Aleksandra Sarcevic, Maria Soledad Pera, Robin Burke, and Katherine Landau Wright. 2025. Recommending With, Not For: Co-Designing Recommender Systems for Social Good. *ACM Transactions on Recommender Systems* (2025).

[15] Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. 2022. Exposure inequality in people recommender systems: the long-term effects. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 194–204.

[16] Miriam Fernandez, Alejandro Bellogín, and Iván Cantador. 2024. Analysing the Effect of Recommendation Algorithms on the Spread of Misinformation. ACM, New York, NY, USA, 159–169. https://oro.open.ac.uk/96966/ Num Pages: 11.

[17] Andres Ferraro, Michael D Ekstrand, and Christine Bauer. 2024. It's Not You, It's Me: The Impact of Choice Models and Ranking Strategies on Gender Imbalance in Music Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems.* 884–889.

[18] Andres Ferraro, Dietmar Jannach, and Xavier Serra. 2020. Exploring longitudinal effects of session-based recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems.* 474–479.

[19] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval.* 249–254.

[20] Emilia Gómez Gutiérrez, Vicky Charisi, and Stephane Chaudron. 2021. Evaluating recommender systems with and for children: towards a multi-perspective framework. In *CEUR Workshop Proceedings. 2021; 2955.* CEUR Workshop Proceedings.

[21] Naieme Hazrati and Francesco Ricci. 2022. Recommender systems effect on the evolution of users' choices distribution. *Information Processing & Management* 59, 1 (2022), 102766.

[22] Naieme Hazrati and Francesco Ricci. 2022. Simulating users' interactions with recommender systems. In *Adjunct proceedings of the 30th acm conference on user modeling, adaptation and personalization.* 95–98.

[23] Naieme Hazrati and Francesco Ricci. 2024. Choice models and recommender systems effects on users' choices. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 109–145.

[24] Suzanne Hidi and K Ann Renninger. 2006. The four-phase model of interest development. *Educational psychologist* 41, 2 (2006), 111–127.

[25] Chih-Wei Hsu, Martin Mladenov, Ofer Meshi, James Pine, Hubert Pham, Shane Li, Xujian Liang, Anton Polishko, Li Yang, Ben Scheetz, et al. 2024. Minimizing live experiments in recommender systems: User simulation to evaluate preference elicitation policies. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2925–2929.

[26] Dominik Kowald, Markus Reiter-Haas, Simone Kopeinik, Markus Schedl, and Elisabeth Lex. 2024. Transparent music preference modeling and recommendation with a model of human memory theory. In *A Human-Centered Perspective of Intelligent Personalized Environments and Systems.* Springer, 113–136.

[27] Monica Landoni, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2024. Good for Children, Good for All?. In *European Conference on Information Retrieval.* Springer, 302–313.

[28] Brett Laursen and René Veenstra. 2021. Toward understanding the functions of peer influence: A summary and synthesis of recent empirical research. *Journal of Research on Adolescence* 31, 4 (2021), 889–907.

[29] Oleg Lesota, Jonas Geiger, Max Walder, Dominik Kowald, and Markus Schedl. 2024. Oh, Behave! Country Representation Dynamics Created by Feedback Loops in Music Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems.* 1022–1027.

[30] Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. 2021. T-RECS: A simulation tool to study the societal impact of recommender systems. *arXiv preprint arXiv:2107.08959* (2021).

[31] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management.* 2145–2148.

[32] Ashlee Milton, Levesson Batista, Garrett Allen, Siqi Gao, Yiu-Kai D Ng, and Maria Soledad Pera. 2020. "Don't judge a book by its cover": Exploring book traits children favor. In *Proceedings of the 14th ACM Conference on Recommender Systems.* 669–674.

[33] Markus Reiter-Haas, Emilia Parada-Cabaleiro, Markus Schedl, Elham Motamedi, Marko Tkalcic, and Elisabeth Lex. 2021. Predicting music relistening behavior using the ACT-R framework. In *Proceedings of the 15th ACM conference on recommender systems.* 702–707.

[34] Alan Said and Alejandro Bellogín. 2018. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction* 28, 2 (2018), 97–125.

[35] Diego Sánchez-Moreno, Ana B Gil González, M Dolores Muñoz Vicente, Vivian F López Batista, and María N Moreno García. 2016. A collaborative filtering method for music recommendation using playing coefficients for artists and users. *Expert Systems with Applications* 66 (2016), 234–244.

[36] Markus Schedl and Christine Bauer. 2017. Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young. In *1st International Workshop on Children and Recommender Systems, in conjunction with 11th ACM Conference on Recommender Systems (RecSys 2017)- Available at: https://doi.org/10.48550/arXiv.1912.11564.*

[37] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval.* 337–341.

[38] Markus Schedl and Bruce Ferwerda. 2017. Large-scale analysis of group-specific music genre taste from collaborative tags. In *2017 IEEE International Symposium on Multimedia (ISM).* IEEE, 479–482.

[39] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2020. Exploring artist gender bias in music recommendation. In *ImpactRS Workshop at ACM RecSys '20,.*

[40] Lawrence Spear, Ashlee Milton, Garrett Allen, Amifa Raj, Michael Green, Michael D Ekstrand, and Maria Soledad Pera. 2021. Baby shark to barracuda: Analyzing children's music listening behavior. In *Proceedings of the 15th ACM Conference on Recommender Systems.* 639–644.

[41] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems.* 154–162.

[42] Robin Ungruh, Alejandro Bellogín, and Maria Soledad Pera. 2025. From Monolith to Mosaic: Uncovering Behavioral Differences for Choice Models in Recommender Systems Simulations. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval.*

[43] Robin Ungruh, Alejandro Bellogín, and Maria Soledad Pera. 2025. The Impact of Mainstream-Driven Algorithms on Recommendations For Children. In *European Conference on Information Retrieval.* Springer.

[44] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research* 31, 1 (2020), 76–101.

[45] Xiangyu Zhao, Long Xia, Lixin Zou, Hui Liu, Dawei Yin, and Jiliang Tang. 2021. Usersim: User simulation via supervised generativeadversarial network. In *Proceedings of the Web Conference 2021.* 3582–3589.