When Preference Is Not Enough

Why Recommender Systems Require Human-Aware Evaluation for Children

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

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

Dominik Kowald
dkowald@know-center.at
Know Center Research GmbH & University of Graz
Graz, Austria

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

Abstract

Children regularly interact with recommender systems, yet little is known about whether the suggestions they encounter *fit* them. Traditional accuracy-based evaluation accounts for user preference, providing an incomplete picture of how well recommenders serve young users. With that in mind, we adopt a human-centric, specifically *child-centric*, evaluation perspective to empirically examine whether recommender systems address children's needs. Focusing on content maturity as a key dimension influencing what is considered fitting as per developmental needs, we probe a range of recommender algorithms on whether their suggestions align with children's developmental maturity and how closely these suggestions reflect the content maturity of previously consumed items. Our analysis showcases that traditional evaluation paradigms fail to uncover dynamics that affect alignment with children's needs.

CCS Concepts

• **Social and professional topics** → **Children**; • **Information systems** → **Recommender systems**.

Keywords

Recommender Systems, Evaluation, Children, Maturity

ACM Reference Format:
Robin Ungruh, Alejandro Bellogín, Dominik Kowald, and Maria Soledad Pera. 2026. When Preference Is Not Enough: Why Recommender Systems Require Human-Aware Evaluation for Children. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3774935.3812719

1 Introduction

Recommender systems (RS) are commonly evaluated in offline settings using accuracy-related metrics that assess how well recommendations match users' past consumption behavior. RS research often assumes that prior consumption is a sufficient indicator of preference and recommendation quality. While this assumption is already debated when it comes to the general population [9], it becomes particularly problematic when evaluating recommendations **children** are exposed to. Children are a non-mainstream user group with distinct needs and vulnerabilities, for whom recommendation quality cannot be reduced to a single, static notion. Developmental stages and evolving maturity further challenge the reliance on homogeneous quality criteria, highlighting that traditional evaluation framings overlook what makes a recommendation *fitting* for a child [4, 8]. Continued adoption of traditional frameworks risks misleading analysis of RS for children, a crucial user base.

In light of these shortcomings, we conduct an empirical exploration that gauges RS from a child-centric perspective. As a concrete starting point, our analysis uses children's maturity, operationalized via age, as a proxy for their needs and questions *to what extent RS fit children's maturity.* To answer this, we leverage traditional *audience labels*, a widely used, normative reference for assessing content maturity, offering coarse-grained but interpretable age-based guidance on suitability [10]. We probe a range of recommendation algorithms (RAs) using both conventional and child-centric evaluation perspectives, the latter informed by maturity alignment. For the latter, we assess (i) whether audience labels assigned to recommendations exceed what is considered appropriate for a child's age and (ii) the degree to which recommendations reflect preferred content maturity inferred from children's prior consumption patterns. This analysis provides a human-aware perspective, combining traditional accuracy-focused metrics with human-centric ones [7] that capture the needs of the target group. By identifying discrepancies between traditional accuracy measures and maturity-focused criteria, we establish how recommendation quality for children is misrepresented under standard evaluation paradigms.

2 Experimental Setup

We describe our experiments below. We share code and extended results: https://github.com/rUngruh/2026_UMAP_age_ratings.

Data We utilize two datasets in our exploration. The well-established **MovieLens-1M (ML)** [5] provides user-movie interactions, which we enrich[1] with audience labels from the Motion Picture Association (MPA)[2] through the OMDB-API[3]. We binarize

This work is licensed under a Creative Commons Attribution 4.0 International License.
UMAP '26, Gothenburg, Sweden
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2311-7/2026/06
https://doi.org/10.1145/3774935.3812719

[1] In case of non-unique matches (75 movies), we annotated items manually using the IMDb (https://www.imdb.com/) website.
[2] https://www.motionpictures.org/
[3] https://www.omdbapi.com/

Table 1: Audience labels and items/interactions per dataset.

		ML		MAL	
Label	Definition	% items	% interactions	% items	% interactions
`Not Rated`	No classification	9.60	3.77	0.40	0.01
`Approved`	General audiences	9.17	7.04	—	—
`G`	All ages	5.01	6.01	11.97	2.16
`PG`	Parental guidance	16.13	21.98	6.50	2.16
`PG-13`	13+	16.28	16.32	52.67	57.65
`R`	17+	43.31	44.65	12.65	27.15
`R+/NC-17`	Adults only	0.48	0.11	10.32	10.35
`Rx`	Pornographic	0.02	0.00	5.49	0.49

the dataset by treating ratings > 3 as positive signals and retain users who rated at least 5 items positively. The processed dataset includes 6,034 users and 575,272 interactions with 3,533 items. In ML, users are annotated with broad age groups, heavily concentrated in the 25–34 age group, with noticeably fewer interactions in younger and older groups; 222 users are children (*Under 18*).

MyAnimeList (MAL)[4] provides user interactions with Anime tracked on the MyAnimeList website. The dataset includes interaction timestamps (i.e., viewing completion time of an anime); we restrict our exploration to interactions that occurred in 2015, the year with the most logged interactions. This allows us to examine age-specific preferences while still capturing a wide range of user behavior and consumption patterns. To account for typical age restrictions on online platforms, we only include interactions where the user was at least 13 years old. We treat users' age on January 1st, 2016, as their age for this experiment, i.e., when they interact with RS. We rely on audience labels obtained from the dataset. In line with ML, we treat ratings > 6 (out of 10) as positive signals and exclude users with fewer than 5 positive interactions. This results in 39,377 users, 1,668,352 interactions with 4,805 items. Overall, the age distribution is bell-shaped, centered around 18–20, with interactions gradually decreasing for younger teens and users in their mid- to late-20s. 16.51% of users are younger than 18[5].

Table 1 shows the audience labels for both datasets along with the number of items per label. We deem all items with labels `R` *formally* unsuitable for children younger than 17. Items annotated with `R+/NC-17` or `Rx` are only suitable for adults (i.e., 18+). As age groups on ML are broader, we treat items with `R`, `NC-17`, and `Rx` as unsuitable for users in the `Under 18` group.

Evaluation Protocol We probe a range of RAs: `MostPop` as a baseline, along with personalized counterparts, i.e., neighborhood-based (`ItemKNN` and `UserKNN`), graph-based (`RP`$^3\beta$), latent factor (`MF2020` and `SLIM`), and autoencoder models (`MultiVAE` and `EASEr`).

For consistency, as ML does not include timestamps, we use a random splitting strategy for both datasets: 70% of each user's interactions are used for training, 20% for validation, and 10% for testing. Using the Elliot framework [1], we conduct hyperparameter tuning using Tree Parzen Estimator, optimizing for *nDCG*@10 for 20 iterations. We validate neural models every 5 epochs and stop early if there is no improvement for 5 validations. We train the final recommender on the union of the train and validation sets using the best hyperparameters, and create $N = 25$ recommendations per user, balancing realistic recommendation list lengths with sufficient coverage to analyze category distributions. Interactions in the combined set are treated as *user profiles* for subsequent analysis, as they constitute the basis for recommendation generation.

We evaluate RAs using *nDCG*, a traditional metric, and maturity alignment criteria. For the latter, we describe each *user profile* based on: (1) *formal maturity*, i.e., the age in the dataset (broad categories in ML, fine-grained ages in MAL); (2) *maturity profile*, i.e., how frequently a user engages with content of each audience label. This is modeled by a distribution across audience labels, normalized at the user level. Formally, the maturity profile of user u is represented by a vector $\mathbf{m}_u$. Analogously, we represent the maturity composition of a recommendation list for u by an RA a as the proportion of items assigned each audience label within the recommendation list; formally $\mathbf{r}_{u,a}$. To measure maturity alignment for an RA a, we use:

- *Audience Label Exceedance* (ALE_u): The proportion of recommendations assigned audience labels above u's formal maturity.
- *Audience Label Difference* ($\Delta AL_u(l)$): The alignment of recommendations with maturity profiles, indicating whether audience label l is amplified or suppressed in recommendations. It is defined as the difference between the proportion of l in $\mathbf{r}_{u,a}$ and the respective proportion in $\mathbf{m}_u$.

3 Results & Discussion

Here, we report and discuss results from our empirical exploration. When comparing measures across ages, we conduct one-way ANOVAs with Tukey's HSD post hoc tests for pairwise differences ($p < .05$). To compare measures across RAs or audience labels, we conduct pairwise t-tests with Bonferroni correction ($p < .05$). Given the magnitude of pairwise comparisons, we highlight salient effects and include the full statistical analysis in our repository.

Traditional Lens. Our initial analysis aligns with conventional practices for evaluating RS based on preference assessed using traditional relevance metrics. Performance across RAs (Fig. 1) tends to be the highest for age groups prominent in the datasets. On ML, performance peaks for users aged 25–34; on MAL, for users aged 15–17. The *nDCG* scores for these groups are significantly higher than for most others. This pattern emerges for personalized RAs and `MostPop`, indicating that performance gains for these users are not specific to personalized methods. Across RAs on MAL, young children (≤ 15) receive recommendations that are significantly less aligned with their preferences than more prominent user groups, but differences remain small. As users age, *nDCG* scores initially improve, but decline again for older adults. This was noted in comparable studies on other domains and RAs [2, 11]: RAs may match children's preferences less accurately than those of mainstream

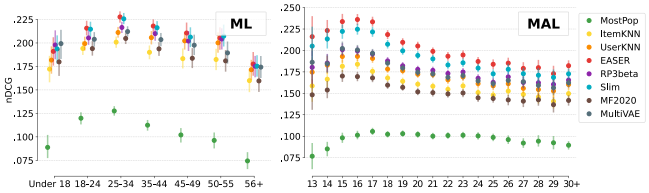

Figure 1: Average nDCG per age group.

[4]We use the cleaned version of the dataset, deemed the most reliable, as it only considers interactions where an Anime was marked as 'completed'. Available at: https://www.kaggle.com/datasets/azathoth42/myanimelist?resource=download

[5]Number of users per age: 13: 138; 14: 476; 15: 1,125; 16: 1,911; 17: 2,853.

users, but differences are typically modest. In fact, *nDCG*-based findings do not position children as 'underserved' users as RAs perform reasonably well for them compared to the entire user base. Notably, variation between RAs is greater on MAL than ML.

Toward a Child-centric Lens. Preference-centered evaluation alone provides no evidence for RA fulfilling children's needs. Hence, we analyze maturity alignment to move beyond what children may *like* toward what *fits* their developmental needs.

Maturity Alignment. *ALE* scores show that recommendations for children frequently exceed formal maturity. Fig. 2 highlights the proportion of 'high-maturity' items in recommendations; i.e., R, R+/NC-17, and Rx items. Recall that for 17-year-olds on MAL, R-rated items are deemed suitable. On ML, the proportion of high-maturity recommendations peaks for ages 18–24 and 25–34 for all RAs (except MostPop showing less variation across ages), with significantly higher proportions of high-maturity items than children. On MAL, we observe a more gradual increase up to age 18; younger user groups receive significantly fewer high-maturity items.

Across RAs, children receive on average 42 to 45% high-maturity recommendations on ML; between 32 and 48% on MAL, where younger children fall toward the lower end of this range. As an example, in the case of users aged 16 in MAL, this means that 11 of the 25 recommendations surpass formal maturity and are considered unsuitable. Across ages, recommendations with audience labels exceeding users' ages remain common rather than exceptional.

To understand the composition of high-maturity recommendations, we examine a concrete case using SLIM on MAL, chosen because it exhibits high variances in *ALE* scores between ages: Most high-maturity items recommended by SLIM are R-rated (approximately 30% of recommendations on average), with an additional 5–10% R+ items. Rx-rated items are rarely recommended ($< 0.1\%$).

Recommendations' audience labels exceeding users' ages indicate that while RAs may perform satisfactorily under preference-centered evaluation, limiting the analysis to this perspective obscures algorithmic tendencies that explicitly affect children.

Child-Recommender Interplay. Our analysis thus far has shown the types of recommendations children receive, but it provides limited insights into how recommendations manifest for children in practice. To uncover why children become exposed to mature content, we examine the interplay between users and RAs and interaction dynamics that shape recommendations.

To study whether recommendations of mature items stem from **prior consumption**, we look at children's maturity profiles. Children frequently consume items with audience labels formally exceeding their ages (Fig. 3). On ML, children consume on average 37.78% high-maturity items, while older groups (18–44) consume

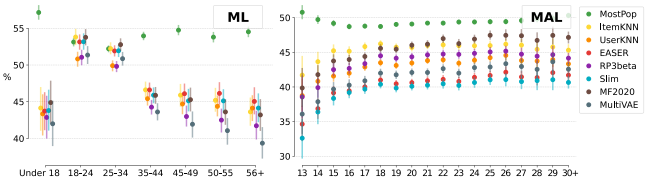

Figure 2: Percentage of high-maturity recommendations.

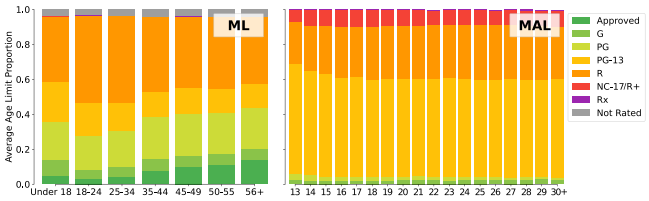

Figure 3: Average maturity profiles per age group.

significantly more of these items. On MAL, children not only frequently consume high-maturity items (already 30.96% of items consumed by 13-year-olds are high-maturity), but older teenagers are already among the users with the highest share of high-maturity items consumed (38.67% of items consumed by 17-year-olds). Significant differences in exposure across age are observed only for users younger than 16, whose consumption of high-maturity items is significantly lower than that of older teens and adults. Most high-maturity items consumed are R-rated ($\sim 70\%$); Rx-rated items account for less than 1% of all interactions.

Despite smaller differences between datasets, consumption of mature items is not exceptional but a consistent pattern among children. On MAL, R-rated content is consumed particularly disproportionately: although it represents only 12.65% of the catalog, it accounts for 27.15% of interactions (Table 1).

As children consume mature items, one might expect personalized algorithms to simply reflect these patterns and thus mirror maturity profiles in RA output. This would mean that $\Delta AL(l)$ scores, which indicate whether certain audience labels are **amplified** or **suppressed** in recommendations, would remain small for all labels. However, across datasets and RAs, this is not the case. Instead, $\Delta AL(l)$ reveals systematic amplification and suppression of audience labels. As patterns are largely consistent between RAs, we illustrate these dynamics using recommendations generated by UserKNN: On ML (Fig. 4, left), $\Delta AL(l)$ is near-zero for various audience labels. However, some amplification and suppression patterns emerge across all users. PG-13 and Approved items are slightly suppressed by UserKNN overall, while PG and R items are amplified, with significant differences relative to other audience labels. For children specifically, R items are particularly amplified and PG-13 items are suppressed. Notably, children are the only group that receives fewer PG items than previously consumed. These dynamics reflect children's consumption patterns (cf. Fig. 3): they show relatively low consumption of R items, which RAs tend to prioritize, while their preferred PG and PG-13 items become less prominent.

On MAL (Fig. 4, right), most labels are suppressed by UserKNN (most saliently PG-13 and R+); R items are the only ones exhibiting amplification. Unlike on ML, this amplification is not significantly stronger for children; age does not significantly affect $\Delta AL(\mathrm{R})$, indicating that intensification of R-rated items is not significantly age-dependent. This *'maturity lift'*, i.e., the increase of R-rated content in recommendations while other categories are subdued, corresponds to an already disproportionate share of interactions with R-rated content (cf. Table 1). Rather than simply reflecting consumption patterns, the item category with already a large share of interactions receives even more prominence, to the detriment of children, for whom this very content may not match their needs.

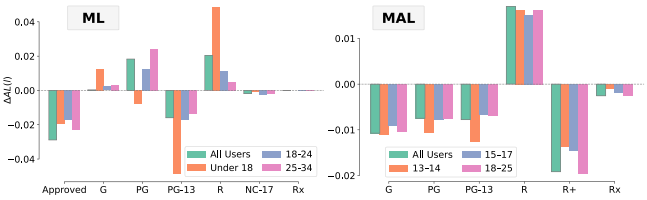

Figure 4: Average ΔAL for prominent age groups (`UserKNN`), indicating differences of audience label proportions in recommendations in comparison to maturity profiles (cf. Fig. 3)

Having examined calibration relative to observed consumption patterns, we now look at **exposure without prior consumption**, specifically children (17 or younger) in MAL who have not consumed any high-maturity items ($N = 174$)[6]. Our analysis of *ALE* scores across RAs indicates that users in this subset, despite never consuming such items, are frequently recommended high-maturity content, albeit at significantly lower levels than other children. Zooming in on `UserKNN` as a RA representative of trends observed, we see that every child without prior consumption of high-maturity content still receives at least one such item in their top-25 recommendations. In fact, on average, 26.41% of `UserKNN`'s recommendations to these users are `R`-rated items (compared to 34.64% for children with prior consumption). Mature content occurs not merely as an amplification of past behavior but emerges consistently in recommendations, even without prior consumption.

4 Concluding Reflections

Conventional evaluation paradigms provide an incomplete picture of recommendation quality, relying on already insufficient proxies [6, 9]. Our work underscores the limitations of this myopic accuracy-centered view: solely measuring how well recommendations align with prior engagement would mislead us into believing that RAs serve most children appropriately. However, with a child-centric lens of age-appropriate content, we show that standard metrics miss critical aspects of quality entirely. Specifically, RAs fail to align with formal maturity and even actively amplify exposure to mature content, including for children who had not previously interacted with such material. These patterns reveal systematic algorithmic tendencies that are obscured under traditional evaluation frameworks. For children who have non-mainstream needs and may be particularly vulnerable, ignoring such human-centric perspectives risks producing reductive insights about recommendation performance. Overall, our results showcase that *preference*-focused evaluation is insufficient to determine whether recommendations are *fitting* for children, highlighting the need for complementary *human-centric* perspectives for evaluations to become truly **human-aware**.

RAs in our study are not **child-centric**; they are neither aware of nor tuned to accommodate children's needs. Within conventional evaluation paradigms, recommending mature content is not considered a failure: RAs typically have no explicit knowledge of users' developmental needs or item suitability and are optimized for engagement and general preferences [6]. Still, recommendations lead to tangible and potentially harmful outcomes for children, raising concerns that remain largely invisible under standard evaluation. To make these dynamics measurable, we focus on content maturity as a proxy for appropriateness. Audience labels are a coarse and imperfect indicator, failing to capture contextual, cultural, or individual vulnerabilities. Yet, they remain the dominant operational mechanism through which platforms enforce "child appropriateness". As such, they are a starting point for more deliberate investigations that consider individual characteristics, vulnerabilities, and needs.

When knowledge about users is limited or when 'fit'-criteria extend beyond simple proxies to include contextual and individual vulnerabilities, **complexity** of human-centric evaluations increases. In turn, there are no simple solutions for issues identified in practice. For example, when age information is unavailable or unreliable, merely filtering recommendations to ensure suitability is deemed to fail. More nuanced approaches to adapting RS and parental controls, grounded in children's actual needs, remain underexplored across platforms accessed by children. Despite such challenges, we show that explicitly child-centric evaluation lenses are necessary; without them, developers of RS lack meaningful incentives to prioritize children's protection. This sentiment aligns with broader calls for human-centric [7] and emerging regulatory goals [3]. These approaches require, in practice, coordinated, multi-stakeholder efforts to advance evaluations that capture what truly matters to users and support the design of systems that serve them. We, thus, advocate for complementary human-centric evaluation lenses and protocols on a quest for safeguarding and ensuring children's well-being.

[6] ML is excluded, as only 12 children have not consumed any high-maturity items.

Acknowledgments

This research was supported by the Austrian FFG COMET program and by Grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033 and "ERDF, a way of making Europe."

References

[1] 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 *Proc. of ACM SIGIR*. 2405–2414.
[2] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proc. of FAccT*. 172–186.
[3] Matteo Fabbri and Ludovico Boratto. 2025. Auditing Recommender Systems for User Empowerment in Very Large Online Platforms under the Digital Services Act. In *Proc. of RecSys*. 51–61.
[4] Emilia Gómez Gutiérrez, Vicky Charisi, and Stephane Chaudron. 2021. Evaluating recommender systems with and for children: towards a multi-perspective framework. In *PERSPECTIVES, co-located with RecSys*.
[5] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM TIIS* 5, 4 (2015), 1–19.
[6] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara fallacy: Towards more impactful recommender systems research. *AI Magazine* 41, 4 (2020), 79–95.
[7] Joseph Konstan and Loren Terveen. 2021. Human-centered recommender systems: Origins, advances, challenges, and opportunities. *AI Magazine* 42, 3 (2021).
[8] Monica Landoni, Theo Huibers, Emiliana Murgia, Mohammad Aliannejadi, and Maria Soledad Pera. 2021. Somewhere over the rainbow: Exploring the sense for relevance in children. In *Proc. of ECCE*, Vol. 42. 1–5.
[9] Alan Said, Maria Soledad Pera, and Michael D Ekstrand. 2025. We're Still Doing It (All) Wrong: Recommender Systems, Fifteen Years Later. In *BEYOND. CEUR-WS.*
[10] Kimberly M Thompson and Fumie Yokota. 2004. Violence, sex, and profanity in films: Correlation of movie ratings with content. *Medscape General Medicine* 6, 3 (2004), 3.
[11] Robin Ungruh, Alejandro Bellogín, Dominik Kowald, and Maria Soledad Pera. 2025. Impacts of Mainstream-Driven Algorithms on Recommendations for Children Across Domains: A Reproducibility Study. In *Proc. of RecSys*. 783–791.