Impacts of Mainstream-Driven Algorithms on Recommendations for Children Across Domains: A Reproducibility Study

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 & TU Graz
Graz, Austria

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

Abstract

Children are often exposed to items curated by recommendation algorithms. Yet, research seldom considers children as a user group, and when it does, it is anchored on datasets where children are underrepresented, risking overlooking their interests, favoring those of the majority, i.e., mainstream users. Recently, Ungruh et al. demonstrated that children's consumption patterns and preferences differ from those of mainstream users, resulting in inconsistent recommendation algorithm performance and behavior for this user group. These findings, however, are based on two datasets with a limited child user sample. We reproduce and replicate this study on a *wider range of datasets* in the movie, music, and book *domains*, uncovering interaction patterns and aspects of child-recommender interactions consistent across domains, as well as those specific to some user samples in the data. We also extend insights from the original study with *popularity bias metrics*, given the interpretation of results from the original study. With this reproduction and extension, we uncover consumption patterns and differences between age groups stemming from intrinsic differences between children and others, and those unique to specific datasets or domains.

CCS Concepts

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

Keywords

Recommender Systems, Reproducibility, Popularity Bias, Mainstream, Children

1 Introduction

Children, a heterogeneous population with unique preferences in media [32, 35, 42] and interaction patterns with online systems [20, 46], are frequently exposed to decisions made by recommender systems (RS) on the various online platforms they use. RS research, however, rarely has them as the protagonist. Typically, design and evaluation of RS are informed by data capturing user-system interactions from a broad user group, where children remain a minority. As the preferences of minorities tend to visibly deviate from the majority—a fact RS may not capture [10, 18, 30, 34]—and with children being vulnerable and in-development users [14, 44], explicit attention must be paid to their preferences and interaction patterns with RS [12, 25, 42]. As most knowledge about RS stems from mainstream users [8, 21, 22, 28, 49], concerns arise as to whether recommendation algorithms (RAs) can adequately accommodate the needs of an underrepresented group like children or if suggestions are inherently skewed toward mainstream actions.

Recently, Ungruh et al. [42] introduced a reference framework to study genre preference deviations between children and mainstream users and how RS treat these user groups. They apply this framework to the MovieLens-1M (ML) [15] and LFM-2b [36] datasets in two experiments—one focused on age-related user preferences, and one on the dynamics between RS and children. Outcomes reveal *deviating preferences* between children and mainstream users in the music and movie genres they consume, and that, while most of the RAs studied produce music recommendations that reasonably align with child preferences, for certain RAs the dominance of mainstream users in the training data *skews* recommendations away from children's preferences.

The at times conflicting findings, exacerbated by the limited datasets used—the main one no longer available—jeopardize result generalizability and call for additional analysis. This prompts us to conduct a reproducibility study; elaborating on how children differ from the mainstream, and whether RAs account for emerging differences. Our motivation is rooted in three main pillars:

1. Children as Non-Mainstream Users. Ungruh et al. [42] highlight that across ML and LFM-2b, children's consumption of items of different genres differs from those of mainstream users. As these outcomes are limited to two datasets and domains, it is unclear whether such deviations are unique to the datasets studied or represent a broader trend observable across datasets.

2. Dominance of User Groups & Deviating RA Behavior. Mainstream users' prominence in data used to train RAs affect how well these systems fare for children, i.e., the quality of the recommendations presented to this group [42]. As this is based on a snapshot from LFM-2b, this takeaway may be limited to the domain studied, particularly considering that music consumption has unique characteristics compared to other common domains in RS research [38]. Additionally, the analysis is restricted to a sample of users and a

The definitive Version of Record was accepted for publication in the *19th ACM Conference on Recommender Systems (RecSys 2025)*, https://doi.org/10.1145/3705328.3748160

limited timeframe, raising critical concerns about the generalizability of findings. It also remains unclear why certain preferences are effectively captured while others are not. The reference work offers interpretations of which interaction patterns may lead to better genre alignment between recommendations and preferences, other aspects are not studied, and assumptions require further validation.
3. Reproducibility Concerns. Thorough reproducibility is a challenging endeavour [6, 11, 39]. For the reference work, this is further complicated by the unavailability of the LFM-2b dataset, making direct reproducibility impossible. But even replicability, conducting the same experiments with different datasets, is not a straightforward process, considering the limited data sources available that include children, required data properties, and available metadata to conduct the study. Ungruh et al. [42] already recognize limitations in data processing and gathering for child-centered analyses; moreover, the reference framework is tailored for a specific use case and datasets, giving cause to inquire about the feasibility of closely replicating the original setup with other datasets.

This Work. Driven by the aforementioned motivations, in this work, we (1) **reproduce** the reference work to probe the results obtained by the original study, (2) **replicate** it by conducting the studied experiments with new datasets and a new domain to broaden insights and generalizability of findings from the reference work, and (3) **extend** the analysis to explore additional facets of children's and mainstream users consumption patterns. In doing so, we broaden the scope of the original study by further exploring child and mainstream preferences within the context of RS in three distinct domains: (1) *movies* using ML, (2) *music* using MLHD [45]—a dataset that has received little attention by the RS community, which becomes particularly relevant as an alternative to LFM-2b, allowing us to verify children's consumption trends in a different dataset in the music domain—, and (3) *books* using Book Crossing (BX), a domain not considered by the original study.

Although children are not a uniform user group, we often refer to children as *one* group to create a foundational understanding of the interplay between children and RS. This simplification aids understanding of children's role as non-mainstream users in a landscape increasingly shaped by ubiquitous RAs. Children are a vulnerable user group. As such, ethical considerations and deliberation are important. Thus, we only utilize publicly available datasets where users voluntarily self-declared age information, i.e., information about users was crawled and aggregated in the used datasets.

Contributions. This work expands the reference framework for an extended picture about children and mainstream users. With our multi-domain study, we probe trends consistent with the reference work as well as those deviating. Generalization to other datasets advances knowledge on how RAs fare for non-mainstream user groups more broadly; auditing their ability to serve *each* user well.

Reproducibility. We provide code to reproduce our experiments (https://github.com/rUngruh/2025_RecSys_Reproducibility). We publish the used sample of MLHD, enabling easier filtering for future studies (https://zenodo.org/records/15394228).

2 Reference Work

The reference work [42] provides a **reference framework** along with associated code to enable reproducibility. This framework has the following key components: a **dataset** in a given domain, a classification of users into distinct **user groups**—including a method to identify one as **mainstream**—and a way to quantify **preference** alignment, with a focus on the alignment between user preferences and recommendations generated by a set of **RAs**. The framework is applied on two experiments: (1) the *Preference Deviation Exploration*, which determines the degree to which groups differ from the mainstream, and (2) the *RS experiment*, which examines the impact of mainstream users' presence in the source data on the quality of recommendations for an underrepresented group.

The reference framework compared the genres of consumed and recommended items of users in different age groups, focused on the user group of children. This was based on two datasets, (1) ML, which captures movie ratings, and (2) LFM-2b, which explores music listening events. By grouping users based on their age, the authors categorize `mainstream users` as users belonging to the most common age group—young adults who contribute the majority of data—`children` are younger minors, while Non-Mainstream Adults (`NMAs`) are adults older than the `mainstream`. In both datasets, `children` are only responsible for a minority of the user-item interactions (2.83% and 7.07%, respectively). To investigate differences in preferences, as well as the degree to which these preferences are captured by common RAs, the study assesses user preferences by the genres of items previously consumed by a user to measure genre alignment, a concept closely related to *miscalibration* [40].

The *Preference Deviation Exploration* surveys the differences between age groups, in particular between `children` and `mainstream users`, across both datasets. Aggregating the preferences of users belonging to a certain age group enabled assessment of how much users within one age group deviate from each other, but also analysis of the degree to which age groups differ from each other. Outcomes showed that, regardless of the dataset, genre consumption differs between age groups, with `children`'s genre preferences differing markedly from those of `mainstream users`.

The *RS experiment* considers the top-50 recommendations created by varied RAs for a sample from LFM-2b. To gauge whether these recommendations align with the preferences of users of different ages, the analysis—by age group—leverages classical performance metrics and the alignment of genres between users' previous consumption and recommendations. To probe the impact of `mainstream users` in the train data on the underrepresented group, a two-step evaluation is adopted: RAs are first trained on a `General Set` that includes the interactions of users of varying ages (where `mainstream users` dominate the data) and once on a `Child Set` that only includes interactions of `children`. As the RAs do not have any `mainstream` data available in this latter setup, performance in the two recommendation scenarios can be compared to assess the impact of `mainstream users` on recommendations for `children`.

3 Reproducibility, Replicability, and Extension

Our work aims to both address constraints of the reference work and contextualize replicated results with additional explanatory factors. As the reference work published associated code, we can follow the reference framework, only making adaptations to improve the clarity of the code and facilitate integration of new datasets. An overview of our reproducibility efforts can be seen in Table 1.

Table 1: Overview of reproducibility (*repr*), replicability (*repl*), and extension (*ext*) efforts across datasets.

Domain	Dataset	New Domain	New Dataset	Pref. Dev. Exploration	RS Experiment
Movies	ML	No	No	*repr + ext*	*repl + ext*
Music	MLHD	No	Yes	*repl + ext*	*repl + ext*
Books	BX	Yes	Yes	*repl + ext*	*repl + ext*

Table 2: Dataset description, including information on the number of child and mainstream users (MS) in the datasets.

Dataset	# Users	% Child	% MS	# Items	# Interactions
ML	6,040	3.68	81.82	3,706	1,000,209
BX	35,029	7.23	77.00	80,785	396,460
MLHD	44,349	8.21	79.89	1,918,414	1,055,574,094

We *reproduce* the *Preference Deviation Exploration* experiment with ML to probe the original analysis and verify if our study leads to the same results. Recall that we exclude LFM-2b as it is no longer available. Further, we *replicate* this experiment on two datasets new to this study: MLHD [45] as an alternative to LFM-2b to explore users' interactions with songs, and Book Crossing (BX) [50], which tracks book interactions—a domain not considered by the reference work. Each dataset was chosen as it includes demographic information about users and can be annotated with item genres, which are leveraged to determine preferences. Together, this enables juxtaposing deviating preferences in age groups across domains.

Inconsistencies in findings of the *RS experiment* and insights limited to a restricted period in LFM-2b prompt replication of this experiment on ML, MLHD, and BX. This allows probing of the stability of outcomes and the effect of mainstream users on recommendations for children across various datasets and domains.

Genre consumption behaviour is *one* aspect that may trigger differences across user groups; the reference work indicates that other consumption characteristics could be distinguishing factors for varying RA behavior between children and mainstream users. As popularity oftentimes affects recommendation quality [2, 3, 21, 23], we explore popularity as a potential factor contributing to deviating RA behavior. To do so, we *extend* the reference framework with a *popularity extension* in both experiments. For the *Preference Deviation Exploration*, we analyze the popularity of items consumed by users of different ages to assess the degree to which users consume items that are (1) overall popular, or (2) popular among users of the same age range. In the *RS experiment*, we gauge if RAs amplify popularity in recommendations compared to users' profiles.

4 Experimental Setup

We discuss the datasets and set up. We explicitly highlight adaptations made to the reference work required to facilitate our study.

Datasets. As in [42], we preprocess the **datasets** as described below and extend them with genre information; we only consider users aged 12 to 65, removing interactions with items with no assigned genre and users without valid age information. An overview of the resulting datasets is presented in Table 2 and Fig. 1.

MovieLens-1M (ML) [15] includes ratings for movies, where each movie is annotated with at least one of 18 genres that we assign equal weighs; users are assigned to one of seven age groups. As per the reference work, we treat users with the label 'Under 18' as children, users aged 18 to 49 as mainstream users, and older users as NMA. Only 2.83% of ratings can be attributed to children and 84.60% to mainstream users.

MLHD [45] includes 27 billion interactions with tracks, gathered from Last.fm (https://www.last.fm/). We utilize MLHD+[1], an improved version of the dataset that allows simple matching to information provided by MusicBrainz. We follow common practice regarding sampling large datasets for experimentation [26, 27], and select a user sample of the dataset. To capture users with consistent interactions over an extended period, we select those whose first recorded interaction occurred in or before 2009 and the last in or after 2013. This five-year window captures the period during which most interactions occurred, as well as the most recent span where user activity was consistently recorded. In line with [22], we exclude users with an unusually high number of interactions, i.e., those whose number of recorded interactions exceeds the mean by more than two standard deviations in the dataset. From the remaining users, we randomly sample 45,000 users—a comparable number of users to those considered in the reference work for the music-related dataset—while preserving the age distribution of the unfiltered dataset to maintain its original demographic structure.

Songs in MLHD are not linked to a genre. Hence, we annotate each song based on artist genres from Allmusic, as done for LFM-2b[2]. For this, we extract artist genres using the MusicBrainz API[3] and match these fine-grained genres with Allmusic genres, annotating artists with at least one of the 20 genres. In line with the reference work, we assume that the genre distribution of each artist extends to their tracks and annotate each song with equally weighted genres.

As the age information is "the age returned by the system at the moment of the data collection (i.e., circa 2013 and 2014)"[4], we assume that each user turned the reported age on January 1st, 2014. This enables us to assess each user's age for each interaction. For age groupings, we follow those used in LFM-2b in the reference work, as MLHD exhibits a similar age distribution: mainstream users are users aged 17 to 29, as this age range accounts for the vast majority of interactions in the dataset. Although 17-year-olds are legally considered children per UNICEF's definition [43], they cannot be treated as a minority in this context due to their high representation in the data. Consequently, we categorize users *under* 17 as children. Users older than 29 are referred to as NMAs. Most interactions in the dataset are from mainstream users (78.89%), while only a 10.80% of interactions can be attributed to children.

Book Crossing (BX)[5] [50] contains over 1 million user-book interactions. As books in BX lack genre annotations, we combine BX with the Goodreads dataset[6] [47, 48], which links books with at least one out of 8 genres. Books are assigned different ISBNs depending on their editions; thus, we turn to the bookdata tool (3.0) [9] to obtain ISBN variations for each book in Goodreads and BX, which we use for merging purposes. We assign equal weighting

[1] https://musicbrainz.org/doc/MLHD+

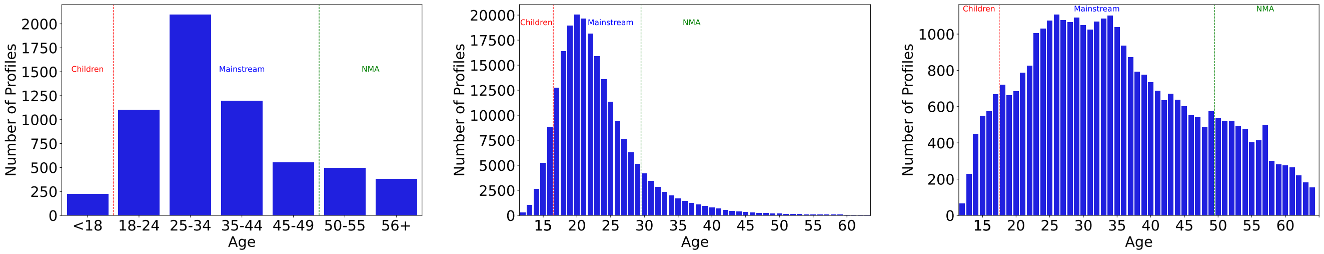
[2] Allmusic updated their genres (https://www.allmusic.com/genres). For comparability to the original study, we use the genres used in [37].

[3] https://musicbrainz.org/doc/MusicBrainz_API#Lookups

[4] https://ddmal.music.mcgill.ca/research/The_Music_Listening_Histories_Dataset_(MLHD)/

[5] https://www.kaggle.com/datasets/syedjaferk/book-crossing-dataset

[6] https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html



(a) Users per age in ML.

(b) Users per age in MLHD.

(c) Users per age in BX.

Figure 1: Size comparisons of the datasets.

to each genre a book is annotated with. Age distributions are akin to those on ML. Thus, we also treat users with age annotated < 18 as `children`, users aged 18 to 49 as `mainstream`, and older users as `NMA`s. On BX, only 2.88% of interactions can be attributed to `children` and 81.29% to `mainstream users`.

Preference Deviation Exploration. We create *user profiles* that capture all items a user consumed throughout a year of their lives. As BX and ML do not include timestamps, we assume that each interaction is associated with the reported age of the user, resulting in one user profile per user; for MLHD, a yearly user profile is created for each year in which a user interacted with tracks.

Genre preferences within user profiles are analyzed using **User Genre Profiles** (*UGP*s), which represent the mean frequency of each genre in a user's profile, accounting for genre weightings [37]. *UGP*s are used as a proxy to assess differences in preferences by accounting for varying consumption patterns. Multiple consumptions of the same item are considered to account for repeated interactions. To gain insights into broader consumption patterns of age groups, **Age Genre Profiles** AGP_{age} represent the average genre consumption of users within a particular age bucket, denoted with *age*. We quantify differences in genre consumption with:

- **In-group Deviation** (IGD_{age}) assesses how much profiles within an age group deviate from each other by measuring the average Jensen-Shannon Divergence (JSD) between an AGP_{age} and each *UGP* of users of the same *age*. This metric is analogous to a standard deviation as it measures the average distance of genre distributions of user profiles to the average distribution across all users in the respective age group.
- **Age Preference Deviation** ($APD_{age1,age2}$) measures the JSD between two *AGP*s.

The reference work compared significant differences between the genre consumption of `children` and `mainstream users` using a MANOVA and matching post-hoc tests. This approach relies on assumptions such as multivariate normality and homogeneity of variance, which may not be fully met in the case of normalized genre proportions. To better accommodate the nature of the data, we adopt a non-parametric approach and use Kruskal–Wallis tests to assess overall group differences per genre, followed by pairwise Mann–Whitney U tests with multiple testing correction ($p < 0.01$).

Popularity Extension. To study the impact of popularity bias between users, we extend the reference framework with:

- **# Interactions:** The number of items a user has consumed.
- **Profile Size:** The number of distinct items in a user profile [23].
- **Profile Popularity:** The average popularity of items in a user's profile, where an item's popularity is defined as the number of users who interacted with it, normalized by the most interacted-with item. This captures a user's tendency to consume items that are broadly popular across the population.
- **Profile Age-Popularity:** The average age-specific popularity of items (the normalized number of same-age users who interacted with each item) in a profile. This captures a user's preferences for items popular among peers in their age group.

RS Experiment. To prepare the data for this experiment, the reference work applies common preprocessing steps associated with music-RS explorations: Restricting data to a specific timeframe, removing one-time listens to avoid spurious interactions [27, 31], binarizing data by only including the first listening event [4], and applying k-core filtering to reduce sparsity in the data [4]. Further, a temporal global split is used [7, 17, 24] and users who lack items in any of the splits are removed. The training set resulting from this is the `General Set`; a subset only including `child` interactions forms the `Child Set`. We probe the **top-**50 **recommendations** created by two unpersonalized baselines—`Random` and `MostPop`—and two personalized RAs—$\mathrm{RP}^3\beta$ [33] and `iALS` [16].

To assess alignment between recommendations and user preferences, users' genre consumptions are captured by *UGP*s as defined previously; here, however, only interactions from the training sets are considered as these are the ones available to the RA. *AGP*s are computed based on these *UGP*s. In line with *UGP*s, **Recommendation Genre Profiles** (*RGP*s) model the average genre distribution of items recommended to a user. To gauge RA performance, traditional performance metrics *nDCG*, *MRR*, *MAP* [11, 13, 41] are used in addition to **Genre Miscalibration** [40]—computed as the JSD between a *UGP* and the respective *RGP*—is computed.

For the **popularity extension**, we add **Popularity Lift** (*PL*): The normalized difference between the Profile Popularity and the average popularity of items in recommendations for a user [2, 23].

For replication across datasets with different sizes and available metadata, we carefully design our setup to ensure comparability with the reference work while accounting for the datasets' unique properties. As the datasets stem from widely different domains, most preprocessing steps are not universally applicable; hence, we discuss details and reasoning for deviations from the original setup.

We can follow the reference framework closely for the *RS experiment* using MLHD. However, due to the size of MLHD, we use a representative subset for the recommendation experiments—a common approach in music recommendation experiments [23]. We first create a subset of 13,000 users who interacted with at least 5 tracks while ensuring that we retain the original age distribution. Then, we restrict our data to the same months as in the reference work (June to October) to avoid differences through seasonal effects. We set the year 2009 for our exploration due to a more reliable number of young users within this year. We retain other preprocessing steps: We only consider user-song interactions where the user has listened to a song at least twice, and we binarize ratings. We remove items with fewer than 10 and users with fewer than 5 interactions. We split by using June to August for training, September for validation, and October for testing. Users lacking items in any of the splits are removed, resulting in a set of 10,325 users with 97,322 items. The `Child Set` includes 1,878 users and 81,674 items.

As interactions with movies and books differ markedly from interactions with music (particularly in the number of consumed items), we adapt our processing steps for ML and BX. For ML, we follow common practices [4] of binarizing the data by only treating ratings > 3 as positive signals, while we keep all interactions for BX as positive signals, as it includes explicit as well as implicit interactions. To reduce sparsity, we follow [4] by applying iterative k-core with $k = 10$ for ML; as users tend to interact with fewer items on BX, we set $k = 5$ here. For BX (due to the lack of timestamps) and ML (due to the short available timeframe [15]), a global temporal split is not applicable. Thus, we split users' interactions into 60% training, 20% validation, and 20% test data. This results in a set of 5,949 users with 2,810 items for ML, with a `Child Set` of 218 users and 1,802 items. The filtered set of BX includes 6,950 users and 16,477 items, and a `Child Set` of 266 users and 2,196 items.

RS Explorations & Hyperparamenter Tuning. We use the Elliot framework for RS explorations and tune the hyperparameters following the original study and [4]. Besides aforementioned metrics, the reference work explored additional metrics to gauge the deviation of *UGP*s and *RGP*s to `mainstream` and `child` *AGP*s. We bypass reporting these results as our focus is on *direct comparisons* between age groups. However, we provide results for these metrics in our Git repository for transparency and completeness.

5 Results

Here, we lay out the outcomes from both experiments.

Preference Deviation Exploration. Fig. 2 shows the *AGP*s of age groups and the *APD* between these. On ML, `children` deviate from all other age groups, with an $APD_{\texttt{child,mainstream}} = 0.013$. Except for the genres `Adventure`, `Horror`, and `SciFi`, the proportion of all genres is significantly different between `children` and `mainstream users`. On MLHD, findings are comparable to insights from the music-related dataset, LFM-2b, in the reference work. While children's preferences are similar, the older a user gets, the closer their preferences align with those of `mainstream users`. Overall, the $APD_{\texttt{child,mainstream}} = 0.0062$, and there are significant differences in the frequencies of all genres except `Electronic` and `Rock` between `children` and `mainstream users`.

As in MLHD, the older a user gets in BX, the closer their preferences align with `mainstream` preferences. However, here, 12-year-olds also stand out as deviating more strongly from children of other ages. For instance, 16-year-olds' preferences align more closely with `mainstream users` than with those of 12-year-olds. Turning to Fig. 2c, it can be seen that particularly the `Children` genre becomes markedly more prominent for `children` than any other group. Overall, with $APD_{\texttt{child,mainstream}} = 0.071$, there are significant differences between `children`'s and `mainstream users`'s genre consumption of all genres except `Fantasy/Paranormal`.

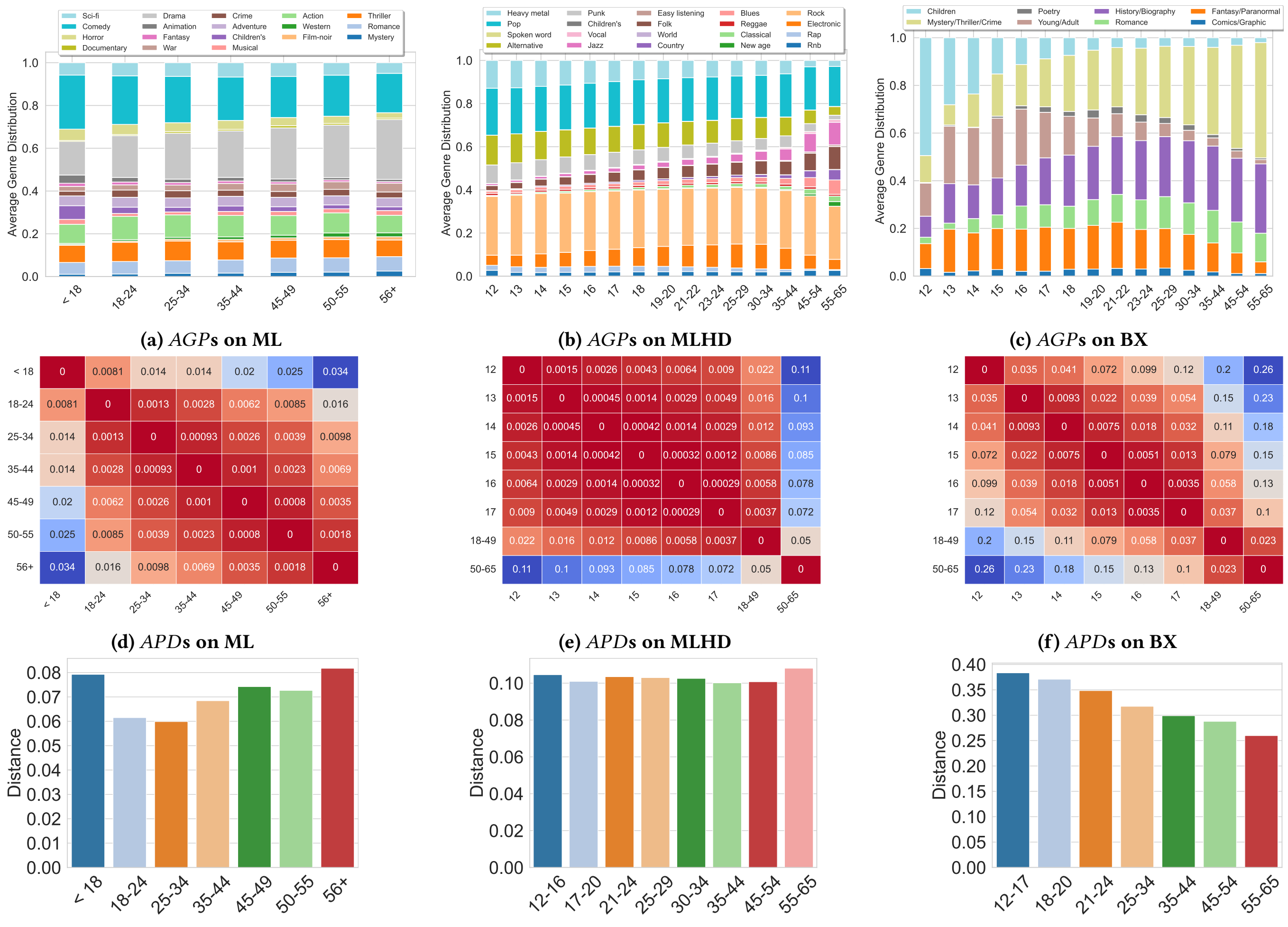
Fig. 2 also shows the *IGD*s across the ages in all three datasets. While no clear differences can be seen between age groups captured by MLHD, results on ML show that while children have a comparably high *IGD*, and the youngest `mainstream` users (18-24) have the lowest value, *IGD* increases the older a user gets. On BX, we find `children` have the highest *IGD*, which decreases with age. Results from the *popularity extension* (Table 3) show consistent trends between ML and BX. `Children` and `NMAs` interact with fewer items, on average, than `mainstream users`, which are also low in popularity. For BX, `children` interact more frequently with items that are popular among their age group. Contrarily, on MLHD, `children` track more listening events than other age groups, but with fewer distinct songs, i.e., they tend to listen repeatedly to a smaller set of items. Further, the items that they interact with are overall more popular among the entire population.

RS Experiment. Our analysis of the quality of recommendations between age groups when trained on the `General Sets` of the datasets yields salient differences between datasets and age groups (see Table 4). On ML, for `MostPop`, `RP`$^3\beta$, and `iALS`, `mainstream users` stand out as a user group that commonly receives better recommendations than `children` or `NMAs`: Most performance metrics as well as genre calibration scores are best for this user group. Metric scores for `children` and `NMAs` are not significantly different across all metrics and algorithms. On BX, performance metrics differences between `children` and `mainstream users` are non-significant across `MostPop`, `RP`$^3\beta$, and `iALS`. `NMAs` stand out with usually significantly worse performance scores. Interestingly, as per *GMC*, `NMAs` receive the best-aligned recommendations, whereas `children` get the worst across all RAs. For both datasets, `mainstream users` stand out as the user group that receives well-aligned recommendations across metrics and RAs. With *GMC* on BX being an exception, recommendations for `mainstream users` are either significantly better than those of other user groups, or differences to other user

Table 3: Results of *popularity extension* per age group[a].

		`children`	`mainstream`	`NMA`
ML	# Interactions/Profile Size	122.568^{m}	$174.368^{c,n}$	127.021^{m}
	Profile Popularity	0.263^{m}	$0.282^{c,n}$	0.260^{m}
	Profile Age-Popularity	0.289^{n}	0.287^{n}	$0.255^{c,m}$
MLHD	# Interactions	$6341.105^{m,n}$	$4758.395^{c,n}$	$4174.021^{c,m}$
	Profile Size	$1244.141^{m,n}$	$1461.833^{c,n}$	$1814.947^{c,m}$
	Profile Popularity	$0.079^{m,n}$	$0.064^{c,n}$	$0.046^{c,m}$
	Profile Age-Popularity	$0.067^{m,n}$	$0.060^{c,n}$	$0.060^{c,m}$
BX	# Interactions/Profile Size	$4.508^{m,n}$	11.949^{c}	11.360^{c}
	Profile Popularity	$0.057^{m,n}$	$0.071^{c,n}$	$0.076^{c,m}$
	Profile Age-Popularity	$0.111^{m,n}$	$0.074^{c,n}$	$0.081^{c,m}$

[a] Significant differences between two groups ($p < 0.01$) are annotated with the corresponding pair (`children` (c), `mainstream` (m), `NMAs` (n)). Note that there are no repeated interactions tracked in ML and BX. Thus, for these datasets, the number of unique interactions (profile size) corresponds to the number of all interactions.



(a) *AGP*s **on ML**

(b) *AGP*s **on MLHD**

(c) *AGP*s **on BX**

(d) *APD*s **on ML**

(e) *APD*s **on MLHD**

(f) *APD*s **on BX**

(g) *IGD* **on ML**

(h) *IGD* **on MLHD**

(i) *IGD* **on BX**

Figure 2: *AGP*, *GMA*, **and** *IGD* **across age groups in different datasets.**

groups are non-significant. In other words, mainstream users, on average, never receive significantly worse recommendations than any other user group. Children's recommendations are commonly less accurate or less well aligned in terms of *GMC*. These insights contradict findings on MLHD. For Random, no significant differences are found (except higher *GMC* for children than for mainstream users); for the other RAs children receive **better** recommendations than other age groups, with *GMC* on MostPop being the exception. In addition, results for metrics between mainstream users and NMAs are mostly non-significant.

Across all datasets MostPop, RP$^3\beta$, and iALS lead to a positive popularity lift, i.e., higher average popularity in the recommendations than in the user profiles. Still, there is diverging behavior between MLHD and the other two datasets. On ML and BX, popularity lift tends to be similar or even higher for children than for other user groups. On MLHD, in contrast, the popularity lift for children tends to be lower than for mainstream users.

Turning our analysis to the differences in recommendation quality for children when trained on the General Set versus the Child Set, we observe that for the personalized RAs training on the Child Set typically leads to **worse** recommendations on ML and BX. On MostPop, no significant changes are found on these datasets except some improvements in terms of *GMC* when trained on the Child Set, and improved *nDCG* on ML.

Similarly, on MLHD, differences between *MRR* and *MAP* scores are non-significant when trained on either set. However, training on the Child Set leads to *worse nDCG* scores and *GMC* for RP$^3\beta$ and iALS. In contrast, recommendation quality increases when trained on the Child Set for MostPop: performance scores are lower; *GMC* increases. For all datasets training on the Child Set does not affect performance scores of Random, but it improves the *GMC*. In terms of popularity, for all RAs except Random, training on the Child Set consistently leads to lower popularity lift for children on ML and BX, and higher popularity lift on MLHD.

6 Discussion

We discuss the obtained results and compare outcomes to the reference work, highlighting replicated as well as deviating findings.

Table 4: Average metrics per age group based on the *RS Experiment*^{b}.

			ML					MLHD					BX				
	Data	Age Group	nDCG$^{\uparrow}$	MRR$^{\uparrow}$	MAP$^{\uparrow}$	$GMC^{\downarrow}$	$PL^{\rightarrow 0}$	nDCG$^{\uparrow}$	MRR$^{\uparrow}$	MAP$^{\uparrow}$	$GMC^{\downarrow}$	$PL^{\rightarrow 0}$	nDCG$^{\uparrow}$	MRR$^{\uparrow}$	MAP$^{\uparrow}$	$GMC^{\downarrow}$	$PL^{\rightarrow 0}$
Random	Child Set	Children	0.018	0.039	0.033	0.141^{*}	-0.557^{*}	0.000	0.001	0.001	0.124^{*}	-0.662^{*}	0.004	0.007	0.007	0.261^{*}	0.165^{*}
	General Set	Children	0.013	0.024	0.024	$0.152^{m,n}$	-0.685^{m}	0.000	0.001	0.001	0.127^{m}	$-0.703^{m,n}$	$0.005^{m,n}$	0.003	0.003	$0.293^{m,n}$	$-0.443^{m,n}$
		Mainstream	0.013^{n}	0.033^{n}	0.030^{n}	$0.123^{c,n}$	$-0.738^{c,n}$	0.000	0.001	0.001	0.121^{c}	$-0.650^{c,n}$	0.002^{c}	0.002	0.002	0.212^{c}	$-0.547^{m,n}$
		NMAs	0.010^{m}	0.023^{m}	0.019^{m}	$0.137^{c,m}$	-0.698^{m}	0.000	0.000	0.000	0.126	$-0.596^{c,m}$	0.001^{c}	0.002	0.002	0.217^{c}	-0.564^{c}
MostPop	Child Set	Children	0.152^{*}	0.237	0.163	0.140^{*}	1.410^{*}	0.013^{*}	0.034^{*}	0.027^{*}	0.136^{*}	5.690^{*}	0.025	0.023	0.020	0.261^{*}	3.509^{*}
	General Set	Children	0.129^{m}	0.208^{m}	0.146^{m}	0.151^{m}	1.768^{m}	$0.011^{m,n}$	0.026^{n}	$0.024^{m,n}$	0.139	$6.429^{m,n}$	0.020	0.023	0.021	$0.308^{m,n}$	$7.961^{m,n}$
		Mainstream	$0.174^{c,n}$	$0.296^{c,n}$	$0.196^{c,n}$	$0.120^{c,n}$	$1.263^{c,n}$	0.008^{c}	0.020	$0.018^{c,n}$	0.135^{n}	$7.751^{c,n}$	0.030^{n}	0.044^{n}	0.037^{n}	$0.199^{c,n}$	6.220^{c}
		NMAs	0.126^{m}	0.213^{m}	0.147^{m}	0.147^{m}	1.594^{m}	0.005^{c}	0.011^{c}	$0.010^{c,m}$	0.146^{m}	$9.229^{c,m}$	0.021^{m}	0.025^{m}	0.022^{m}	$0.181^{c,m}$	5.864^{c}
$RP^3\beta$	Child Set	Children	0.222^{*}	0.340^{*}	0.215^{*}	0.084^{*}	0.769^{*}	0.039^{*}	0.085	0.067	0.059^{*}	0.709^{*}	0.028^{*}	0.035^{*}	0.031^{*}	0.211^{*}	0.144^{*}
	General Set	Children	0.287	0.418^{m}	0.265^{m}	0.062	0.543	$0.043^{m,n}$	$0.083^{m,n}$	$0.069^{m,n}$	$0.053^{m,n}$	$0.601^{m,n}$	0.103^{n}	0.125	0.111^{n}	$0.169^{m,n}$	0.379
		Mainstream	0.308^{n}	$0.477^{c,n}$	$0.296^{c,n}$	0.059^{n}	0.593^{n}	0.033^{c}	0.066^{c}	0.053^{c}	0.060^{c}	$0.822^{c,n}$	0.083^{n}	0.125^{n}	0.098^{n}	$0.119^{c,n}$	0.456
		NMAs	0.281^{m}	0.435^{m}	0.277^{m}	0.063^{m}	0.542^{m}	0.030^{c}	0.051^{c}	0.045^{c}	0.062^{c}	$1.183^{c,m}$	$0.063^{c,m}$	0.083^{m}	$0.067^{c,m}$	$0.102^{c,m}$	0.455
iALS	Child Set	Children	0.197^{*}	0.334	0.203^{*}	0.073^{*}	0.214^{*}	0.033^{*}	0.067	0.060	0.060^{*}	1.811^{*}	0.034^{*}	0.045^{*}	0.043^{*}	0.210^{*}	0.508^{*}
	General Set	Children	0.292^{m}	0.400^{m}	0.246^{m}	0.054^{m}	0.377^{m}	$0.038^{m,n}$	$0.082^{m,n}$	$0.065^{m,n}$	$0.042^{m,n}$	$0.826^{m,n}$	0.106^{n}	0.123	0.105	$0.160^{m,n}$	1.242^{m}
		Mainstream	$0.322^{c,n}$	$0.481^{c,n}$	$0.295^{c,n}$	$0.047^{c,n}$	$0.311^{c,n}$	0.030^{c}	0.063^{c}	0.050^{c}	0.050^{c}	$1.083^{c,n}$	0.080^{n}	0.112^{n}	0.092^{n}	$0.115^{c,n}$	0.926^{c}
		NMAs	0.302^{m}	0.449^{m}	0.272^{m}	0.055^{m}	0.363^{m}	0.026^{c}	0.047^{c}	0.041^{c}	0.052^{c}	$1.278^{m,n}$	$0.060^{c,m}$	0.081^{m}	0.069^{m}	$0.096^{m,n}$	1.006

[b] Significant differences between two groups ($p < 0.01$) are annotated with the corresponding pair (children (c), mainstream (m), NMAs (n)). An asterisk (*) on a Child Set row denotes significant differences in the recommendations for children between the Child Set and the General Set.

Children as Non-Mainstream Users. The reference work uncovered differences in preferences, assessed through genre consumption behavior, which this work amplifies, painting a broader picture. There are clear differences in genre consumption across ages in all datasets. Findings on ML agree with those in [42], which was anticipated given the direct reproduction of the *Preference Deviation Exploration*. Results obtained on MLHD resemble those on LFM-2b in the reference work, i.e., the proportions of genres in *AGP*s align closely; trends in *APD* and *IGD* are similar, highlighting that preference deviations in the original study are reflective of music interactions more generally. Nonetheless, as both datasets are based on interactions from *Last.fm*, key characteristics of users in both datasets will mainly be reflective of the platform's users. Our analysis on BX uncovers findings on a dataset and even a domain unexplored in the reference work. Deviations from children to the mainstream are particularly noticeable here, and younger users show higher preference deviations within their age groups (Fig. 2i).

While the key trend of the reference work—namely, salient differences in genre preferences between children and mainstream users—is confirmed by our work, the *popularity extension* highlights distinctions between the domains examined. In ML and BX, children consume fewer items and particularly less popular ones than those consumed by mainstream users. In the music domain, informed by results from MLHD, children prefer fewer items that are popular in general, which they listen to repeatedly. This finding emphasizes the need for domain-specific considerations when assessing preferences of underrepresented user groups, as generalizations across domains may obscure important nuances in user preferences. Differences in preferences highlight that children differ in key preferences from adult users. While deviations are even more severe to NMAs, the deviations to mainstream users highlight a potential oversight of current RS research: The datasets used to uncover preferences, interaction patterns, or probe how well a system fares are mainly based on mainstream users as these are the most prominent user group (see Fig. 1). However, other user groups such as children—the main focus of the reference work and this study—may deviate and thus be overlooked.

Dominance of User Groups and Deviating RA Behavior. Key differences in behavior between age groups, identified in the reference work and confirmed by our findings, raise the question of whether RAs can capture these distinctions, particularly considering that mainstream users dominate the data used to train such algorithms. The need to thoroughly explore this concern with the *RS experiment* is exacerbated by the discovery of previously laid out domain-related differences in consumption patterns of children.

RA performance is directly affected by these differences: The tendency from the reference work that children receive recommendations by personalized RAs that are as good or even better than those of mainstream users is in line with outcomes of the replicated experiment on MLHD. However, outcomes from the experiments on ML and BX show that mainstream users mostly receive accurate and well-aligned recommendations, often to the detriment of recommendation quality for children and NMAs. This contrast may be explained by the tendency of children to prefer popular items in the music domain and not in others. On ML, MostPop performs significantly worse for children than for mainstream users according to all metrics, and on BX, *GMC* scores are markedly higher for children, indicating that genres of popular recommendations do not align with their preferences. On MLHD, on the other hand, recommendations by MostPop are, depending on the metric, highly suitable for children, matching outcomes of the reference work on LFM-2b. This preference for popular items on MLHD is reflected by the popularity extension, where children exhibit less popularity lift than other user groups on MLHD, but not on others. Due to their preference for already popular items, children are impacted less by the popularity bias of the RAs.

Differences between these datasets highlight the importance of research approaches that acknowledge different user groups in different domains. In domains like music, where children prefer popular music and engage with it more intensively, popularity-biased recommenders may incidentally perform better for this group than for others. In contrast, in domains like movies or books, where children prefer less popular items, systems trained on mainstream data may struggle to produce equally aligned recommendations for children. Good performance in one domain, as highlighted in the reference work, does not necessarily translate to another, making multi-domain perspectives crucial if RS research truly attempts to acknowledge and serve underrepresented users like children.

The reference work emphasized inconsistent behavior between RAs. RP$^3\beta$ leveraged data from other groups to better match children's preferences (i.e., RP$^3\beta$ did not perform as well on Child Set). In contrast, iALS benefited from the focus on children's interactions in the Child Set, yielding similar performance and improved *GMC* [42]. Such inconsistencies cannot be found in our study. Instead, RP$^3\beta$ and iALS both fare significantly less well for children when trained on the Child Set across all datasets on most metrics. This may indicate one out of two reasons: First, as assumed by the reference work, children are indeed 'difficult users' [34], users that are more challenging to recommend to due to deviating preferences; higher *IGD*s (Fig. 2), particularly on ML and BX, support this assumption as such measures can be used to determine difficult users [5, 34]. To recommend suitable items to them, an RA may require additional information; RAs utilized in our study can leverage mainstream users' data to create recommendations that align better with what a child likes. Second, as the approach developed by the reference work to gauge the effect of mainstream users on recommendations for children leads to a reduction of training data available to the RA, it reduces the number of interactions that can be leveraged to create fitting recommendations. This may affect RA performance negatively.

MostPop leads to no significant differences and, for some metrics, improvement for children when training on the Child Set. Recommending items popular *amongst* children instead of among all users can be effective to serve this group. However, this narrow focus on popular items may have other downsides, potentially reducing fairness, diversity, or novelty in recommendations [1, 19].

Reproducibility Concerns. We reflect on the changes made to the original setup to enable our experiments. A majority of datasets used by the RS research community does not include age-related information, and if they do, information is coupled with uncertainty: Neither of the datasets used in this study provides entirely accurate information about when age-related data was collected. Thus, the data available serves as the best proxy of users' age.

As in Ungruh et al. [42], we focused on datasets from entertainment domains. In these, we assume that children are mostly 'free' to choose what to listen to, read, or watch. We assume that this is reflected by the data available. However, in other domains, where they may have less agency (e.g., e-commerce, education, or tourism), this assumption may not hold. Therefore, the reference framework may be limited to studies in comparable domains.

The datasets selected are the only ones that, to our knowledge, include demographic information; yet it is not always possible to adopt them 'as is' to the reference setup. BX and MLHD do not provide genre information, requiring external datasets and APIs for genre annotations. This led to the exclusion of numerous items without reliable genre information. Although we ensured a sufficiently large and balanced set of items across age groups, removing unannotated items raises concerns, particularly as these items may be central to capturing unique or niche preferences. Further, while genre distributions are suitable to measure preferences in the domains studied, simplifications (such as assuming that artists' genres reflect to each song) may not capture all nuances, and genre equivalents may not exist to sufficiently compare users' preferences in other domains like tourism. Properties of some datasets, such as datasets being too large to handle efficiently, non-sequential structure of the data, or unavailable timestamps, limited our ability to fully mimic the original setup. We adopted alternative strategies grounded in common practices from prior RS studies [4, 23] and recognize that such deviations showcase challenges of consistent evaluation of datasets with differing structures.

7 Conclusion & Future Work

In this work, we reproduced, replicated, and extended the findings of Ungruh et al. [42], focused on children as a non-mainstream user group. By broadening the analysis across datasets and domains, we confirmed key trends and uncovered new insights into the interplay between children and RS. Of note, despite children being a minority, RAs can perform well for this user group, and the dominant interactions of the mainstream do not necessarily impact children negatively. Yet, this effect is not conclusive across datasets, metrics, and RAs. In fact, our study spotlights that children are indeed a non-mainstream user group with preferences deviating from those of the mainstream for which RAs may fail to account for.

Even if a RA provides suggestions that align with child preferences, this does not mean that the recommendations are necessarily 'good' for them. They might still fail in other relevant aspects for children such as age-appropriateness [29]. Building on the reference work, we provide insights about how current RAs fare when it comes to children. Still, insights are limited by the small number of systems studied, the narrow focus of quality metrics, and a simplified view of children as a homogeneous group. Therefore, we see emerging lessons learned as a foundation to move toward child-aware RS, ones that recognize children as part of the audience and strive to provide recommendations that are not only accurate, but recommend items truly *fitting* to their users. Future research should further leverage the presented extended framework to other non-mainstream user groups (e.g., older adults, niche-interest users, or culturally marginalized communities) to understand how RS can better fare for *all* users, not just the statistical majority.

Acknowledgments

This research was supported by the Austrian FFG COMET program. Further, it was supported by grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033. We thank Michael Ekstrand, Rares Boza, and Gabriel Vigliensoni for general technical support and insights into the datasets.

References

[1] Himan Abdollahpouri and Robin Burke. 2021. Multistakeholder recommender systems. In *Recommender systems handbook*. Springer, 647–677.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *RMSE workshop held in conjunction with the 13th ACM Conference on Recommender Systems* (2019).

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in

recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 119–129.

[4] 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.

[5] Alejandro Bellogín. 2011. Predicting performance in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 371–374.

[6] Alejandro Bellogín and Alan Said. 2021. Improving accountability in recommender systems research through reproducibility. *User Modeling and User-Adapted Interaction* 31, 5 (2021), 941–977.

[7] Pedro G Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction* 24 (2014), 67–119.

[8] 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 *Conference on fairness, accountability and transparency*. PMLR, 172–186.

[9] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*.

[10] Farzad Eskandanian, Nasim Sonboli, and Bamshad Mobasher. 2019. Power of the few: Analyzing the impact of influential users in collaborative recommender systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 225–233.

[11] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021).

[12] 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.

[13] Asela Gunawardana, Guy Shani, and Sivan Yogev. 2022. Evaluating Recommender Systems. In *Recommender Systems Handbook: Third Edition*. Springer US, 547–601.

[14] David J Hargreaves, Adrian C North, and Mark Tarrant. 2015. How and why do musical preferences change in childhood and adolescence. *The child as musician: A handbook of musical development* (2015), 303–322.

[15] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015).

[16] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *8th IEEE international conference on data mining*. 263–272.

[17] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2023. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems* 41, 3 (2023), 1–27.

[18] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2025. Unmasking Gender Bias in Recommendation Systems and Enhancing Category-Aware Fairness. In *Proceedings of the ACM on Web Conference 2025*. 5127–5138.

[19] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction* (2024), 1–58.

[20] Ahmet Sami Konca. 2022. Digital technology usage of young children: Screen time and families. *Early Childhood Education Journal* 50, 7 (2022), 1097–1108.

[21] Dominik Kowald and Emanuel Lacic. 2022. Popularity bias in collaborative filtering-based multimedia recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–11.

[22] Dominik Kowald, Peter Muellner, Eva Zangerle, Christine Bauer, Markus Schedl, and Elisabeth Lex. 2021. Support the underground: characteristics of beyond-mainstream music listeners. *EPJ Data Science* 10, 1 (2021), 14.

[23] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 35–42.

[24] Dominik Kowald, Paul Seitlinger, Simone Kopeinik, Tobias Ley, and Christoph Trattner. 2013. Forgetting the words but remembering the meaning: Modeling forgetting in a verbal and semantic tag recommender. In *International Workshop on Mining Ubiquitous and Social Environments*. Springer, 75–95.

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

[26] 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.

[27] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing item popularity bias of music recommender systems: are different genders equally affected?. In *Proceedings of the 15th ACM conference on recommender systems*. 601–606.

[28] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. 2021. Leave no user behind: Towards improving the utility of recommender systems for non-mainstream users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 103–111.

[29] Sonia Livingstone and Kim R Sylwander. 2025. There is no right age! The search for age-appropriate ways to support children's digital lives and rights. *Journal of Children and Media* 19, 1 (2025), 6–12.

[30] Masoud Mansoury, Himan Abdollahpouri, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. 2020. Investigating potential factors associated with gender discrimination in collaborative recommender systems. In *The thirty-third international flairs conference*.

[31] Alessandro B Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666.

[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] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2016. Updatable, accurate, diverse, and scalable recommendations for interactive applications. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016).

[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] 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)*.

[36] 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.

[37] 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.

[38] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2021. Music recommendation systems: Techniques, use cases, and challenges. In *Recommender systems handbook*. Springer, 927–971.

[39] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine* (2025).

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

[41] Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. 2021. Quality metrics in recommender systems: Do we calculate metrics consistently?. In *Proceedings of the 15th ACM conference on recommender systems*. 708–713.

[42] 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, 67–84.

[43] UNICEF. 2019. The convention on the rights of the child: The children's version. https://www.unicef.org/child-rights-convention/convention-text-childrens-version

[44] Patti M Valkenburg and Joanne Cantor. 2001. The development of a child into a consumer. *Journal of Applied Developmental Psychology* 22, 1 (2001), 61–72.

[45] Gabriel Vigliensoni and Ichiro Fujinaga. 2017. The music listening histories dataset.. In *ISMIR*. 96–102.

[46] Clara Virós-Martín, Mireia Montaña-Blasco, and Mònika Jiménez-Morales. 2024. Can't stop scrolling! Adolescents' patterns of TikTok use and digital well-being self-perception. *Humanities and Social Sciences Communications* 11, 1 (2024).

[47] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*. 86–94.

[48] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. *arXiv preprint arXiv:1905.13416* (2019).

[49] Ziwei Zhu and James Caverlee. 2022. Fighting mainstream bias in recommender systems via local fine tuning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1497–1506.

[50] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *WWW'05*.