The Impact of Mainstream-Driven Algorithms on Recommendations for Children

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

DOI
10.1007/978-3-031-88714-7_5

Publication date
2025

Document Version
Final published version

Published in
Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Proceedings

The Impact of Mainstream-Driven Algorithms on Recommendations for Children

Robin Ungruh[1(✉)], Alejandro Bellogín[2], and Maria Soledad Pera[1]

[1] Delft University of Technology, Delft, The Netherlands
{R.Ungruh,M.S.Pera}@tudelft.nl
[2] Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain
alejandro.bellogin@uam.es

Abstract. Recommendation algorithms are often trained using data sources reflecting the interactions of a broad user base. As a result, the dominant preferences of the majority may overshadow those of other groups with unique interests. This is something performance analyses of recommendation algorithms typically fail to capture, prompting us to investigate how well recommendations align with preferences of the overall population but also specifically a "non-mainstream" user group: children—an audience frequently exposed to recommender systems but rarely prioritized. Using music and movie datasets, we examine the differences in genre preferences between Children and Mainstream Users. We then explore the degree to which (genre) consumption patterns of a mainstream group impact the recommendations classical algorithms offer children. Our findings highlight prominent differences in consumption patterns between Children and Mainstream Users; they also reflect that children's recommendations are impacted by the preference of user groups with deviating consumption habits. Surprisingly, despite being under-represented, children do not necessarily receive poorer recommendations. Further, our results demonstrate that tailoring training specifically to children does not always enhance personalization for them. These findings prompt reflections and discussion on how recommender systems can better meet the needs of understudied user groups.

Keywords: Recommender Systems · Children · Consumption Behavior

1 Introduction

Recommender Systems (RS), essential tools for personalizing online experiences, cater to users with diverse tastes and preferences. However, these systems are typically designed and evaluated across broad demographic categories, overlooking the distinct characteristics of specific groups [12,33]. Popular datasets used in RS research reflect a skewed view of user populations: $\sim 31\%$ and $\sim 28\%$ of users are female, and 8% and 4% are under 18 years old in LFM-2b [45] and MovieLens-1M [20], respectively. Such imbalances can lead RS to misrepresent the behavior or preferences of underrepresented groups, and consequently be less

C. Hauff et al. (Eds.): ECIR 2025, LNCS 15574, pp. 67–84, 2025.
https://doi.org/10.1007/978-3-031-88714-7_5

effective for these groups [2,13,25,26,33]. This is a known issue, especially for users with niche interests, as recommendation algorithms (RAs) tend to capture popular preferences more accurately than niche ones [16]. Ideally, RAs would consider users' individual preferences, but items favored by the majority receive more interactions and are consequently suggested more frequently—regardless of users' specific interests [13,24].

An often understudied group is children (individuals up to 18 years of age [51]). Despite their common use of online platforms [39,42] and exposure to RS, research rarely prioritizes them. As per preliminary studies based on short-term observations, we know that children are a distinct type of human [6] with unique interests [36,47]; and when it comes to RS their interaction and consumption patterns differ from those of adults [5,44]. These insights, however, are seldom reflected in the design and evaluation of RS explicitly for children [e.g., 17,41]. Furthermore, across popular platforms, RS are rarely tailored to specific user groups; they are deployed with a broad user range in mind. Consequently, children's interactions may be overshadowed by adults', who typically represent mainstream users—users responsible for a large majority of available data. Sidelining children's preferences could lead to a poor user experience and skewed recommendations. This uncovers a critical research gap for children (and other minority groups): the need for a comprehensive exploration that examines the preferences of diverse users in a nuanced way and based on behavior observed over an extended period of time.

To address this gap, we anchor our work in two research questions: **(RQ1)** *Do item preferences differ between* `children` *of varied ages and* `mainstream`*?* **(RQ2)** *Do common RAs suggest items to* `children` *that deviate from their initial preferences due to the dominance of profiles of* `mainstream`*?* To address these RQs, we study user interactions in two popular domains for RS deployment–movies and music–by utilizing the well-known LFM-2b [45] and MovieLens-1M [20] datasets. We conduct a two-phase empirical exploration where we assess the preferences of individuals based on consumption patterns of items of different genres, modeling and comparing the types of media that users prefer across ages. We then build on emerging findings to probe whether common RAs skew their recommendations because of the dominance of mainstream users, neglecting younger audiences. We recognize that children do not form a monolithic user group. However, to establish a foundational understanding of this user group, we follow accepted practices [13,35,37] and view children as a generalized user group while also examining sub-groups whenever the metadata in aforementioned datasets allows us to do so.

This work has a direct impact on the evaluation and deployment of child-aware RS and implications of interest for underrepresented groups: RAs assessment should consider not only their overall performance for their user base but also their ability to genuinely *serve* the interests of *all* users, including those whose niche behavior or preferences may be overshadowed by dominant profiles. Further, our analysis exposes the challenges of data pre-processing for offline evaluation. Adopting common pre-processing steps instead of being explicitly mindful of non-traditional user groups may affect their representation and, con-

sequently, mislead study outcomes for this population. To enable reproducibility, we publish all associated code in a public repository[1].

2 Related Work

Children are often overlooked in RS research, with most works in this area focusing on children's consumption patterns [12,17,26,38]. Existing research analyzes the characteristics of books read by children [15,36]; it also examines genre and song features in children's music consumption [44]. The efforts, however, are based on limited samples and timeframes that may not fully reflect children's actual interactions. More so, they discuss implications for the design of RS tailored specifically to children.

Turning to RS intended for a wide user range, research indicates that not all user groups are served equally well, with children potentially receiving less suitable recommendations according to lower performance scores [13]. Schedl and Bauer [44] also note that music RS designed for "general" users tend to perform worse for children, likely due to the influence of adult user profiles skewing recommendations. While traditional metrics like accuracy evaluate algorithmic performance, they overlook crucial factors such as user satisfaction, engagement, and the system's ability to accommodate diverse preferences [10,21]—key considerations for underrepresented groups.

RAs are effective if they suggest items in line with users' previous consumption behavior, fostering a sense of consistency and relevance that may better engage users over time [29]. Unfortunately, this is seldom the case for minority groups. Chaney et al. [10] note that RAs suffer from homogenization, where diverse users are treated similarly despite deviating consumption patterns. For instance, female users whose interactions with systems are underrepresented in datasets receive recommendations that are less accurate [13,32] and also deviate from their interactions [31]. This raises concerns about how children might be unequally served and unfairly treated by RAs.

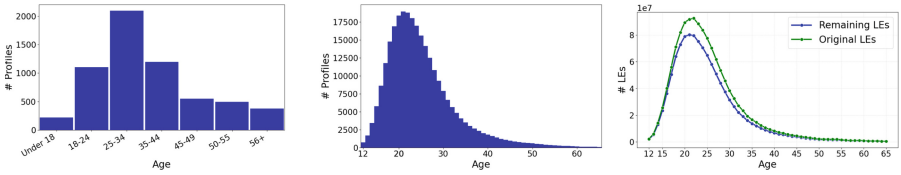
3 Analyzing Deviations Between Preferences

In phase one of the exploration, we analyze user interactions with items of different genres to scrutinize consumption patterns. We focus on the differences between `Children` and `Mainstream Users`—the latter responsible for a majority of recorded interactions despite many individuals of different ages who might use such systems.

3.1 Experiment Setup

We anchor our study on two datasets. **MovieLens 1M** [20], widely used in RS research, includes user demographics and consists of $1,000,209$ ratings of $3,706$

[1] https://github.com/rUngruh/PreferenceAnalysis.



(a) Users per age group in ML.

(b) Profiles per age at listening in LFM-2b.

(c) Impact of matching genres to songs on number of LEs

Fig. 1. Size comparisons of the datasets.

movies by 6,040 users (an average of 165.60 rated items per user). Each movie is annotated with at least one of 18 genres. As per dataset metadata, each user (and therefore their associated ratings and consumption patterns) belongs to one of 7 age groups (see distribution in Fig. 1a). Most ratings (84.60%) are recorded by users aged 18 to 49, which we treat as `mainstream`, given their overwhelming representation. `Children` (Under 18) only make up for 2.83% of ratings; the remaining ratings are from `Non-mainstream Adults (NMA)` aged 50+, with no further age distinctions provided for this group.

LFM-2b [34,45] is a large dataset with fine-grained user demographics and user-item interactions. It includes 2,014,164,872 Listening Events (LEs) from 120,322 users on 50,813,373 songs, recorded from February 2005 to March 2020. Based on metadata provided by the dataset creators, each user is associated with a self-reported age as of October 31, 2013. For simplicity, we assume that each user turned this age on this date, allowing us to estimate their age for each recorded LE with an error margin of ±1 year. Unlike ML, which only offers broad age categories, this detailed information allows us to pinpoint the age at which a user consumed a particular song, enabling more nuanced insights. We exclude users without a valid age, those under 12 (to account for social media age limits), or those over 65 (retirement age)–the latter exclusion as we study younger user groups. This results in 46,005 users and 1,337,596,535 LEs.

LFM-2b associates songs with 2,808 micro-genres. As these are too specific for meaningful comparisons, we instead utilize genre information from the LFM-1b UGP [46] dataset, which maps 219,022 artists to at least one out of 20 genres from Allmusic (https://www.allmusic.com/genres). We assume that the genre of an artist also extends to each of their songs, enabling us to identify the genre for 25,719,981 songs. By excluding songs without genre information, the dataset used for analysis includes 1,131,465,529 LEs by 45,601 users. Leveraging the dataset's long timespan and detailed user age information, we create yearly *user profiles* for each user. Each year starts on October 31st, and a single profile represents one year of a user's life, including only Listening Events (LEs) from that year. This results in 275,232 unique user profiles, the majority representing young adults (Fig. 1b). Maintaining the consumption behavior captured in the original data is key to guaranteeing the validity of our exploration. Therefore, prior to deciding on the adoption of a specific filtering step, we purposely gauge if certain user groups are disproportionally affected, as this would distort the representation of their preferences. Based on the proportion of removed LEs by

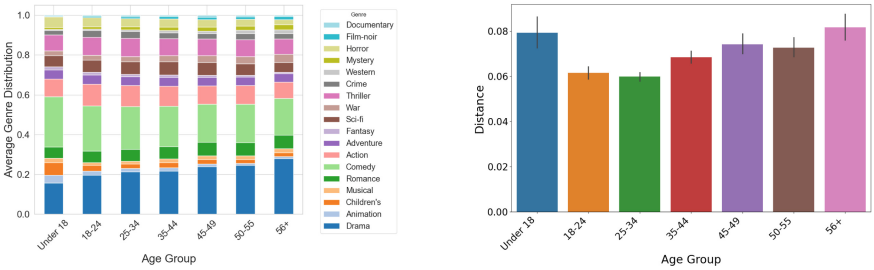
users' age at the time of the LEs illustrated in Fig. 1c, we see that relatively few LEs are removed. Particularly for younger user groups, who are at the center of our exploration, most of their LEs remain. For most users, more than 80% of items are retained, enabling us to capture interactions with many items—all with annotated genre information.

An analysis of the profiles obtained for each year highlights that users provide listening information over multiple years: On average, 6.04 yearly profiles per user are obtained. A large majority of users (42,816) have more than 1 profile (i.e. listening to music for more than 1 year), and more than half of the users (23,919) provide listening histories for more than 5 years. On average, each user has recorded 22,978.65 LEs. Each yearly profile includes on average 3,385.20 LEs and 266,063 profiles include more than 10 LEs. A majority of the LEs (73.11%) in the dataset comes from younger user groups, particularly those aged 17 to 29. We refer to this age group as `Mainstream Users`. Although 17-year-olds are technically children [51], their prevalence in the dataset precludes them from being considered a minority group. Consequently, for discussions involving LFM-2b, we refer to users aged 12 to 16 as `Children`, who account for 7.07% of all LEs. `NMA`s (30 to 65) are responsible for 14.71% of recorded LEs.

To capture individual users' consumption history based on the genres of items they interact with, we define **User Genre Profiles** for users grouped on an age level. As items can have multiple genres, we model each item as a uniform distribution of its genres, i.e. weights sum to 1. A UGP is a normalized distribution reflecting the mean frequency of each genre in the users' consumption history. If an item is consumed multiple times, all occurrences are counted to convey the higher frequency of repeated interactions. We create an **Age Genre Profile** (*AGP*) for each age group to represent the "average" genre consumption of users in that group. The consumption profile for a specific age group AGP_{age}, where *age* marks the group (for example, `Children, Mainstream`, or 17-year-olds), is the average of the UGP of each user in this age group. While more fine-grained approaches for creating user profiles exist [9], our technique offers a simple, interpretable, and comparable method for defining user profiles, following [46]. More complex profiling methods are left as future work.

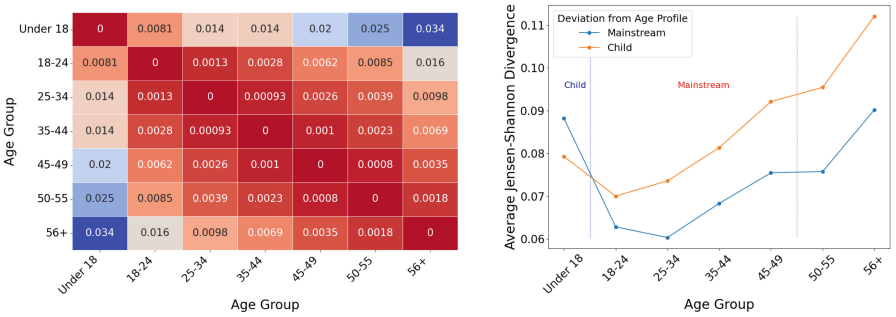
We analyze differences in users' consumption histories as a proxy for their preferences. To begin understanding age-related differences, the extracted profiles capture *snapshots* of users' consumption, deliberately excluding developmental factors and changes in individual preferences over time. To detect salient differences in genre consumption patterns of users within and between age groups, we leverage the Jensen-Shannon Divergence (JSD)„ which provides a bounded and interpretable measure of similarity between distributions [30]. Primarily focusing on how `Children` and `Mainstream Users` differ, we compare distributions from three different perspectives:

– **In-group Deviation**, the average spread of different preferences within an age group computed as the average JSD between an *AGP* and each *UGP* within the respective age group.



(a) Genre Distribution of *AGP*s. (b) In-group Deviation for age groups.

Fig. 2. Analysis of the genre distribution of different age groups on ML.



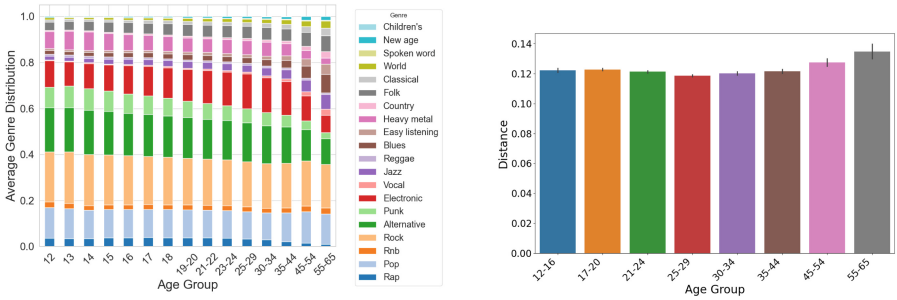
(a) Age Preference Deviations. (b) Mainstream/Child Deviations.

Fig. 3. Comparisons between genre distribution of different age groups on ML.

- **Age Preference Deviation**, the pairwise distance between *AGP*s using JSD.
- **Mainstream/Child Deviation**, the average JSD between *UGP*s of a certain age and the AGP_{Child} or $AGP_{\text{Mainstream}}$.

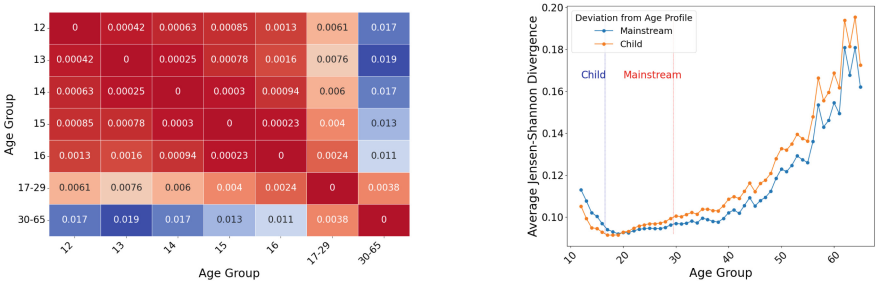
3.2 Results

Analyzing *AGP*s for **MovieLens**, illustrated in Fig. 2a, we note that 'Drama' is consumed more frequently the older the user, and genres such as 'Comedy' and 'Children's' appear more frequent in the *AGP*s of younger users. However, the Age Preference Deviation between AGP_{Child} and $AGP_{\text{Mainstream}}$ is 0.013 and the MANOVA [50] analyzing the effect of belonging to the `Children` or `Mainstream` user group on the proportion of different genres in the *UGP* is *not significant* ($p > .01$). This suggests that the genre proportions in the UGP are not meaningfully influenced by age group in this dataset. Notably, there are fewer `Children` UGPs ($N = 222$) relative to other age groups, which may influence the results of the test. Figure 2b highlights that `children` profiles differ markedly



(a) Genre Distribution of AGPs. (b) In-group Deviation for age groups.

Fig. 4. Analysis of the genre distribution of different age groups on LFM-2b.



(a) Age Preference Deviations. (b) Mainstream/Child Deviations.

Fig. 5. Comparisons between genre distribution of different age groups on LFM-2b.

from each other while `mainstream` profiles tend to be closer to the respective AGPs. As per Age Preference Deviation in Fig. 3a, UGPs of users younger than 17 typically deviate from those of all other age groups. The older the users are, the more their interests differ from those of younger users. Figure 3b highlights that `Children` differ the most from the $AGP_{\text{Mainstream}}$ and are closer to the AGP_{Child}. The older a user gets, the more their UGP differs from the AGP_{Child}.

Although preference trends on **LFM-2b** in Fig. 4a are barely discernible and the Age Preference Deviation between AGP_{Child} and $AGP_{\text{Mainstream}}$ is 0.0041, the MANOVA analyzing the effect of age group membership (`Children` or `Mainstream`) on the frequency of different genres in the UGP is significant ($p < .01$). Further analysis using Tukey's HSD [1] indicates that significant differences ($p < .01$) exist in the proportion of 'Rap', 'Alternative', 'Punk', 'Vocal', 'Jazz', 'Blues', 'Easy Listening', 'Country', 'Classical', 'World', and 'New Age'. Despite both datasets having a similar number of genres, In-group Deviation in LFM-2b is higher across all age groups than in ML and remains similar between all age groups (Fig. 4b); there is no significant difference ($p > .01$) between the In-group Deviation of `Children` or `Mainstream`.

Figure 5a shows that AGP_{12} to AGP_{17} deviate from $AGP_{\text{Mainstream}}$, with even greater divergence from AGP_{NMA}. However, as users grow up, their con-

sumption gradually aligns more closely with `Mainstream Users`. As seen in Fig. 5b, users 19 or younger have *UGP*s that align more with AGP_{Child} than with $AGP_{\text{Mainstream}}$. For users aged 20 and older, the *UGP*s align more with $AGP_{\text{Mainstream}}$. Overall, age groups that are close together appear to have similar listening patterns, while higher age difference indicates bigger deviations between the consumption of certain genres.

Results across the 2 datasets show that young adults exhibit more consistent consumption patterns than children and represent the majority of users driving interactions, regardless of the domain. `Children` emerge as a distinct and underrepresented group with divergent preferences. Despite their unique tastes, the relatively fewer `Children` profiles may cause their preferences to be overlooked among the `Mainstream`, highlighting the difficulty in effectively capturing `Children`'s preferences. Although we probe `Children`'s preferences during specific intervals, children are not a static group. They are in a developmental phase where individual tastes and interests are continuously changing [19,52], which may require more nuanced approaches to personalization. This emphasizes the complexity of addressing the needs of niche user groups in broader systems.

4 Investigating Recommender Impact Across Age Groups

In phase two of this exploration, we build on insights from Sect. 3 that reveal differing consumption behaviors for `Children` and `Mainstream`. As statistically significant differences in users' genre consumption history pertained only to LFM-2b and considering the more detailed information LFM-2b provides, we focus further analysis on this dataset. Specifically, we investigate whether RAs (i) capture unique consumption patterns, (ii) serve users from different age groups differently, or (iii) skew recommendations toward previous consumption of `Mainstream Users` at the expense of individual preferences.

4.1 Experiment Setup

We inspect two non-personalized RAs: **Random** and **MostPop**; alongside two personalized ones: **iALS** [22], a matrix factorization model which is trained based on implicit data which is frequently used as a non-neural baseline [3], and $\mathbf{RP}^3\beta$ [40], a graph-based recommendation model which is a well-performing alternative to other benchmark algorithms despite its simplicity [14]. To examine RA behavior, we focus on a specific timeframe—June 1st to October 30th, 2013—to gain insights into the broader effects of `Children` and `Mainstream Users` in an RS environment. This period was chosen because it encompasses a dense portion of the dataset, as the majority of LEs occurred in 2012 and 2013. By narrowing our scope to this timeframe, we ensure sufficient data for robust analysis while carefully verifying that the age distribution in this subset closely matches the overall distribution shown in Fig. 1b. As commonly done for experiments involving LFM-2b, we exclude user-song interactions where a user has listened to a song only once [28,34]. Additionally, we binarize ratings by including the

first listening event for each user-song interaction, disregarding multiple listens [3]. Users who interacted with fewer than 5 songs or songs with fewer than 10 interactions are excluded to reduce sparsity in the dataset.

To evaluate how RAs fare, we apply a temporal global split [7], one of the most realistic and strict splitting methods [23]. LEs from June to August 2013 are used for training, September 2013 for validation, and October 2013 for testing. Users lacking items in any of the splits are removed, resulting in a subset of 18,065 users and 159,900 items.

To model users' genre consumption and compare it to genres present in their recommendations, we define *UGP*s (as in §3.1), i.e. genre distributions of users' consumption history. In this case, however, items in the profiles are restricted to those in the train set, as these are the interactions available to the RAs. The *AGP*s are computed based on these *AGP*s, as described in Sect. 3.1. We define **Recommendation Genre Profiles (*RGP*)s** to model genre distributions within a recommendation list akin to *UGP*s, but using solely the items in that list.

For RA performance analysis on individual age groups, we use nDCG, MRR, and MAP [18]. To study similarities between genre profiles and compare *UGP*s and *AGP*s with the respective *RGP*s to determine how well the genres of the recommendations align with those of users' consumption, we rely on:

- **Genre Miscalibration (*GMC*)**, the JSD between a *UGP* and the respective *RGP*. This builds on the use of JSD as a calibration measure [27], highlighting the alignment between the genres of the recommended items with users' previous consumption patterns [29,49].
- **Recommendation-Mainstream Deviation (*RMD*)** and **Recommendation-Child Deviation (*RCD*)**, the average JSD between *RGP*s and the $AGP_{\text{Mainstream}}$ or AGP_{Child}, respectively.

We first assess the ability of RAs to cater to the preferences of users across different age groups, accounting for performance metrics and genre-profile similarities among `Children`, `Mainstream Users`, and `NMA`s. We then investigate whether recommendations for `Children` are influenced by the dominance of `Mainstream Users` interactions by training the RAs using data from the entire user base as well as data restricted to `Children`—a recommendation scenario in which the interactions of non-`Children` do not influence outcomes. For this, we use General-Set, which reflects the user distribution dominated by `Mainstream` profiles, and Child-Set, a filtered version of General-Set that includes only the 1,215 `Children` profiles (users aged 16 or younger), respectively.

For RA deployment, we use the Elliot framework [4] and follow the configurations suggested by Anelli et al. [3] for hyperparameter tuning and training (see repository). We generate **50 recommendations** per user using each RA, a common cutoff that allows us to build informative *RGP*s.

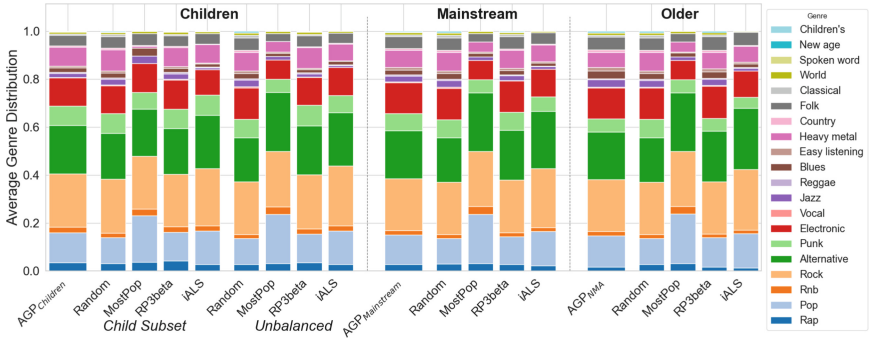


Fig. 6. *RGP*s and average RGPs across age groups and train sets.

4.2 Results

We present an overview of experimental results in Fig. 6 and Table 1. We begin our analysis by comparing the recommendations based on the General-Set for different age groups and use a one-way ANOVA [48] to measure the effect of the age group on each assessment measure. If the ANOVA is significant ($p < .01$), we compare pairs using Tukey's HSD with $p < .01$. Differences in RA performance between age groups are not significant for MostPop or Random. However, MostPop does outperform Random across all age groups, which is expected given its well-documented ability to suggest suitable items [8,11]. Recommendations generated by non-personalized RAs show a considerable deviation from users' previous consumption patterns (GMC). Yet, the low RCD and RMD values produced by MostPop and Random suggest that recommendations offered by these baselines are close to the "average" profiles captured by the data. These findings underscore that RCD and RMD are useful to account for the amount of personalization produced by an RA. The similarity in results across age groups for MostPop and Random traces back to their non-personalized nature. Interestingly, the GMC is significantly smaller for `Children` in comparison to `Mainstream` and `NMA`s when creating recommendations with MostPop, indicating that `Children`'s consumption is rather aligned with genres that are popular across a wide audience base.

Analysis of the personalized RAs reveals that `Children` are better served than `Mainstream Users` and `NMA`s, as evidenced by the MAP scores for RP$^3\beta$ and all performance metrics for iALS. In terms of genre distributions, personalized RAs show lower GMC scores compared to non-personalized ones, indicating better genre calibration, a key property for effective recommendations [49]. On the other hand, RCD and RMD scores are higher for RP$^3\beta$ and iALS compared to non-personalized RAs. Hence, instead of matching the "average" user profiles, personalized methods better cater to users' individual consumption patterns. While there are no significant differences for RP$^3\beta$ in terms of genre-profile similarities between age groups, iALS performs better at GMC for `Children` compared to `Mainstream` or `NMA`s. In line with this, the RCD is also smaller

Table 1. Results per age group: `Children` (c), `Mainstream` (m), `NMAs` (n). Significant differences between two groups are annotated with the corresponding pair. An asterisk (*) on Child-Set row denotes significant differences in the recommendations for `Children` between Child-Set and General-Set.

	Data	Age Group	nDCG$^{\uparrow}$	MRR$^{\uparrow}$	MAP$^{\uparrow}$	$GMC^{\downarrow}$	RCD	RMD
Random	Child-Set	Children	.0002	.0003	.0003	0.1009*	0.0164*	0.0169*
	General	Children	.0001	.0002	.0002	0.1033^{n}	0.0182	0.0161
		Mainstream	.0002	.0006	.0006	0.1083	0.0184	0.0161
		NMAs	.0001	.0003	.0003	0.1124^{c}	0.0183	0.0160
MostPop	Child-Set	Children	.0112*	.0297*	.0237*	0.1084*	0.0351*	0.0350*
	General	Children	.0061	.0162	.0127	$0.1001^{m,n}$	0.0225	0.0274
		Mainstream	.0055	.0125	.0108	$0.1085^{c,n}$	0.0226	0.0273
		NMAs	.0052	.0093	.0086	$0.1133^{c,m}$	0.0230	0.0276
RP$^3\beta$	Child-Set	Children	.0026*	.0070*	.0064*	0.0745*	0.1627*	0.1657*
	General	Children	.0146	$.0369^{n}$	$.0325^{m,n}$	0.0596	0.1131	0.1144
		Mainstream	.0129	.0269	$.0234^{c,n}$	0.0623	0.1659	0.1665
		NMAs	.0104	$.0210^{c}$	$.0185^{c,m}$	0.0609	0.1643	0.1624
iALS	Child-Set	Children	.0270	.0551	.0455	0.0646*	0.1030*	0.1072*
	General	Children	$.0269^{m,n}$	$.0544^{m,n}$	$.0430^{m,n}$	$0.0685^{m,n}$	$0.1139^{m,n}$	0.1172
		Mainstream	$.0185^{c,n}$	$.0395^{c,n}$	$.0327^{c,n}$	$0.0795^{c,n}$	$0.1196^{c,n}$	0.1218
		NMAs	$.0174^{c,m}$	$.0309^{c,m}$	$.0284^{c,m}$	$0.0842^{c,m}$	$0.1177^{c,m}$	0.1188

for `Children` than for `Mainstream Users`, resulting in suggestions for `Children` that align more closely with the AGP_{Child}.

Personalized RAs performing best for `Children` might come as a surprise; so is the lower deviation from their consumption patterns, with no apparent skew toward `Mainstream` interactions. Upon further scrutiny, we attribute this to the tendency of RAs to suggest popular items, which might benefit `Children`, i.e. children, particularly those aged 12 to 13, tend to have more popular items in their test sets. Further, the genres of popular items are better suited for `Children` than for `Mainstream Users` or `NMAs`, as indicated by GMC scores of MostPop.

We juxtapose the results for `Children` based on recommendations generated by RAs trained on General-Set and Child-Set; we perform paired t-tests ($p < .01$) on the scores for `Children` between the two training sets. When limiting the training data to `Children` profiles, we observe no significant difference in Random's performance, although both GMC and RCD are lower, while RMD is higher. This likely reflects the smaller item corpus, restricted to `Children`'s previously consumed items. MostPop performs better when trained on the Child-Set because it specifically captures what is popular among `Children`. This focus on `Child`-specific data allows the RA to better align with `Children`'s preferences, as only 46% of the top 50 most popular items in the General-Set overlap with those in the Child-Set. Interestingly, MostPop's GMC scores are worse, and its recommendations deviate more from both AGP_{Child} and $AGP_{\text{Mainstream}}$ when using the Child-Set compared to the General-Set.

Trained on the Child-Set, $RP^3\beta$ results in higher *GMC*, *RCD*, and *RMD* scores. Despite theoretically being better suited to match `Children` profiles due to the focused train set, $RP^3\beta$ struggles to accommodate their preferences, deviating more from both `Children`'s and `Mainstream Users`' past consumption. Instead, the RA better suits `Children` if interactions of `Mainstream Users` are present in the train data too, highlighting that the algorithm can accommodate diverse user preferences if a wide range of different user interactions is available.

Unlike $RP^3\beta$, iALS benefits from being trained exclusively on child interactions. Performance scores remain comparable to those achieved with the General-Set, remaining high for nDCG, MRR, and MAP. Notably, the *GMC*, *RCD*, and *RMD* for `Children` are lower than when training on the General-Set, indicating better calibration of genres. This suggests that focusing on `Children`'s consumption patterns enables more personalized recommendations that align more closely with the genres `Children` have previously consumed. At the same time, these recommendations reflect genre distributions that are more consistent with both `Children`'s and `Mainstream Users`' average patterns. This highlights that when trained on the General-Set, the RA reflects the unbalanced nature of users in the dataset, leading to recommendations that are less in line with `Children`'s unique preferences.

5 Discussion

Results detailed in Sect. 3.2 confirm nuanced differences in consumption patterns across age groups, with younger children deviating more from `Mainstream Users` and gradually aligning with them as they grow older. While prior works uncovered distinct consumption patterns of children [44,47], we build on these findings by examining two datasets, with LFM-2b offering a long-term view of consumptions. Further, we quantify how children's consumption patterns differ from those of the `Mainstream Users`, offering detailed, data-driven insights. Given that `Children` represent a minority of users, these differences raise concerns about whether RAs cater effectively to their preferences. Our findings underscore that the dominance of adult preferences in a combined dataset may overshadow those of children, motivating our second experiment to investigate how RAs trained on a broad user base impact `Children`'s experiences. As previous studies have shown inconsistent results in how well systems perform for children [13,44], we carefully examine a snapshot that reflects long-term consumption patterns, offering insight into how music RAs *actually* serve different user groups. We also measure algorithm performance and assess how well recommended genres align with users' past consumption, incorporating calibration as a criterion for recommendation quality [49]. The findings presented in Sect. 4.2 initially challenge expectations that children might be underserved due to the influence of majority groups. `Children` often emerge as being better served than `Mainstream Users`. We posit that this phenomenon relates to popularity bias: `Children` tend to prefer items that are not only popular among group members but also among the entire population, whereas `Mainstream Users` have

more diverse preferences and niche tastes. Further investigations of the specific items and levels of popularity favored by children are needed to understand this dynamic.

In analyzing how well RAs perform when trained on the Child-Set, we observe that not all RAs benefit equally from this focus on the specific user group. While iALS leverages `Children`'s preferences effectively and performs comparably when trained on either set, $RP^3\beta$ performs significantly worse when trained on the Child-Set according to all metrics. Algorithms are often designed and evaluated on datasets representing a broad user base, with unbalanced representations of users with varying and diverse characteristics. Not all algorithms can accommodate the preferences of a minority group by focusing on their interactions. Instead, considering a wide range of users and diverse interactions can be more suitable for improving recommendations for minority groups with certain RAs. Said and Bellogín [43] find that only some user groups benefit from RA training tailored to their profiles. Such "easy users" typically exhibit high coherence in the types of items they prefer (e.g., genres). Since children do not directly benefit from focused training, they seem to be more "difficult", requiring additional data to receive well-suited recommendations. The fact that RAs still achieve highly fitting recommendations for children when trained on the General-Set highlights distinct dynamics and properties of this group, warranting further investigation.

These findings have important implications for RS designed for a broad user base. Systems must consider whether all user groups are treated equally, especially when dealing with a diverse population. At the same time, systems tailored to minority groups, such as children, need to carefully select algorithms that can effectively capture focused preferences with fewer user interactions. Our results suggest that $RP^3\beta$ may not be suitable for such cases, while RAs like iALS could be more appropriate for capturing children's preferences. This insight also invites broader consideration of other minority groups and users with special consumption patterns. Although our analysis did not specifically focus on `NMA`s when preprocessing users, we observe that they are often served less effectively by different algorithms.

Limitations. MovieLens-1M, well-established in the research community, has the advantage of being widely used but is limited in size, e.g., only 222 child users, which hinders generalizability. Additionally, its broad demographic categories restrict insights into how preferences evolve with age or how users transition between age groups. LFM-2b offers in contrast more granular demographic data and allows for observations pertaining to preference changes over time. However, as of April 2024, LFM-2b is no longer publicly available, posing challenges for future studies. Additionally, the simplified approach of assigning genres based on artist annotations serves as a proxy and has its limitations. Future research should explore more nuanced methods for quantifying calibration while maintaining feature comparability. For instance, Lesota et al. [27] examines calibration by a song's country of origin. Similarly, using features such as tempo or mood as consistency criteria could provide alternative and more detailed perspectives.

Although established, LFM-1b UPG's genre annotations to label songs lead to removing several items in LFM-2b without annotation, which could affect user representation across age groups (§3.1). For younger children, however, the number of removed LEs remains relatively small, allowing us to preserve their preferences. Note that the removed songs are often less popular and may reflect distinctive preferences. This challenge extends to Sect. 4, where common preprocessing steps lead to removing songs with few interactions, impacting how distinctive remaining consumption histories might be for children's preferences. We addressed this by carefully assessing the effects of preprocessing on both genre and user distributions, minimizing the impact on children's unique preferences and ensuring robust representation. Despite endeavors for generalization, our exploration is limited to a specific timeframe and few RAs. Further validation is needed across different time periods and with a wider range of RAs.

6 Concluding Remarks

In this work, we analyzed how mainstream-driven RAs impact recommendations for children—a group often overlooked despite their uniqueness and large presence. Children rarely get the spotlight when systems are evaluated, so we address this gap by utilizing the only available datasets with demographic information that include them. Under careful consideration of how to preserve the distinct characteristics of their interactions, we explored and compared their consumption patterns, enabling us to make inferences about their preferences. Further, we assessed how well these preferences are captured by RAs and explored the impact of mainstream-driven systems on the alignment of recommendations to children's preferences. Our results highlight that such explorations are not a straightforward process. The observed user groups, the training data, and the RAs themselves all contribute to different outcomes. Our findings underscore the critical need for nuanced evaluations when developing and deploying child-aware RS, as well as the need to extend such considerations to other minority user groups.

Acknowledgments. We want to thank Markus Schedl and Stefan Brandl for their support in using the LFM-2b dataset and their guidance regarding the collection of demographic data. Further, we would like to thank Michael Ekstrand for his insights regarding findings reported in [13]. This work was supported by Grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF, a way of making Europe."

References

1. Abdi, H., Williams, L.J.: Tukey's honestly significant difference (hsd) test. Encycl. Res. Des. **3**(1), 1–5 (2010)
2. Alonso-Cortés, M., Cantador, I., Bellogín, A.: Recommendation fairness in eparticipation: listening to minority, vulnerable and nimby citizens. In: European Conference on Information Retrieval, pp. 420–436. Springer, Heidelberg (2024). https://doi.org/10.1007/978-3-031-56066-8_31
3. Anelli, V.W., Bellogín, A., Di Noia, T., Jannach, D., Pomo, C.: 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, pp. 121–131 (2022)
4. Anelli, V.W., et al.: 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, pp. 2405–2414 (2021)
5. Beel, J., Langer, S., Nürnberger, A., Genzmehr, M.: The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. In: Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J. (eds.) TPDL 2013. LNCS, vol. 8092, pp. 396–400. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40501-3_45
6. Bilal, D.: The mediated information needs of children on the autism spectrum disorder (asd). In: Proceedings of the 31st ACM SIGIR Workshop on Accessible Search Systems, Geneva, Switzerland, pp. 42–49. ACM, Geneva (2010)
7. Campos, P.G., Díez, F., Cantador, I.: Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. User Model. User-Adap. Inter. **24**, 67–119 (2014)
8. Cañamares, R., Castells, P.: Should i follow the crowd? a probabilistic analysis of the effectiveness of popularity in recommender systems. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 415–424 (2018)
9. Cantador, I., Bellogín, A., Vallet, D.: Content-based recommendation in social tagging systems. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 237–240 (2010)
10. Chaney, A.J., Stewart, B.M., Engelhardt, B.E.: How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 224–232 (2018)
11. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 39–46 (2010)
12. Ekstrand, M.: Challenges in evaluating recommendations for children. In: International Workshop on Children & Recommender Systems (2017). https://md.ekstrandom.net/pubs/kidrec-eval-challenges.pdf
13. Ekstrand, M.D., et al.: 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, pp. 172–186. PMLR (2018)

14. Ferrari Dacrema, M., Boglio, S., Cremonesi, P., Jannach, D.: A troubling analysis of reproducibility and progress in recommender systems research. ACM Trans. Inf. Syst. (TOIS) **39**(2), 1–49 (2021)
15. Gao, S., Ng, Y.K.: Analyzing the preferences and personal needs of teenage readers to make book recommendations. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 463–468 (2021)
16. Ghazanfar, M.A., Prügel-Bennett, A.: Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. Expert Syst. Appl. **41**(7), 3261–3275 (2014)
17. Gómez Gutiérrez, E., Charisi, V., Chaudron, S.: Evaluating recommender systems with and for children: towards a multi-perspective framework. In: CEUR Workshop Proceedings 2021, vol. 2955. CEUR Workshop Proceedings (2021)
18. Gunawardana, A., Shani, G., Yogev, S.: Evaluating recommender systems. In: Recommender Systems Handbook: Third Edition, pp. 547–601. Springer, Heidelberg (2022). https://doi.org/10.1007/978-1-0716-2197-4_15
19. Hargreaves, D.J., North, A.C., Tarrant, M.: How and why do musical preferences change in childhood and adolescence. The child as musician: a handbook of musical development, pp. 303–322 (2015)
20. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM Trans. Interact. Intell. Syst. (TIIS) **5**(4), 1–19 (2015)
21. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. (TOIS) **22**(1), 5–53 (2004)
22. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 263–272. IEEE (2008)
23. Ji, Y., Sun, A., Zhang, J., Li, C.: A critical study on data leakage in recommender system offline evaluation. ACM Trans. Inf. Syst. **41**(3), 1–27 (2023)
24. Klimashevskaia, A., Jannach, D., Elahi, M., Trattner, C.: A survey on popularity bias in recommender systems. In: User Modeling and User-Adapted Interaction, pp. 1–58 (2024)
25. Kowald, D., Muellner, P., Zangerle, E., Bauer, C., Schedl, M., Lex, E.: Support the underground: characteristics of beyond-mainstream music listeners. EPJ Data Sci. **10**(1), 1–26 (2021). https://doi.org/10.1140/epjds/s13688-021-00268-9
26. Landoni, M., Huibers, T., Murgia, E., Pera, M.S.: Good for children, good for all? In: European Conference on Information Retrieval, pp. 302–313. Springer, Heidelberg (2024). https://doi.org/10.1007/978-3-031-56066-8_24
27. Lesota, O., Geiger, J., Walder, M., Kowald, D., Schedl, M.: Oh, behave! country representation dynamics created by feedback loops in music recommender systems. In: Proceedings of the 18th ACM Conference on Recommender Systems, pp. 1022–1027 (2024)
28. Lesota, O., et al.: Analyzing item popularity bias of music recommender systems: are different genders equally affected? In: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 601–606 (2021)
29. Liang, Y., Willemsen, M.C.: The role of preference consistency, defaults and musical expertise in users' exploration behavior in a genre exploration recommender. In: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 230–240 (2021)
30. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991)

31. Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R.: Feedback loop and bias amplification in recommender systems. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 2145–2148 (2020)
32. Mansoury, M., Abdollahpouri, H., Smith, J., Dehpanah, A., Pechenizkiy, M., Mobasher, B.: Investigating potential factors associated with gender discrimination in collaborative recommender systems. In: The Thirty-Third International Flairs Conference (2020)
33. Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., Yilmaz, E.: Auditing search engines for differential satisfaction across demographics. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 626–633 (2017)
34. Melchiorre, A.B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., Schedl, M.: Investigating gender fairness of recommendation algorithms in the music domain. Inf. Process. Manag. **58**(5), 102666 (2021)
35. Milton, A., Allen, G., Pera, M.S.: To infinity and beyond! accessibility is the future for kids' search engines. In: IR for Children 2000-2020: Where Are We Now? – Workshop co-located with the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021). https://doi.org/10.48550/arXiv.2106.07813
36. Milton, A., Batista, L., Allen, G., Gao, S., Ng, Y.K.D., Pera, M.S.: "don't judge a book by its cover": exploring book traits children favor. In: Proceedings of the 14th ACM Conference on Recommender Systems, pp. 669–674 (2020)
37. Milton, A., Murgia, E., Landoni, M., Huibers, T., Pera, M.S.: Here, there, and everywhere: building a scaffolding for children's learning through recommendations. In: 1st Workshop on the Impact of Recommender Systems co-located with the 13th ACM Conference on Recommender Systems (ACM RecSys 2019) (2019). https://ceur-ws.org/Vol-2462/short2.pdf
38. Milton, A., Pera, M.S.: Evaluating information retrieval systems for kids. In: 4th International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (KidRec '20), co-located with the 19th ACM International Conference on Interaction Design and Children (IDC '20) (2020). https://doi.org/10.48550/arXiv.2005.12992
39. OfCom, U.: Children and parents: Media use and attitudes report 2023. Office of Communications London, London (2023)
40. Paudel, B., Christoffel, F., Newell, C., Bernstein, A.: Updatable, accurate, diverse, and scalable recommendations for interactive applications. ACM Trans. Interact. Intell. Syst. (TiiS) **7**(1), 1–34 (2016)
41. Pera, M.S., Ng, Y.K.: Automating readers' advisory to make book recommendations for k-12 readers. In: Proceedings of the 8th ACM Conference on Recommender Systems, pp. 9–16 (2014)
42. Pew Research Center: Teens and internet, device access fact sheet (2024). https://www.pewresearch.org/internet/fact-sheet/teens-and-internet-device-access-fact-sheet/. Accessed 9 Apr 2024
43. Said, A., Bellogín, A.: Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. User Model. User-Adap. Inter. **28**(2), 97–125 (2018). https://doi.org/10.1007/s11257-018-9202-0

44. Schedl, M., Bauer, C.: 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) (2017). https://doi.org/10.48550/arXiv.1912.11564
45. Schedl, M., Brandl, S., Lesota, O., Parada-Cabaleiro, E., Penz, D., Rekabsaz, N.: 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, pp. 337–341 (2022)
46. Schedl, M., Ferwerda, B.: Large-scale analysis of group-specific music genre taste from collaborative tags. In: 2017 IEEE International Symposium on Multimedia (ISM), pp. 479–482. IEEE (2017)
47. Spear, L., et al.: Baby shark to barracuda: analyzing children's music listening behavior. In: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 639–644 (2021)
48. St, L., Wold, S., et al.: Analysis of variance (anova). Chemom. Intell. Lab. Syst. **6**(4), 259–272 (1989)
49. Steck, H.: Calibrated recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 154–162 (2018)
50. Tabachnick, B.G., Fidell, L.S., Ullman, J.B.: Using Multivariate Statistics, vol. 6. Pearson, Boston (2013)
51. UNICEF: The convention on the rights of the child: The children's version (2019). https://www.unicef.org/child-rights-convention/convention-text-childrens-version#:~:text=A%20child%20is%20any%20person%20under%20the%20age%20of%2018
52. Valkenburg, P.M., Cantor, J.: The development of a child into a consumer. J. Appl. Dev. Psychol. **22**(1), 61–72 (2001)