# Challenges in Evaluating Recommendations for Children

Michael D. Ekstrand
People and Information Research Team (PIReT)
Dept. of Computer Science, Boise State University
Boise, Idaho 83725-2055, USA
michaelekstrand@boisestate.edu

## ABSTRACT

Recommender systems research and development cannot advance without robust evaluation strategies. While many evaluation strategies have proven effective for deploying and testing recommenders for general audiences, child-oriented recommendations pose unique challenges, and adult-oriented evaluation strategies do not necessarily translate.

In this position paper, I briefly describe several of the challenges I see in evaluating recommender systems for children, how they relate to similar problems for general audiences, and why existing solutions from the recommender systems community are insufficient. Significant progress in building compelling, useful, and personalized information experiences for children will require new developments in evaluating their effectiveness.

## KEYWORDS

recommender systems, evaluation

**ACM Reference format:**
Michael D. Ekstrand. 2017. Challenges in Evaluating Recommendations for Children. In *Proceedings of International Workshop on Children & Recommender Systems at RecSys '17, Como, Italy, August 2017 (KidRec),* 2 pages.
https://doi.org/

## 1 INTRODUCTION

Recommender systems crucially rely on evaluation strategies. Data and metrics are required to tune algorithms and train models and we need means of determining recommender effectiveness. However, many of the methods developed for general audiences are not not likely to work as well when applied to children, and there are additional challenges around privacy and multiple stakeholders that arise in child-oriented recommendations.

The challenges to evaluating recommender systems for children include:

- Lack of available data sets, for both practical and legal reasons
- Limited attention span or literary abilities for surveys
- Multiple stakeholders in a child's information experience

These problems do not only affect recommender systems; they also affect child-oriented search engines and other information access tools. In this paper, I consider these challenges and existing solutions from general-audience recommender evaluation.


*KidRec, August 2017, Como, Italy*


## 2 EVALUATING WITH ADULTS

Evaluating recommendations for adults is not an easy task, but is reasonably well-understood. Public data sets from several domains, including movies [5], books [12], and consumer products [6], enable offline evaluation of recommender effectiveness [4], and A/B tests are common when targeting "standard" Internet users [9], applied thoughtfully, can produce meaningful and robust results.

## 3 CHALLENGES WITH CHILDREN

Specifically targeting children — or in some cases, groups of users known to contain children — introduces several complexities into the recommender evaluation process.

### 3.1 Lack of Data

Child-facing recommendation lacks the standard benchmark data sets that have propelled research in recommendations and information retrieval for children. Data collected from children is rightly subject to strict privacy controls; U.S. law limits collection of data from children under 13. Therefore, there are not many public services that tailor to children, and those that do cannot release data sets. Some data is available in a limited fashion under confidentiality agreements, such as that used by Pera and Ng [10], but it is generally small and not publicly available, limiting both robustness and replicability.

While offline evaluation is limited in its ability to predict online user response, the availability of standard data sets has enabled a great deal of research on recommendation techniques and remains a crucial pre-filtering step to test and optimize algorithms before deploying them to users, and research studies involving users often using existing data sets to train the recommenders users will experience. The lack of such data makes research progress in recommendations for children difficult.

### 3.2 Limited Survey Abilities

Many evaluation techniques depend on some form of survey responses from users. There are at least two crucial challenges when performing survey research with children: they are unlikely to want to take a lengthy survey (it is difficult enough to get adults to take a survey sufficiently long to rigorously measure multiple factors), and they may not have sufficient reading or cognitive skills to understand and reason about the survey responses or their understanding of the items. Adults can answer surveys about resources for children, but as we will see, that provides an incomplete picture of relevance or interest.

Carefully-designed, simplified surveys can be used with children, but common survey methods for recommender systems [8] will need significant adaptation. This not only hinders survey-based

evaluation, but also methods that depend on survey data such as graded relevance [7].

### 3.3 Limited Assessment Abilities

General-audience recommender systems are typically trained and evaluated on data sets consisting of ratings, clicks, plays, purchases, and similar activities. While there are significant limitations to the assumptions we make about such data (does clicking an article really mean the user finds it interesting?), these the situation is worse when children are concerned.

Children are in the process of acquiring the basic skills needed to assess a resource and determine whether it will meet their needs or desires. Domain knowledge is known to affect user response to search result pages [1], but children often lack even general information literacy skills. It seems likely that click logs from children interactions with a system are likely to be even noisier with respect to resource relevance or interest than those from general users, and children are unlikely to be able to provide robust ratings particularly when attempting to accommodate non-taste factors such as educational value or information accuracy.

### 3.4 Multiple Stakeholders

Recommender system evaluation has historically focused on its impact on the user (accuracy, satisfaction, click through rate, etc.) or, through this impact, its effect on business outcomes (e.g. sales volume and user retention). However, in child-oriented recommendations, there are several stakeholders to consider on the user side of the equation alone. The child themselves has interests and information needs; they likely have a caretaker (e.g. parent or guardian) who has input into the kinds of material suitable for the child; and in many settings, such as elementary classrooms, a teacher who wants recommendations to support particular learning outcomes. Effective evaluation needs to clearly identify the stakeholders and assess their interests; the outcomes of the other RecSys 2017 workshop on *Value-Aware and Multi-Stakeholder Recommendation* may provide insights into achieving this.

## 4 THE PROMISE

These problems are significant, but there is great promise in addressing them. In addition to extending recommendation technology to meet the particular needs and desires of an often-overlooked user demographic, recommendation may be able to significantly enhance the educational and other life experiences of children.

Directly embracing the multi-stakeholder nature of child-targeted recommendation will likely enable many of these applications. A book recommender, for example, could take into account both the child's taste (the traditional operation of consumer-oriented recommender systems) and learning objectives set by their school or teachers. It could also account for values or constraints expressed by the child's caretakers, for example with regards to the amount of violence contained in their reading material or treatment of subjects that the caretaker knows will cause the child distress.

## 5 CONCLUSION AND THE PATH FORWARD

One does not simply evaluate recommender systems for children through off-the-shelf application of existing evaluation strategies. Data sets are not commonly available — nor should they be, given current legal and ethical standards for handling data from children — for offline evaluation. In online evaluation, we must pay particular care to the capabilities and interests of children, in addition to the consideration of their interests intrinsic to the recommendation process, and in many applications should engage with multiple stakeholders to ensure that recommendations advance the child's interests in multiple ways.

I expect that, in advancing the ability to deploy recommender systems to meet the particular information needs of children at various stages of life, we will need to make significant advances in design and evaluation, relying heavily on techniques that are quite new to the recommender systems research literature. *Participatory design* [11] seems particularly promising, as it engages multiple stakeholders across the design and evaluation process and is being employed with children and their caregivers, but recommender systems research and development rarely employs it. Participatory design has several potential benefits for general-audience systems [3]; it seems even more useful for navigating the complex landscape of building experiences that work well for children. Directly engaging children and their adults [2] will enable fascinating new applications.

## REFERENCES


[1] Michael J Cole, Xiangmin Zhang, Chang Liu, Nicholas J Belkin, and Jacek Gwizdka. 2011. Knowledge Effects on Document Selection in Search Results Pages. In *Proc. SIGIR 2011*. ACM, New York, NY, USA, 1219–1220. https://doi.org/10.1145/2009916.2010128

[2] A Druin. 2002. The role of children in the design of new technology. *Behav. Inf. Technol.* 21, 1 (1 Jan. 2002), 1–25. https://doi.org/10.1080/01449290110108659

[3] Michael D Ekstrand and Martijn C Willemsen. 2016. Behaviorism is Not Enough: Better Recommendations Through Listening to Users. In *Proc. RecSys 2016 (RecSys '16)*. ACM, New York, NY, USA, 221–224. https://doi.org/10.1145/2959100.2959179

[4] Asela Gunawardana and Guy Shani. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *J. Mach. Learn. Res.* 10 (2009), 2935–2962. http://jmlr.org/papers/v10/gunawardana09a.html

[5] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (Dec. 2015), 19:1–19:19. https://doi.org/10.1145/2827872

[6] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proc. WWW 2016*. International World Wide Web Conferences Steering Committee, 507–517. https://doi.org/10.1145/2872427.2883037

[7] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and Predicting Graded Search Satisfaction. In *Proc. WSDM 2015*. ACM, New York, NY, USA, 57–66. https://doi.org/10.1145/2684822.2685319

[8] Bart Knijnenburg, Martijn Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Model. User-adapt Interact.* 22, 4-5 (1 Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[9] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2008. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18, 1 (2008), 140–181. https://doi.org/10.1007/s10618-008-0114-1

[10] Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating Readers' Advisory to Make Book Recommendations for K-12 Readers. In *Proc. RecSys 2014*. ACM, New York, NY, USA, 9–16. https://doi.org/10.1145/2645710.2645721

[11] Douglas Schuler and Aki Namioka (Eds.). 1993. *Participatory Design: Principles and Practices*. CRC / Lawrence Erlbaum Associates.

[12] Cai-Nicolas Ziegler, Sean McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proc. WWW 2005*. ACM, Chiba, Japan, 22–32. https://doi.org/10.1145/1060745.1060754