# Personalizing an Online Parenting Library: Parenting-Style Surveys Outperform Behavioral Reading-Based Models

**Mark P. Graus**
Eindhoven University of Technology, IPO 0.20
5600 MB Eindhoven, the Netherlands
m.p.graus@tue.nl

**Martijn C. Willemsen**
Eindhoven University of Technology, IPO 0.17
5600 MB Eindhoven, the Netherlands
m.c.willemsen@tue.nl

**Chris C. P. Snijders**
Eindhoven University of Technology, IPO 1.[20]
5600 MB Eindhoven, the Netherlands
c.c.p.snijders@tue.nl

## ABSTRACT

The present study set out to personalize a digital library aimed at new parents by reordering articles to match users' inferred interests. The interests were inferred from reading behavior as well as parenting styles measured through surveys. As prior research has shown that parenting styles are related to how parents take care of their children, these styles are likely to be related to what content a parent is interested in. The present study compared personalization based on parenting styles against other types of personalization.

We conducted a user study with 106 participants, in which we compared the effects of four different approaches of personalization to our users' reading behavior and user experience: a non-personalized baseline, personalization based on reading behavior, personalization based on parenting styles measured through surveys, and a hybrid personalization based on both reading behavior and parenting styles. We found that while the reading behavior was not significantly influenced by different types of personalization, participants had a better user experience with our survey-based approach. They indicated they perceived a higher level of personalization and satisfaction with the system, even though in terms of objective metrics this approach performed worse.

### ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g., HCI): User Interfaces; H.3.3 Information Storage and Retrieval: Information Search and Retrieval

### Author Keywords

Personalization; Parenting; User Experience; Cold Start; Psychological Traits; Psychological Models; User Models

## INTRODUCTION

Becoming a parent is for many a big challenge in life. New parents have to get used to a new set of responsibilities and have to learn a whole new set of care-taking skills, ranging from practical (such as changing diapers) to more emotional (such as recognizing and reacting to a child's emotions). There are numerous ways to acquire these skills: parents can get advice from relatives, or alternatively rely on vast amounts of books, websites, videos, and other types of media.

Parents have different styles of parenting and as such some topics may be very relevant to a parent, while others are completely irrelevant. In this sense, helping parents find their way in content related to the parenting domain is similar to personalization areas such as movie or book recommendations. A challenge in personalizing content on parenting is that first-time parents have to find their own way in a domain that is completely new to them. Parents may not have a clear view yet on the range of alternative ways of taking care of a child that match their styles. It might not be easy for them to judge what content is relevant and they might read content that is not in line with their parenting styles or interests. As such, there might be a discrepancy between what content new parents read and what is actually relevant to them. As a result personalization based on reading behavior (as is common) might not provide the desired results.

An additional challenge is that parenting is an activity people are very committed to and about which they hold strong beliefs. As a result, new parents might find certain types of content extremely irrelevant, to the point of being offended. A mother who struggled and eventually gave up breastfeeding might be hurt by receiving unwanted breastfeeding advice. Being wrong in personalization in this domain thus has a bigger impact than in other domains.

We aim to help parents in finding relevant content by personalizing a digital library of information articles on parenting. Because the content is aimed at new parents, we think a discrepancy can exist between reading behavior and reading interests and parenting styles measured through surveys might provide more reliable information for predicting reading interests. To investigate this, we personalize a library using both behavior data and survey data.

### Research Question and Hypotheses

The current paper aims to investigate how a library comprising articles on parenting can be improved by personalizing the


*HUMANIZE '18*, March 11, 2018, Tokyo, Japan

order in which the articles are presented[1]. A screenshot of what the library interface looked like can be found in Figure. 1. In addition the paper investigates if and how parenting styles can contribute to this personalization. The main research question thus is "How does personalization based on parenting styles compare to personalization based on reading behavior in terms of user behavior and user experience?"

We try to answer this research question by investigating the effects of personalization based on survey responses measuring parenting styles (explained in Section 1.4) and more conventional ways of personalization that rely on behavior data. Specifically we compare the effects of survey-based personalization with personalization based on reading behavior, personalization based on both reading behavior and survey responses and a non-personalized baseline. We are interested in the effects of this personalization both in terms of influenced behavior (e.g. does personalization based on surveys increase the number of articles users read?) and in terms of user experience (e.g. does personalization based on surveys result in a higher satisfaction with the digital library?). To investigate the effects on the user experience we adopted the User-Centric Evaluation Framework for personalized systems by Knijnenburg and Willemsen [11]. We designed a UX survey with items aimed to measure different aspects of the user experience.

With the survey we aimed to measure three aspects of the user experience specifically and formulated survey items to do so: the perceived level of personalization ("The library shows articles I find interesting"), the system satisfaction ("It was easy to find relevant/interesting articles"), and reading satisfaction ("I enjoyed reading the items I read"). We hypothesize that the different ways of personalization influence the perceived level of personalization. A higher level of personalization should lead to a higher system satisfaction, which should lead to a higher reading satisfaction. The higher system satisfaction is also expected to increase the amount of reading by the user.

In terms of improving user satisfaction and increasing reading behavior we hypothesize the following order in the different personalizations, from worst to best:

- The non-personalized library
- The library personalized based on just reading behavior
- The library personalized based on just survey responses
- The library personalized based on both reading behavior and survey responses

The remainder of this section will introduce the theoretical background on which this study is based.

### Personalization

Personalization is the process of altering a system to fit to the needs and/or preferences of an individual [15]. Examples

[1]The content and the design of the library were taken from Philips' uGrow app. Available for iOS https://itunes.apple.com/app/ugrow-healthy-baby-development/id1063224663 and Android https://play.google.com/store/apps/details?id=com.philips.cl.uGrowDigitalParentingPlatform

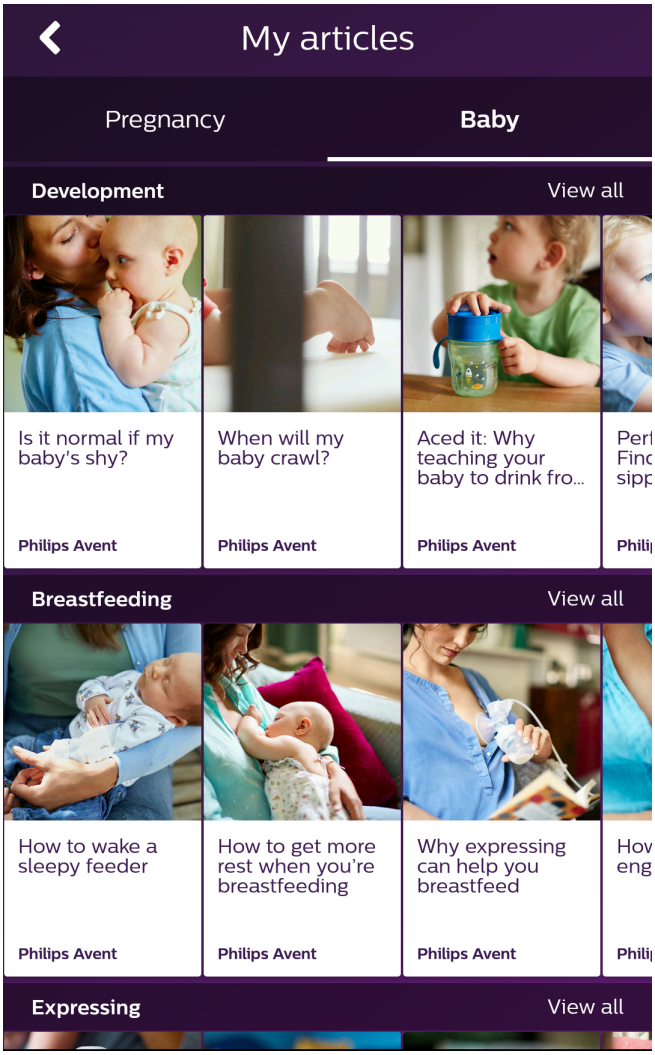


**Figure 1. uGrow 'My Articles' page**

of personalization can be found on numerous websites, for example in the form of recommendations on Amazon, or as filters on social media feeds such as Twitter and Facebook. In general the goal is to alter a system in a way that it caters to the individual needs of a user to influence user behavior or user experience. A typical goal of influencing behavior is to make users consume more content in a media browsing system or purchase more items in a webshop, while a typical goal of influencing user experience is to make it easier for users to reach their goals.

Personalization can be implemented in many different ways [19], but the most widely adopted methods rely on historical data describing how users interact with a system, and combine these data across users to make predictions on what content a user will find relevant. The system is subsequently altered so that the user is exposed to more of the content he/she is likely to find relevant.

A standard problem related to this approach to personalization is the cold start problem [18]. More specifically, three cold

start problems exist: the system cold start, the user cold start, and the item cold start. The system cold start occurs when not enough data are available within the system as a whole to make predictions. The user and item cold start occur when there are not enough interaction data available corresponding to respectively the user or the item, so that no predictions can be made for respectively the user or the item.

In the context of parenting an additional challenge occurs. Apart from being new to a system, (some) parents are also new to being parents and they might find it hard to identify what content is relevant to them. This can result in a mismatch between the content they read and the content that they are actually interested in. In systems in which user evaluations of content are not being tracked explicitly, assuming that content is appreciated because it was read may well lead to inaccurate predictions about user preferences. Because of this, a library aimed at parents might benefit from relying on other types of data for personalization.

### Personalization and Psychological Traits

Many psychological traits have been incorporated in personalization applications. Hauser et al. [9] personalized an online tool to compare contracts for mobile phones based on cognitive styles (i.e. the way in which individuals prefer to process information) and showed that providing users information in a way that matches their cognitive style (e.g. textual versus visual information) increases buying propensity. Germanakos and Belk [7] found that adapting an online learning environment to the working memory capacity of its students resulted in higher test scores.

Similarly, Fernandez-Tobias et al. [5] showed that incorporating personality in collaborative filtering algorithms allowed them to better predict recommendations across domains (e.g. recommending movies based on someone's music listening behavior). They did this by extending the SVD++ algorithm [12], an algorithm used to predict ratings that users will assign to items. Fernandez-Tobias et al. used a part of the myPersonality dataset[2] comprising 160k users and in total just over 5 million likes over 16k items (consisting of books, movies or music artists). The personality traits (the five factor model with the traits openness to experience, conscientiousness, extraversion, agreeableness and neuroticism [14]) were available for all users and were used to predict likes. Their results showed that incorporating the personality information substantially improved the extent to which likes on Facebook could successfully be predicted.

These studies demonstrate that personalization can benefit from considering and incorporating personal characteristics (such as personality traits or cognitive styles). In the case of parenting, parenting styles are psychological traits that are likely to play a role in what content parents find relevant. In the present study we measure parenting styles and subsequently use them for personalizing the online library.

[2]Available from http://mypersonality.org/wiki/doku.php?id=download_databases

### Parenting Styles

Zhao [20] performed a literature review on research on parenting with the goal of understanding how scholars operationalize and measure parenting styles. Zhao was in particular interested in how parenting styles relate to the actual care-taking behavior and as such, the review was primarily focused on research that comprised both questionnaires and a behavioral aspect. She found that parenting as a whole is a combination of cognitive factors, the physical task of taking care of a baby, and the interplay between the two (cf. [2]). Zhao in addition found that researchers conceptualize parenting styles as individual differences along two cognitive dimensions: structure (i.e. how important parents think structure is for their children) and attunement (i.e. how much parents value reacting to a child's needs and how able they are at reading those needs) [1, 3, 16]. Prototypical parenting styles are the resulting combinations of scores along these two dimensions (high attunement/high structure, high attunement/low structure, low attunement/high structure and low attunement/low structure). Other cognitive factors that have been identified in literature to play a role are parental distress, perceived self-efficacy, and the perceived difficulty of the child.

The cognitive factors allegedly have an interplay with how parents actually take care of their children. To validate these parenting styles and investigate how they relate to care-taking behavior, Zhao [20] conducted a survey study in which she measured parenting styles and asked respondents to self-report on how they take care of their children. The analysis of the survey data showed support for the conceptualization of parenting styles along the previously mentioned dimensions of structure and attunement. In addition it showed that parenting styles are related to the actual care-taking behavior of parents. For example, parents scoring low on attunement are less likely to engage in breast-feeding and more likely to opt for bottle-feeding. As parenting styles are related to how parents take care of their children, they are likely to be useful predictors for what type of content parents are interested in. For example, parents that find structure important put their children to bed at a fixed bedtime instead of waiting for the kid to become sleepy. As a result they might be more conscious of the fact that their child does not fall asleep easily and will thus be more interested in content on how to get a baby to sleep well than people that value flexibility over structure and wait for their child to get sleepy.

## STUDY DESIGN

To investigate our research question we designed a user study that consisted of two main parts, with a first part aimed at collecting initial data to be used for personalizing the "My Articles" page and a second part aimed at investigating the effects of the different ways of personalization on the reading behavior and user experience.

During the first part, participants were asked to complete a survey to measure their parenting styles, after which they were invited to browse the non-personalized library (i.e. a library with a fixed order of articles). The responses to the surveys were stored for personalization later. The information regarding what articles participants read during the browsing

phase was used for personalization based on reading behavior. The order of articles for the second part of the study was calculated in one out of four ways (described in more detail in section 3). For each participant we selected at random which set of predictions was used to personalize the library.

In the second part of the study the participants were re-invited to interact with their now personalized digital library. Subsequently, participants evaluated their experience with the system through our UX survey. We first report on the initial phases of the study.

### Initial Data Collection: Survey and Reading Behavior

We implemented the online library on a website that was accessible through browsers on computers and mobile phones. We recruited participants through posts in online forums dedicated to parenting and through Facebook ads targeting parents in the United Kingdom and United States with children younger than two years old. In total 234 parents clicked on the link to participate in the study. All participants that completed the entire study were compensated with $4.50 or £3.50 of shopping credit for amazon.co.uk or amazon.com. The ad campaign and data collection took place in May and June 2017.

The first part of the study consisted of two steps. In the first part people were asked to complete the survey to measure parenting styles. After completing the survey, the participants were presented with the digital library and invited to browse through it and read the articles that they were interested in. The participants were invited to read as many articles as they wanted for as long as they wanted and to click a link labeled "I've finished reading" once they felt like they read enough. After clicking this link participants were asked to submit their email address for the second step of the study.

In total 181 participants completed the survey (15 men/166 women, 99 first time parents, with an average age of the baby 11.39 (SD: 7.96) months). On average the whole session lasted just over 6 minutes (378 seconds, SD: 279.80 seconds). The survey consisted of 15 items of the original survey of Zhao [20]. For the five cognitive (structure, attunement, maternal self-efficacy, parental disstress and perceived difficulty of the child) we selected per factor the two items with most extreme factor loadings. We added items concerning the demographics of the parent (gender, level of education, whether they were first time parents) and child (gender, age) that had had large effects on the self-reported behavior in the original analysis. The factor scores for our participants were calculated by using the factor loadings from the original survey and are displayed in Figure 2. These scores show similar distributions and correlations as the factors in the original survey.

The interface of our library was made to have the look and feel of the original library (see Fig. 1) as much as possible. As in the original interface, the articles are subdivided in categories that are displayed in rows. Within the category rows the articles are displayed horizontally. The user is able to scroll up and down to different categories and left and right within categories to the different articles. As in the original interface, the order of articles and categories was fixed: every participant had exactly the same order of categories and articles.

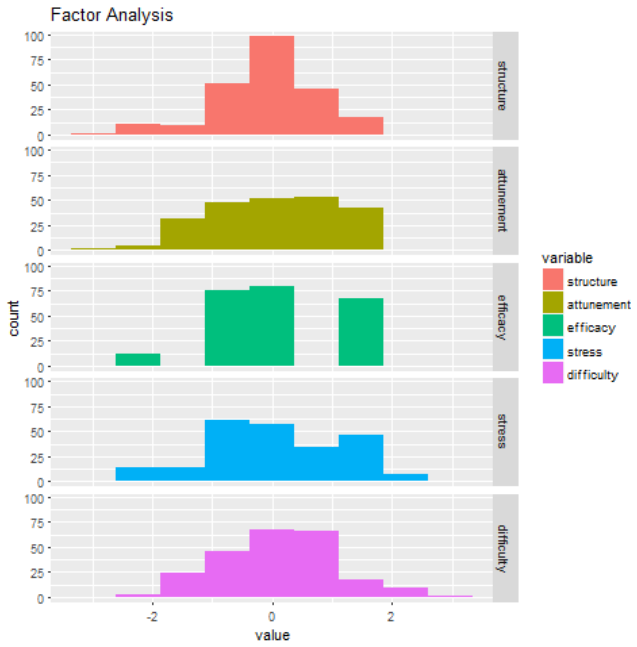


**Figure 2. Distributions of the 5 factor scores measured through the first survey.**

The initial part of the data collection was concluded with offering the participants to freely browse the online library. Participants opened on average 2.23 articles (SD: 3.37 articles) from 1.25 categories (SD: 1.51 categories). These data and the survey responses were used to calculate relevance predictions for the individual participants.

## CALCULATING RELEVANCE PREDICTIONS

Based on the data collected in the first step of the study we calculated per participant four different relevance rank predictions for all articles. As a baseline we used the non-personalized General Top-N. The three other ways of predicting differed in what data from the first step were used. A survey-based ordering was based on the data from the survey responses of the participants and on reading behavior at the aggregated level. A reading-based ordering used only data regarding the articles that people had read in the first step. Finally, a hybrid ordering used both the survey responses and the individual reading behavior. The way these orderings were calculated is described in the following sections.

### Survey-Based Predictions

We used the survey responses collected in the first step to predict relevance of the different articles for the participants in our study. To do this, the participants were subdivided in segments, by performing median splits on the 2 cognitive factors: attunement and structure. The user segment was then defined to be the combination of these two scores, resulting in four segments. We considered incorporating the three other factors measured in the study (self-efficacy, parental distress and perceived difficulty of the child), but given the number of users in our dataset adding additional factors resulted in segments that became too small.

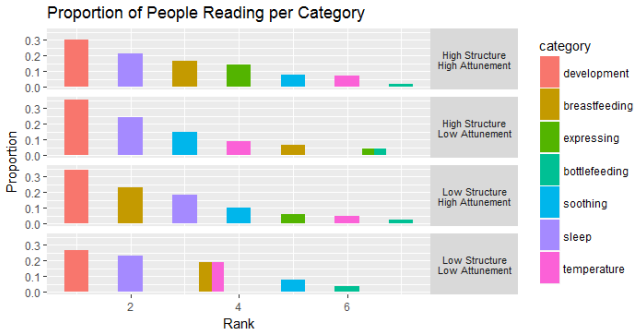


**Figure 3. Article Categories Ranked on Popularity per Segment**

As the participants read on average just over 2 articles, there was not enough data to show differences on the level of individual articles (i.e. articles were not read often enough to allow for enough variance), but participants from different segments did prefer different categories, as can be seen in Figure 3. When investigating these predictions, the popularity order for these categories seems to make sense intuitively. For example, the breastfeeding category is predicted to be more popular for segments with high attunement, which is congruent with the relationship with breast-feeding and high attunement in the original survey [20].

As a result we decided to sort the categories based on the attunement-structure segment and sort the articles within each category based on general popularity. That is, the survey-based predictions only personalized the order of the categories, not the articles within each category. We tried basing segments on other factors than attunement and structure, but the resulting predictions were not as easily interpretable as the predictions based on these segments.

### Reading-Based Predictions

For the conditions based on reading behavior alone, we used the Bayesian Personalized Ranking Matrix Factorization (or BPRMF) algorithm implemented in MyMediaLite [6, 17] to predict relevance. BPRMF is an extension to classic matrix factorization [13] that allows it to calculate recommendations from positive only feedback instead of rating data.

Conventional matrix factorization attempts to complete the matrix $R$ with dimensionality of $U$ (number of users) and $I$ (number of items). In this matrix the cells represent ratings the user has given to the corresponding item. This matrix is decomposed into two $k$-dimensional sub-matrices $P$ and $Q$ in which the rows of $P$ and $Q$ represent respectively users and items in a k-dimensional latent feature space. These matrices are constructed so that the predicted rating $\hat{r}_{ui}$ is calculated by taking the inner product $p_u * q_i$ (see Equation 1).

$$\hat{r}_{ui} = \mathbf{q}_i * \mathbf{p}_u \tag{1}$$

Rendle et al. [17] extended this matrix factorization into BPRMF to allow using positive only feedback to calculate per user a ranking of the articles from highest predicted relevance to lowest predicted relevance. In the current study, the positive only data describe whether or not a user read an article in the first step, and the predictions would indicate what items a user is most likely to read. In order to translate this positive-only, binary feedback into a ranking, pairs of items are semi-randomly selected per user, where each pair consists of an item that the user has interacted with and one with which the user has not interacted. The assumption is that the first is preferred over the second. Sampling a large number of pairs per user, results in a ranking that can be used in matrix factorization and the resulting model then calculates a relative relevance score instead of a rating.

### Hybrid Predictions

The BPRMF algorithm was extended to combine reading behavior and the individual parents' user attributes inferred from the survey for the calculation of hybrid predictions. The BPRMF algorithm was extended similarly to how Fernandez-Tobias et al. [5] extended the SVD++ [12] algorithm to incorporate personality in predictions.

Where the original BPRMF algorithm uses two matrices $P$ and $Q$ to calculate predictions, our user attribute aware BPRMF algorithm uses a third matrix $Y$. $Y$ describes the user attributes on the same $k$ latent features the users and articles are expressed in. In our case we used high and low scores for the five cognitive factors from our parenting style survey as user attributes. We decided again to use the median splits per factors to assign each user a high or low score for each factor in order to prevent overfitting. Every user has thus 5 user attributes and the relevance predictions are similar to the original BPRMF algorithm with an additional matrix in which user attributes are represented. The predicted relevance is then calculated according to equation 2.

$$\hat{r}_{ui} = \mathbf{q}_i * \left( \mathbf{p}_u + \sum_{a \in A(u)} \mathbf{y}_a \right) \tag{2}$$

This model is fit using stochastic gradient descent. Each iteration consists of two steps. In the first step the $P$ and $Q$ matrices are fit, while leaving the $Y$ matrix constant. In the second step the $Y$ matrix is fit, while leaving the $P$ and $Q$ matrix constant. We implemented this algorithm in the MyMediaLite library [6].

### Calculated Relevance Predictions

In total the dataset contained 221 users and 508 reads[3]. For each user predictions using the four methods described above were calculated. The predictions were then sorted in two steps. First the 7 categories were ordered based on the article with the highest predicted relevance (a strategy called min-rank that has been shown to work well in similar circumstances [4, 21]). Within the categories the articles were ordered based on predicted relevance.

The algorithms for the reading-based and hybrid predictions required the tuning of a set of regularization hyperparameters, which we carried out using Bayesian Optimization. The

[3]We included reading data from a pilot study to ensure we had enough data to calculate predictions

| | 5-fold Cross Validation | | | | Post-hoc Comparison | | | |
|---|---|---|---|---|---|---|---|---|
| algorithm | AUC | prec@5 | prec@10 | NDCG | AUC | prec@5 | prec@10 | NDCG |
| baseline | 0.840 | 0.083 | 0.065 | 0.424 | 0.706 | 0.146 | 0.104 | 0.477 |
| survey | - | - | - | - | 0.650 | 0.060 | 0.062 | 0.353 |
| reading | 0.832 | 0.079 | 0.061 | 0.411 | 0.767 | 0.176 | 0.114 | 0.522 |
| hybrid | 0.769 | 0.080 | 0.059 | 0.404 | 0.807 | 0.214 | 0.126 | 0.561 |

**Table 1. Performance Metrics calculated through 5-fold Cross Validation and a post-hoc performance analysis**

Bayesian Optimization was performed using 5-fold cross validation, using AUC as the target measure. Once optimal values for the hyperparameters were established, the predictive models were constructed and the predictive performance (i.e. the reading-based and hybrid recommendations) was investigated through 5-fold cross validation also. Table 1 shows these performance metrics of the three algorithms under the column '5-fold Cross Validation'. The performance metrics appeared to be adequate[4]. However, the baseline, reading-based, and hybrid predictions are calculated on the level of the individual articles, they cannot be easily compared to the survey-based predictions that are calculated first on the category level and then within the categories on an individual article level. In order to make a fair comparison, we performed a post-hoc analysis by recalculating the performance metrics for the sets of recommendations to correspond to the survey-based predictions. We did this by calculating the lists of recommendations and sorting all lists first by category based on the minimum predicted rank of the article within that category and subsequently sorting the articles within their categories based on the predicted relevance for the individual articles. We then calculated performance metrics by using the actual reading behavior as ground truth. The outcome of these recalculations can be found under the columns 'Post-hoc Comparison' in Table 1. These numbers indicate the most accurate predictions for the hybrid predictions, followed by the reading-based predictions, the survey-based predictions and finally the non-personalized baseline. Based on these metrics we would expect the hybrid predictions to be most in line with what participants will read, and the survey-based least. This order is different from the order in the k-fold cross validation metrics because no k-fold cross validation was applied to be able to compare with the survey-based recommendations (i.e. the train and test set were identical).

## RE-ENGAGING WITH THE NOW PERSONALIZED SYSTEM

The second part of the study was used to investigate our research question and test our hypotheses. To this end participants were re-invited to interact with the website, where they were now shown the library personalized in one out of four ways (selected at random). The invitations were sent out after all predictions were calculated, which means that the time between finishing the first part and starting the second part differed between participants (median 42.6 days. SD: 15.4 days). In this step the interface was personalized by reordering both the categories and the articles within the categories. The categories were ranked based on the minimum predicted relevance rank (or highest predicted relevance) within the category, so that the category with the article with the highest predicted relevance was shown on top. This way of sorting categories has been shown to be one of the best strategies in terms of reducing browsing time [4]. Within categories the articles were ordered by predicted relevance rank, with the article with the lowest predicted relevance rank to the left of the list.

Participants were allowed to browse the library freely during which we measured what articles the participants opened. Participants were shown a link labeled "I have finished reading" that would take them to the survey as soon as they felt they read enough. The survey contained 11 items aimed at measuring Perceived Level of Personalization, System Satisfaction, and Reading Satisfaction.

### Participants

All 181 users from the first part were invited to join the second part of the study via email. Of the 181 users we sent invitations to, 150 visited the second part and 121 completed the study. A number of cases were removed, for either trying to complete the study with multiple email addresses (3 users), having missing data in the survey (1 user), or finishing the second part of the study in less than 50 seconds (11 users). For our final data analysis we ended up with 106 users (9 men/97 women, 50 first time parents, mean (SD) age of the baby 10.63 (8.45) months)

These users were distributed roughly equally over conditions (baseline: 29, survey: 29, reading-based: 22, hybrid: 26). In addition, there appeared to be no bias in response rate for the different parenting style segments of the participants, with response rates of .56 for the low structure/high attunement segment, .65 for the high structure/low attunement segment, .73 for the high structure/high attunement segment and .60 for the low structure/low attunement segment($\chi^2(3) = 3.239$, $p = 0.356$).

### Results

To gain insight in how the different methods of predicting relevance influenced the final recommendations participants received, we calculated the difference of the recommendations with the general Top-N in terms of Spearman Rank Correlation. The (Spearman) correlation coefficient $\rho$ indicates to what extent lists are similar, with a value of 1 if the order is identical and -1 if they are in reverse order. The results are shown in Figure 4 and they reveal that the available reading data does not allow personalization that differs a lot from the baseline condition (as the correlation between reading-based and baseline is 0.91 on average). Personalization based on the

[4]An overview of the different metrics and how to interpret them can be found in [8].

survey-based predictions is quite different from the baseline predictions, with an average correlation of 0.37. The hybrid predictions fall somewhere in between the reading-based and survey-based predictions with a correlation of 0.74. These numbers indicate that the additional data of parenting styles allows for personalization that deviates more from the baseline than personalization based on reading behavior alone.

One possible explanation of the reading-based personalization not differing much from the baseline is insufficient data. As there are a limited number of users (221 users, see Section 2.1), that read a limited number of articles (2.44 articles on average) from a library with a limited number of articles (102) that was presented in a fixed order. As such the dataset might not contain enough variance between users' reading behavior to fully benefit from collaborative filtering. What argues against this is the fact that the reading-based and hybrid recommendations appear to outperform the survey-based predictions in terms of prediction accuracy (see Table 1).

### Reading Behavior

Participants read on average 2.72 articles (SD: 4.28 articles), but 42 participants (39.6%) did not read any articles. The descriptives for the number of article reads per condition are shown in Table 2. The different conditions had no significant influence on the number of articles people read, as negative binomial regressions with the condition as independent variable and the number of reads as dependent variable showed no significant difference across conditions. This implies that no support is found for the hypotheses regarding the effect of our experimental manipulations on how participants interact with their personalized libraries.

| | condition | Mean | SD | min | max | N |
|---|---|---|---|---|---|---|
| 1 | baseline | 2.448 | 3.501 | 0 | 13 | 29 |
| 2 | survey | 2.517 | 4.032 | 0 | 16 | 29 |
| 3 | reading | 4.273 | 6.670 | 0 | 31 | 22 |
| 4 | hybrid | 2.038 | 2.289 | 0 | 9 | 26 |

**Table 2. Article Reads Per Segment**

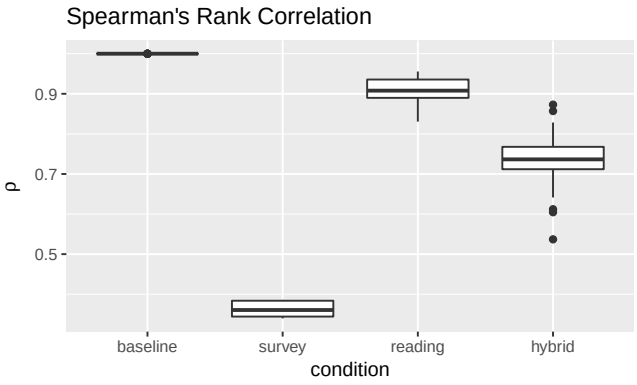


**Figure 4. Boxplots of Spearman's Rank Correlation with General Top-N per Condition**

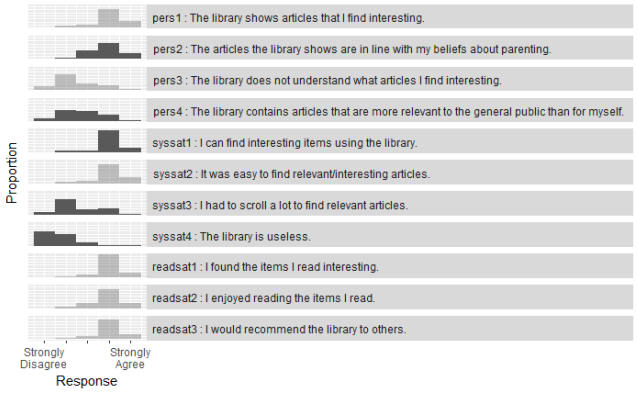


**Figure 5. Survey Items and Response Distributions. The light-grey items have been omitted from the analysis because of poor factor loadings.**

### User Experience

As per the user-centric evaluation framework by Knijnenburg and Willemsen [11] all survey items were submitted to a structural equation model (SEM). The responses to the individual items can be seen in Figure 5, with items belonging to Perceived Level of Personalization (pers1-pers4), System Satisfaction (syssat1-syssat4), and Reading Satisfaction (readsat1-readsat3). The three items for reading satisfaction show very low variance among each other, which lead to these three items not fitting in the model. This might have been caused by the fact that the reading behavior did not differ across conditions as we manipulated only the order in which the articles were presented, and not the actual content in the library. Therefore, people were actually able to read the same articles regardless of experimental condition and thus the reading satisfaction might be similar. Apart from the items on Reading Satisfaction, two of the remaining items (pers1 and syssat2) explained little variance and were also removed from the analysis.

Despite the fact that participants did not read a large amount of articles, the interface did allow participants to get a general idea of the library by looking at the categories and the article titles. Nevertheless, we do feel that the participants who actually read articles are better able to evaluate the library. To account for this we introduced an additional (dummy) variable labeled 'Read' indicating whether or not people read any articles.

A SEM was constructed using the remaining six survey items measuring two latent constructs (Perceived Personalization and System Satisfaction), the experimental conditions, and the variable describing whether or not people read as exogenous variables. The two latent factors had high correlation, but the model showed good fit (with $\chi^2(36) = 44.447$, $p = .158$, CFI = .984, TLI = .974, RMSEA = .047, 90% CI: $[0.000, .088]$). For each participant we used this model to calculate the scores on these latent factors to be used for the remainder of the analysis.

As the final model consists of only two latent constructs (Perceived Personalization and Systems Satisfaction) that are highly correlated, there is no clear underlying structural model to test anymore. For the analysis we could either combine both

factors into one overall latent factor, or analyze both factors separately. We chose to do the latter as both factors might still capture different nuances of the user experience, despite their high correlation.

We analyzed the effect of our manipulation on the factor scores of both constructs through linear regressions, with the factor scores as dependent variables and the experimental condition as independent variable. As additional moderator we included the dummy variable representing whether or not people read articles.

The average factor scores per condition for the two measured constructs can be found in Figure 6. The image shows an increase in both Perceived Personalization and System Satisfaction for the survey-based condition. The effects are higher for the participants that did not read (represented in the red bars) and lower for the participants that did (represented in the green bars).

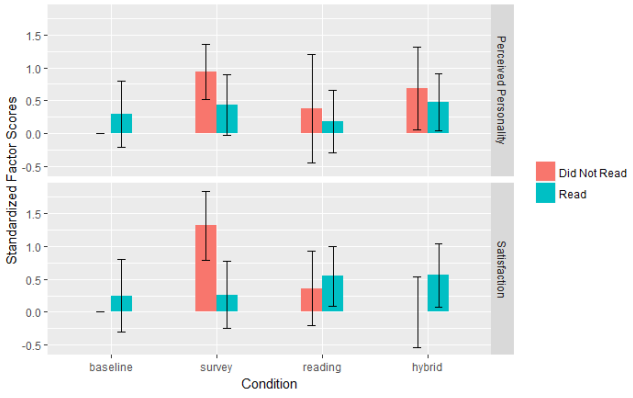


**Figure 6. Marginal effects on Perceived Personalization (top row) and System Satisfaction (bottom row) for the different conditions. The error bars correspond to +/- 1 standard error. Separate bars are shown for participants that read no articles (red) and at least one (green). Scores are standardized: a score of +1 implies 1SD higher than the baseline (baseline recommendations for a user that did not read). Error bars are one standard error of the mean.**

The regression models in Table 3 show these effects as well. Regression model (1) shows a positive and significant effect on Perceived Personalization for participants in the survey-based condition, indicating that these participants had the feeling the library catered more to their interests[5]. An additional, albeit not statistically significant, effect the table shows is an effect with a significance level of $p < 0.1$ for the increased perceived level of personalization in the condition with hybrid personalization. Although caution is needed when interpreting this effect, it describes a trend towards participants experiencing a higher level of personalization with the hybrid personalization.

In terms of System Satisfaction the patterns are slightly different. Participants that received the survey-based personalization were more satisfied with the system, as can be seen in model (2) in Table 3. Model (3) reveals how this effect holds up for participants that read versus participants that did not. It shows a negative interaction effect for the participants that received survey-based personalization and read at least one article, which suggests that only the people that do not read any articles actually perceive a higher system satisfaction; for those who do read at least one article the effect is strongly reduced.

In conclusion support is found for the hypothesis that survey-based personalization outperforms the non-personalized baseline, while no evidence was found that the reading-based and hybrid personalization did so. The lack of effect in terms of reader experience are in line with the comparison of the different predicted rankings in terms of Spearman's Rank Correlation, that showed a high similarity between the reading-based and non-personalized baseline. This comparison further showed that the survey-based personalization was most different from the baseline, which is also reflected in the user experience (albeit stronger for the people that did not read than the people that read). The hybrid conditions falls in between the survey-based and reading-based and similarly the effects on user experience appear to fall in between the effects of the survey-based and reading-based recommendations.

## CONCLUSION AND DISCUSSION

This study set out to compare personalization based on psychological traits measured through a survey to personalization based on reading data. Through a user study we compared different methods against a non-personalized baseline and showed that personalization based on survey information about parenting styles resulted in a significantly higher experienced user satisfaction and perceived level of personalization despite a lower objective performance, whereas using only historical reading behavior or the combination of historical reading behavior and measured parenting styles did not. Our findings speak to the potential usefulness of including data regarding characteristics of users (collected through an initial survey or otherwise) in personalization to alleviate the cold start problem. While the actual reading behavior for users was not influenced, an improved user experience may increase the probability for users to return to the library later on.

The fact that using the survey data for personalization also outperformed the condition where recommendations were based on both survey data and reading behavior is likely caused by the fact that the hybrid recommender - given how we had implemented it - came up with suggestions that were relatively close to the baseline condition. Hybrid predictions that would have assigned more weight to the survey data might have faired better. In any case, we do see that personalization based on surveys captures the interests better, or at least increase the reported user satisfaction, and that they lead to a more different order in which articles are presented than based on the reading behavior alone.

From a system owner point of view it is worth noticing that the survey-based predictions were very straightforward to calculate and implement compared to the reading-based and hybrid predictions. In addition, after completing the short survey the user can immediately benefit from personalization. Both the

[5]Because the factor scores are calculated through a Structural Equation Model they are normally distributed with a mean of 0 and SD of 1. Participants in the condition with survey-based personalization thus had a perceived level of personalization of 0.563 SD higher than participants in the baseline.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Perceived Personalization | | System Satisfaction | | | |
| | (1) | | (2) | | (3) | |
| | $\beta$ | (SE) | $\beta$ | (SE) | $\beta$ | (SE) |
| survey | $0.563^{*}$ | (0.241) | $0.673^{**}$ | (0.243) | $1.334^{***}$ | (0.334) |
| reading | 0.140 | (0.260) | 0.406 | (0.261) | 0.479 | (0.460) |
| hybrid | $0.438^{\bullet}$ | (0.248) | 0.273 | (0.249) | $-0.030$ | (0.391) |
| Read | | | | | 0.196 | (0.328) |
| survey:Read | | | | | $-1.279^{**}$ | (0.464) |
| reading:Read | | | | | $-0.158$ | (0.556) |
| hybrid:Read | | | | | 0.389 | (0.498) |
| Constant | 0.057 | (0.171) | 0.097 | (0.171) | $-0.004$ | (0.236) |
| Observations | 106 | | 106 | | 106 | |
| $R^2$ | 0.062 | | 0.072 | | 0.186 | |
| Adjusted $R^2$ | 0.034 | | 0.045 | | 0.128 | |
| Residual Std. Error | 0.919 (df = 102) | | 0.923 (df = 102) | | 0.882 (df = 98) | |
| F Statistic | 2.236 (df = 3; 102) | | 2.649 (df = 3; 102) | | $3.200^{**}$ (df = 7; 98) | |
| *Note:* | | | | | $^{\bullet}$p<0.1; $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 | |

**Table 3. Regression Tables for Experimental Manipulation and Read on Perceived Level of Personalization and System Satisfaction. The regression coefficients are the standardized $\beta$s and values between parentheses are standard errors.**

reading-based and (to a lesser extent) the hybrid predictions require reading behavior from the user before they can be calculated. Admittedly providing explicit feedback in the form of a survey demands more effort than the implicit feedback provided through the natural interaction of reading. However, the higher user experience suggests there might be a trade-off between the costs of user effort and the benefits of accurate personalization.

Another interesting finding is that the effects of personalization on user experience disappeared as soon as participants started reading articles. A possible explanation for this observation can be the number of articles people see in the second part that they have already read in the first part. Seeing articles one has already read may contribute to a higher perceived level of personalization and satisfaction with the library as a whole, while reading these articles might actually be detrimental for the user experience. In other words, what looks good might not necessarily be what helps the user and as such it might be worthwhile to investigate the factors that influence user satisfaction of a personalized system before and after consumption and to see if and how these are different from each other. From a more general perspective this raises the question whether and how personalization needs to anticipate possible changes and differences in the perception of recommendations as the user progresses. Alternatively it might indicate that the process of evaluating personalization is different and depends on whether the user is evaluating through observing or through experiencing.

### Shortcomings and Future Work

While the findings of this study indicate that using surveys as a basis for personalization can improve personalized systems, the specific application in which we tested our hypotheses might limit the extent to which this finding can be generalized. Participants in our study interacted with the system twice. One time for an initial data collection and a second time for the evaluation. This difference might have lead to a discrepancy, as in the first session people were exploring the system and possibly paying attention to other aspects than in the second session. For example, in the first session people were getting used to the way of navigation in the library and getting acquainted with the system and its usability may have been an issue. In the second session, participants are more likely to have evolved past this stage, and they can now focus more on what it is that they want to read. This would imply that data in the first session is describing behavior of participants who are getting to know a system, and as a result models trained on this data will generate recommendations based on what an exploring user will typically read, which may not be appropriate to personalize a library for a participant who already knows and is actively using a system.

As mentioned in the results section, it is unsure how our findings hold up in a setting with a bigger library and more interaction data (both in terms of number of users and in terms of interactions per user). With only 102 articles in a fixed order, behavior for participants in the initial data collection may not have differed enough from each other (yet) to allow the personalization based on reading behavior to produce predictions that are personalized sufficiently. The fact that these personalizations stayed relative close to the non-personalized baseline can be interpreted this way. The survey-based recommendations on the other hand combined data from users with similar parenting styles and as a result were able to differentiate themselves more from the non-personalized baseline. Having more articles and perhaps also a somewhat longer initial period will allow for behavior with more differences between users, allowing to more effectively leverage the predictive power and complexity of reading-based personalization, which in turn will provide more insight into the conditions that play a role

in how personalization based on behavior compares to personalization based on psychological traits. However, our results show that in this situation with limited reading data a short survey delivers good data for initial personalization.

In line with the previous argument, it is important to realize that in terms of data per user, our participants only interacted with the system once and read 2.23 articles on average. They might still have been in their cold start phase and there may not have been enough information about the users' reading behavior to provide useful recommendations. What argues against this is that both the hybrid and reading-based models had higher prediction accuracy than the survey-based recommendations. Given these observations it would be worthwhile to perform a study that controls for the amount of feedback collected from the participants. Having more feedback per participant allows to investigate how the number of interactions per user affects the performance of the different personalization approaches, similar to how Kluver and Konstan [10] investigated the effects of number of interactions on predictive accuracy.

Apart from the amount of data per user, the amount of data available within the system as a whole may be another factor that plays a role which method of personalization works best. Evaluating how survey-based and reading-based personalization compare over time, as more data enter the system as a whole or per user, would provide valuable insight in which approach works best when. One could imagine a system that starts out from personalization based on measured psychological traits that transitions into a system based more on behavior or a hybrid system. Investigating this effect would require a more longitudinal study, where users are invited to a personalized library at multiple moments, to see whether and how the different approaches are affected by the cold start.

Apart from the drawback of a low number of participants for calculating relevance predictions, the low number also limited the statistical power of our statistical analysis of the effects of personalization. While young parents are active on the internet, they are hard to approach. In the current study we did not manage to detect effects of personalization on reading behavior and only differences between some of the experimental conditions. The effects caused by the personalization might have been smaller than the statistical power of our analysis allows us to detect. Conducting a study with more participants would allow us to detect these possibly smaller effects.

In conclusion, the current paper demonstrates that measuring psychological traits for the sake of personalization is worthwhile and might well lead to increased user satisfaction, but additional work is needed to establish under which conditions this approach is valuable.

## REFERENCES


1. B Arnott and Amy Brown. 2013. An Exploration of Parenting Behaviours and Attitudes During Early Infancy: Association with Maternal and Infant Characteristics. *Infant and Child Development* 22 (2013), 349–361. DOI: http://dx.doi.org/10.1002/icd.1794
2. Diana Baumrind. 1966. Effects of Authoritative Parental Control on Child Behavior. *Child Development* 37, 4 (dec 1966), 887. DOI:http://dx.doi.org/10.2307/1126611
3. Jay Belsky and Sara R. Jaffee. 2015. The Multiple Determinants of Parenting. In *Developmental Psychopathology*. Number April. John Wiley & Sons, Inc., Hoboken, NJ, USA, 38–85. DOI: http://dx.doi.org/10.1002/9780470939406.ch2
4. Gianluca Demartini, Paul-Alexandru Chirita, Ingo Brunkhorst, and Wolfgang Nejdl. 2008. Ranking Categories for Web Search. In *Advances in Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, 564–569. DOI: http://dx.doi.org/10.1007/978-3-540-78646-7_56
5. Ignacio Fernández-Tobías, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Iván Cantador. 2016. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction* 26, 2-3 (jun 2016), 221–255. DOI:http://dx.doi.org/10.1007/s11257-016-9172-z
6. Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. MyMediaLite: A Free Recommender System Library. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011)*.
7. Panagiotis Germanakos and Marios Belk. 2016. *Human-Centred Web Adaptation and Personalization*. Springer International Publishing, Cham. 336 pages. DOI: http://dx.doi.org/10.1007/978-3-319-28050-9
8. Asela Gunawardana and Guy Shani. 2015. *Evaluating Recommender Systems*. Springer US, Boston, MA, 265–308. DOI: http://dx.doi.org/10.1007/978-1-4899-7637-6_8
9. J. R. Hauser, G. L. Urban, G. Liberali, and M. Braun. 2009. Website Morphing. *Marketing Science* 28, 2 (mar 2009), 202–223. DOI: http://dx.doi.org/10.1287/mksc.1080.0459
10. Daniel Kluver. 2012. How Many Bits Per Rating ? *Proceedings of the 6th ACM conference on Recommender systems - RecSys '12* (2012), 99–106. DOI: http://dx.doi.org/10.1145/2365952.2365974
11. Bart P. Knijnenburg and Martijn C. Willemsen. 2015. Evaluating Recommender Systems with User Experiments. In *Recommender Systems Handbook*. Springer US, Boston, MA, 309–352. DOI: http://dx.doi.org/10.1007/978-1-4899-7637-6_9
12. Yehuda Koren. Factorization Meets the Neighborhood : a Multifaceted Collaborative Filtering Model. (????). DOI: http://dx.doi.org/978-1-60558-193-4/08/08
13. Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* (2009), 42–49.
14. Robert R. McCrae, Jr. Paul T. Costa, and Thomas A. Martin. 2005. The NEOâĂŞPIâĂŞ3: A More Readable

Revised NEO Personality Inventory. *Journal of Personality Assessment* 84, 3 (2005), 261–270. DOI: http://dx.doi.org/10.1207/s15327752jpa8403_05 PMID: 15907162.

15. Bamshad Mobasher. 2007. Data mining for web personalization. *The adaptive web* (2007), 90–135. http://link.springer.com/chapter/10.1007/978-3-540-72079-9

16. Stephanie L Prady, Kathleen Kiernan, Lesley Fairley, Sarah Wilson, and John Wright. 2014. Self-reported maternal parenting style and confidence and infant temperament in a multi-ethnic community: Results from the Born in Bradford cohort. *Journal of Child Health Care* 18, 1 (2014), 31–46. DOI: http://dx.doi.org/10.1177/1367493512473855

17. Steffen Rendle, Wolf Huijsen, and Karen Tso-Sutter. 2008. *State-of-the-art Recommender Algorithms*. Technical Report. www.mymediaproject.org

18. Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 02* 46, Sigir (2002), 253–260. DOI: http://dx.doi.org/10.1145/564376.564421

19. K. R. Venugopal, K. G. Srinivasa, and L. M. Patnaik. 2009. *Algorithms for Web Personalization*. Springer Berlin Heidelberg, Berlin, Heidelberg, 217–230. DOI: http://dx.doi.org/10.1007/978-3-642-00193-2_10

20. Tiange Zhao. 2016. *Investigating the relationship between parenting beliefs and parenting practice for in-app personalization*. Master thesis. Eindhoven University of Technology. https://pure.tue.nl/ws/files/46944250/855031-1.pdf

21. Zheng Zhu. 2011. *Improving Search Engines via Classification*. Ph.D. Dissertation.