# *Can a Recommender System Support Treatment Personalisation in Digital Mental Health Therapy? A Quantitative Feasibility Assessment Using Data from a Behavioural Activation Therapy App*

# Can a Recommender System Support Treatment Personalisation in Digital Mental Health Therapy? A Quantitative Feasibility Assessment Using Data from a Behavioural Activation Therapy App

Robert Lewis
MIT Media Lab, Massachusetts Institute of Technology
Cambridge, MA, USA
roblewis@media.mit.edu

Craig Ferguson
MIT Media Lab, Massachusetts Institute of Technology
Cambridge, MA, USA
fergusoc@media.mit.edu

Chelsey Wilks
Department of Psychological Science, University of Missouri-St Louis
St. Louis, MO, USA
chelseywilks@umsl.edu

Noah Jones
MIT Media Lab, Massachusetts Institute of Technology
Cambridge, MA, USA

Rosalind W. Picard
MIT Media Lab, Massachusetts Institute of Technology
Cambridge, MA, USA

## ABSTRACT

Recommender systems have the potential to improve the user experience of digital mental health apps. Personalised recommendations can help users to identify therapy tasks that they find most enjoyable or helpful, thus boosting their engagement with the service and optimising the extent to which it helps them to feel better. Using a dataset containing 23,476 ratings collected from 973 players of a mental health therapy game, this work demonstrates how collaborative filtering algorithms can predict how much a user will benefit from a new therapy task with greater accuracy than a simpler baseline algorithm that predicts the average rating for a task, adjusted for the biases of the specific user and specific task. Collaborative filtering algorithms (matrix factorisation and k-nearest neighbour) outperform this baseline with a 6.5-8.3% improvement in mean absolute error (MAE) and context-aware collaborative filtering algorithms (factorisation machines) outperform with a 7.8-8.8% improvement in MAE. These results suggest that recommender systems could be a useful tool for tailoring recommendations of new therapy tasks to a user based on a combination of their past preferences, the ratings of similar users, and their current context. This scalable approach to personalisation – which does not require a human therapist to always be in-the-loop – could play an important role in improving engagement and outcomes in digital mental health therapies.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Applied computing** → **Psychology**; **Health care information systems**.


*CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA*

ACM ISBN 978-1-4503-9156-6/22/04…$15.00
https://doi.org/10.1145/3491101.3519840

## KEYWORDS

digital mental health, recommender systems, behavioural activation, collaborative filtering, factorisation machines, machine learning

**ACM Reference Format:**
Robert Lewis, Craig Ferguson, Chelsey Wilks, Noah Jones, and Rosalind W. Picard. 2022. Can a Recommender System Support Treatment Personalisation in Digital Mental Health Therapy? A Quantitative Feasibility Assessment Using Data from a Behavioural Activation Therapy App. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3491101.3519840

## 1 INTRODUCTION

The demand for mental health services is greatly outpacing the supply of trained mental health professionals [7]. Digital mental health – i.e., the provision of psychological treatment through digital channels – has been touted as an accessible and scalable route to address this heightened interest. By translating psychotherapies into digital formats, consumers of mental health services (i.e., clients) can access services and engage with interventions without requiring human therapists to be *in-the-loop*, thus allowing many more clients to be treated.

However, several challenges exist when creating digital versions of psychotherapies, and treatment *personalisation* – i.e., adapting a therapy to the specifics of an individual – is a prominent one. Many psychotherapies consist of a large catalogue of treatment components, for which the optimal sequencing of these items is not fixed across clients, but rather should be adapted to their medical history, preferences, and context. Examples of such psychotherapies include behavioural activation (BA) – commonly used for treating mood disorders such as depression and in particular the subtype of anhedonia [9, 21] – and dialectical behavioural therapy (DBT) – considered the *gold-standard* for treating borderline personality disorder and with established evidence for reducing suicidal behaviour [11, 24, 31]. In face-to-face treatments, a therapist will adjust the content, timing, and dosage of therapy tasks to clients based on ongoing assessments; in their digital formats,

one must find a way to algorithmically mimic this therapy curation. In the absence of good therapeutic curation, clients are left to their own devices to identify their own therapy tasks, which may lead to disengagement or degeneration. Specifically, if a client cannot find therapy tasks that work well for them they are likely to lose interest or belief in the therapy and churn from the app, thus foregoing treatment that might help them to improve their mental health. This lack of personalisation – which is thought to foster client engagement with digital therapy [25, 33, 34] – may in part explain why retention rates with digital mental health apps are so low, with a recent review of 93 mental health apps (with median total installs of 100,000) suggesting median Day-N retention rates were only 3.9% (interquartile range, IQR: 10.3%) and 3.3% (IQR: 6.2%) at 15 days and 30 days after installation, respectively [6].

Recommender systems (RS) hold promise as a technology to provide personalisation in therapy apps. Their utility in social media and retail contexts is well-known, where their algorithms are primarily tuned to increase engagement and consumption by recommending personalised lists of items to users based on their preferences, context, and what similar users have enjoyed [14, 39, 40]. While RS algorithms in these contexts are primarily optimised to benefit the system's implementors (e.g., by increasing page visits and thus *pay-per-click* advertising revenues), we propose that, with careful redesigns of their optimisation criteria, these same algorithms can be modified to primarily benefit the system users. For example, in the context of digital mental health – where user outcomes should be prioritised over revenue generation – an RS algorithm could instead be optimised to recommend therapy tasks to clients that might help them to feel happiest, most productive, or most relaxed given their current context. On finding such tasks, a client is more likely to improve their well-being in the short-term and persist with the treatment in the long-term, thus increasing the likelihood that they recover to a state of better mental health.

Looking further ahead, the concept of digital *micro interventions* for mental health has been proposed [5]. *Micro interventions* have highly focused objectives (e.g., completing a brief gratitude exercise to increase in-the-moment appreciation, or performing a short mindful breathing exercise to increase self-awareness) and it is suggested that they should be self-contained, such that a client can engage with them without needing to work through an overarching therapy framework. Indeed, many of the therapy tasks in BA and DBT can be considered *micro interventions*, though the concept is therapy agnostic in the sense it allows many items from different therapies to be combined into a single catalogue. Thus, one can foresee a future where platforms (e.g., *digital therapeutic pharmacies* or *digital apothecaries* [29]) offer an array of *micro interventions* to clients from a variety of different providers. As more and more items are added to their catalogues, the burden of choice increases on their clients, likely resulting in information overload and users either choosing suboptimal items or churning. Analogous shifts led to the proliferation of online entertainment content (e.g., news, movies, music) which resulted in the first generation of recommender systems as information retrieval systems that helped users choose what to consume. Hence, it seems reasonable that a new generation of RS – that are carefully designed to benefit their users – could provide important therapy personalisation services in this next generation of mental health care.

Therefore, assessing the viability of a recommender system that can curate therapy items in both individual therapies such as BA and DBT, and in *digital therapeutic pharmacies*, is the focus of this paper. While previous work have assessed specific *content-based* and *contextual bandit* RS algorithms on smaller mental health datasets (e.g., [4, 32, 38]), this work is the first to present the accuracy of various *collaborative filtering* algorithms on a large behavioural activation therapy dataset (>20,000 ratings from ≈1,000 users). We conduct an offline assessment and discuss how its results suggest that there is promise in using collaborative filtering for therapy curation. We hope these results will stimulate discussion on how further techniques and concepts from the RS community can be modified to improve engagement and outcomes in digital mental health. As future work, we intend to implement the best performing algorithms from this offline analysis into a live recommender system that will be evaluated in a user study.

## 2 RELATED WORK

### 2.1 Recommender System Algorithms

Recommender systems can be considered as having *collaborative filtering* (CF), *content-based*, *knowledge-based*, and *context-aware* architectures, where in practice the best performing algorithms are often a hybrid of these components [3]. *Collaborative filtering* algorithms make recommendations to the target user based on how their rating history compares to those of other users. Popular approaches include neighbourhood-based methods such as k-nearest neighbours and model-based methods such as matrix factorisation. *Content-based* systems make recommendations to users based on the properties of the items they have rated highly in the past, and *knowledge-based* systems make recommendations using decision rules programmed *a priori* by a human expert. It is often found that CF algorithms outperform content-based and knowledge-based algorithms when there are at least a few ratings per user and per item in the ratings matrix (i.e., it is not a *cold-start* scenario). Finally, *context-aware* recommender systems take into account the context of users when they consume and rate items and use this information to refine recommendations in future (e.g., if it is 9am on a Monday vs. 12pm on a Saturday, what should I recommend to the user?). Several paradigms exist to make an algorithm *context-aware* including *pre-filtering*, *post-filtering*, and *contextual modeling*, where the former two effectively filter the input data or predictions before or after using a non-context-aware algorithm, while the latter actually incorporates the contextual information into the rating prediction function [3]. *Factorisation machines* are an archetypal algorithm for *contextual modeling*, which combine the benefits of collaborative filtering with the ability to incorporate both context information and characteristics of the users and/or items [36]. Deep learning methods have further improved the accuracy of CF and context-aware RS algorithms [15, 44, 46]. Finally, *contextual bandits* are growing in popularity [3, 22] as a type of RS that permits *online learning* of user preferences (i.e., updating their parameters while continually serving recommendations to a user); they are particularly beneficial in scenarios where new users and items constantly arrive in the system, e.g., news articles on web pages.

## 2.2 Recommender Systems for Mental Health

Recent works have considered the role of RS in the context of mental health though the field is still nascent. For example, *IntelliCare* – a suite of 12 apps for depression and anxiety – employed a basic RS in their "Hub" that suggests a short list of apps to users on a weekly basis, where the recommendations are made at random[1] [8, 27]. *PopTherapy* is another relevant system, that uses *contextual bandits* to recommend interventions derived from popular web apps to users (e.g., learn about active constructive responding by watching this YouTube video with a friend) with the goal to reduce their stress levels [32]. *MOSS* recommends cognitive behavioural therapy micro interventions (CBT) to reduce depression and takes into account user context features such as day, location and smartphone usage [43]. Finally, *MUBS* supports behavioural activation therapy and uses item ratings as well as item content features to tailor recommendations, and it was found to motivate users in a study involving a cohort of 17 patients with depression [38]. An offline analysis of *contextual bandit* RS algorithms was also recently performed on historical data from 114 users of an emotion regulation app, reporting benchmarks on algorithms and contextual features that might lead to optimal RS performance in this therapy context [4]. A recent paper provides further perspective on why recommender systems could be vital tools for improving engagement and outcomes with digital mental health interventions [42], and a review paper further surveys instances where recommender systems have been implemented in digital mental health and broader *mHealth* contexts [10]. We also note that RS algorithms are often discussed in the context of the mHealth *just-in-time-adaptive-intervention* (JITAI) framework [30]. Our research builds on this prior literature and is the first to benchmark the performance of several different *collaborative filtering* RS algorithms on a large behavioural activation dataset (>20,000 ratings from ≈1,000 users).

# 3 PROOF-OF-CONCEPT EXPERIMENTS

## 3.1 Data from a Therapy App

The experiments in this paper use data from a free mobile game with an embedded behavioural activation therapy module that gives players in-game rewards for regularly choosing and completing real-world therapy tasks [12]. The game was released publicly in 2020 and has since registered over 10,000 users. In the game, players are prompted on a daily basis to select a therapy task from a list of 76 different options from categories of "Basics", "Fitness", "Fun", "Social", "Art" and "Other" (e.g., "take a shower", "do 5 mins of stretching", "write a diary entry", "text a friend or family member", etc.). On completing the task, players are asked to rate to what extent it helped them to improve their mood, answering the question "How do you feel after your adventure?" on a 5-point scale of: "Worse", "Not As Good", "The Same", "A Little Better", "Much Better". In addition to logging their rating of the task, the timestamps of when the user chose the task and when they rated it are logged by the game. This allows the calculation of several context features related to time (e.g., is it morning, afternoon, evening, or night-time? And is it a weekday or a weekend?). Furthermore, metadata on the different therapy tasks are available, including their category, their perceived effort (low, medium or high), and how long they are expected to take (3-60 minutes). However, the game has a strict privacy policy[2] and so no identifying information was collected from the players, including no demographics or identifiable locations[3]. The detailed design of the game is presented in past work [12]. For the purpose of the RS analysis presented here, Table 1 summarises the properties of the dataset. Furthermore, distributions of the ratings are shown in Figure 1a, with the user and item long-tails displayed in Figures 1b and 1c, respectively.

## 3.2 Experimental Settings

*3.2.1 Data Sampling and Evaluation Approach.* The experiments of this paper use the task rating described in Section 3.1 as an *explicit feedback* rating, with values on a 5-point scale from 1-5 where 1 corresponds to "Worse" and 5 corresponds to "Much Better" (see Section 3.1 for full rating scale). Accordingly, the utility of the RS algorithms are assessed for their ability to accurately predict missing ratings in the *matrix completion problem*, a standard offline evaluation paradigm in the RS literature that has been used in large competitions such as the *Netflix Prize* [3, 20]. To this end, the generalisation error of the algorithms are calculated by segmenting the data into training, validation and testing sets. The user-item ratings are sorted by timestamp and then the last 2 items for each user are added to the testing set, the next 2 from last are added to the validation set (which is used to select optimal hyperparameters), and the remaining ratings are used to train the algorithm (this *user-based temporal split* of rating data is a standard approach [26]). Given users can perform a task more than once – which, for example, may result in a rating for that task being present in both the training and testing set – we take further measures to filter the testing and validation set so that there is no *leakage* between them[4]. We first filter the testing set to only contain ⟨user, task⟩ pairs not present in the training set. We then filter the validation set so that it only contains ⟨user, task⟩ pairs not present in either the training or testing set. As such, the RS algorithms are assessed for their ability to use the training data to accurately predict the ratings of tasks that a user has not tried before.

The mean absolute error (MAE) and root mean squared error (RMSE) are reported for each algorithm (given the latter squares the error, it gives relatively higher weights to large errors). The experiment is performed 10 times and the errors are reported as the average error value over these 10 experiments. This controls for the effect of random initialisation of model parameters on model performance. Hyperparameter tuning was performed and details of this are provided in Section 3.2.2. Furthermore, before sampling as described above, the data was filtered to only contain users who

[1] Though it is noted that the long-term plan for this work is to make recommendations with RS algorithms that use logs of user data to identify apps that the person will most likely use and find useful [27].

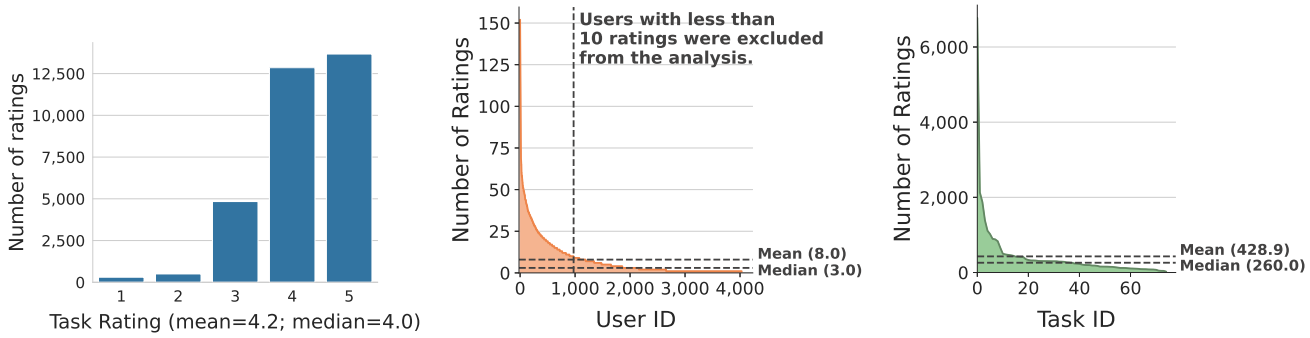
[2] https://guardians.media.mit.edu/privacy/

[3] The only location information stored is the timezone of the user's device, which allows the extraction of the aforementioned time-of-day features.

[4] NB: this is an additional notion of *leakage* separate to the standard one we must always control for in machine learning experiments. Indeed one should expect that a user's rating of the same therapy task at different times will vary considerably by their context and so controlling for this leakage is arguably not essential, especially if there are context features in the model. However, given the intention in this experiment is to test how accurately these algorithms predict ratings for tasks a user has not tried before, we chose to control for it anyway.

**Table 1: Properties of the rating data from the behavioural activation therapy game. The sets $U$, $I$ and $R$ are the different users, therapy task ratings (i.e., items), and unique ratings identified by <user, item, timestamp>, respectively. The analysis in this paper is conducted with Dataset II so that sufficient data can be allocated to the validation and testing sets while still giving the models some data from each user to be trained on (i.e., avoiding the *cold-start* scenario).**

| Dataset | $\|U\|$ | $\|I\|$ | $\|R\|$ | $\frac{\|R\|}{\|U\|}$ | $\frac{\|R\|}{\|I\|}$ |
|---|---|---|---|---|---|
| I. All Ratings Data | 4,032 | 75 | 32,171 | 8.0 | 428.9 |
| II. Data From Users with 10 or More Ratings | 973 | 75 | 23,476 | 24.1 | 313.0 |




**(a) Therapy task rating distribution (i.e., number of tasks rated at each level)**

**(b) User "long-tail" (i.e., number of task ratings by user)**

**(c) Item "long-tail" (i.e., number of ratings by therapy task)**

**Figure 1: Distributions of the rating data from the therapy app: (a) item ratings are clearly skewed towards values of 4 ("I feel a little better") and 5 ("I feel much better"), which is not uncommon in other RS contexts; (b) some users rate many more items than others; and (c) similarly, some items are much more popular than others (i.e., they are performed and rated far more frequently). Please see Section 3.2.1 for a discussion of why some user ratings were excluded.**

provided 10 or more task ratings (reducing the data to 973 users and 23,476 ratings) so that all users had at least 6 ratings in the training set. Ensuring each user has at least 6 ratings in the training set ensures the algorithms are not attempting to predict for users they have never seen before in the testing set, a setting referred to as the *cold-start* scenario which is a well-known challenge for RS algorithms and is typically evaluated in separate experiments (which are out of scope for this paper). Ratings for the "Other" therapy task – where a player is free to choose what they do – are also excluded from the analysis given it is not known what players do if they chose this and thus it would be hard to predict accurately.

It is important to briefly comment on the offline nature of the analysis presented here. The properties of RS algorithms can be assessed in both online and offline settings, where online evaluation protocols consist of a user study (e.g., A/B testing algorithms to understand their effect on user outcomes), while offline evaluations are cheaper and faster to conduct as they only require historical rating data from users [3, 4]. The online assessment of an algorithm will often lead to the clearest understanding of its suitability for a group of users, as one can ask clarifying questions about the items they were recommended (e.g., were they helpful, understandable, actionable etc.), and furthermore one can test how consuming recommended items changed distal outcome variables (e.g., does completing the recommended tasks help clients to improve their their mental health over time). However, offline assessments of algorithms still answer important questions about the viability of RS algorithms (e.g., can they accurately predict how much a user will like an item) and, as they are easier to conduct, they are often an important first step in evaluating candidate RS algorithms before they are subsequently assessed in online settings. Therefore, the results presented in this paper provide a first step towards evidencing if RS algorithms are a viable solution in mental health therapy. However, further design and online assessments would be required to create a recommender system that could go live in this context.

*3.2.2 Models Assessed.* Two simple baseline models are assessed to contextualise the performance of the more advanced algorithms.

(1) ***Random***: represents drawing random samples from a normal distribution defined by the mean and standard deviation of the rating values in the training set.
(2) ***BaselineOnly***: a more advanced baseline, where the prediction for a specific user-item pair is adjusted to take into account if the user and/or item give or receive systematically higher or lower ratings than the mean (i.e., it adjusts for bias). It is defined by $\hat{r}_{ui} = \mu + b_u + b_i$ where $\hat{r}_{ui}$ is the predicted rating, $\mu$ is the training set mean, and $b_u$, $b_i$ are the training set biases for user $u$ and item $i$, respectively, [3, 19].

A set of *collaborative filtering* algorithms is then assessed, including two that can incorporate context data:

(3) **k-nearest neighbour (*KNN*)**: which estimates the missing ratings by taking the similarity-weighted sum of the $k_{nn}$ nearest neighbours to the query rating.

(4) **k-nearest neighbour with user baseline adjustment (*KNN-wBaseline*)**: this is an extension of the *KNN* model but with additional adjustments for the user and item biases, $b_u$ and $b_i$, [3, 19].
(5) **Funk singular value decomposition (*SVD*)**: is implemented per the Funk SVD / matrix factorisation (MF) paradigm with adjustments for the user and item biases, $b_u$ and $b_i$ [13, 20]. The model learns by minimising a squared error function and its hyperparameters are the number of latent factors ($k_{latent}$), the regularisation parameter ($\lambda$), the learning rate ($\gamma$), and the number of training epochs ($e$).
(6) **SVD++ algorithm (*SVD++*)**: an extension of *SVD* [18] that incorporates information on the tasks the users rated irrespective of rating value (i.e., *implicit feedback*) to learn additional latent factor parameters in addition to those parameters learned for the *explicit* rating values (cf. *SVD* above).
(7) **Factorisation machine with context (*FM-Context*)**: factorisation machines can be considered as a generalisation of *SVD/MF* where more than 2 variables can be included as predictors [36]. They are thus an elegant solution for incorporating additional contextual variables into a collaborative filtering model. A *two-way* (i.e., degree $d=2$) factorisation machine is used for the *FM-Context* model. Let $\hat{r}_{uic}$ be the rating to predict given the user $u$, item $i$, and context $c$, and $\boldsymbol{x}$ be a vector of binary predictors including the item ID, user ID, and different context and item attribute features. $\boldsymbol{w}$ and $\mathbf{v}$ are then parameters that the model can learn, with the former corresponding to biases for the $n$ features (+1 for the global bias, $w_0$) and the latter corresponding to the latent factors ($k$ dimensional vectors, with a vector for each unique feature value for the $n$ features):

$$\hat{r}_{uic} = \hat{y}(\boldsymbol{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j \quad (1)$$

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle := \sum_{f=1}^{k} \mathrm{v}_{i,f} \cdot \mathrm{v}_{j,f} \quad (2)$$

**These context features are included in *FM-Context***: time of day (morning/afternoon/evening/night-time); day of week (Mon-Sunday); and is weekend (yes/no). Additionally, the following item attributes are included: task category ("Basics", "Fitness", "Fun", "Social", "Art"); effort level of task (low/medium/high); and length of task (<10mins, 10-30mins, ≥30mins). Its hyperparameters are the same as *SVD*.
(8) **Field-aware factorisation machine with context (*FFM-Context*)**: an extension to *FM-Context* which allows a latent factor vector to be learned for each feature value's interaction with all other *fields* thus giving the model more flexibility in the relations it can learn between its inputs (where a field can be understood as a feature column, and instances of this feature column may assume different feature values) [17]. **This model uses the same context features as *FM-Context* and has the same hyperparameters**.

Given that users can perform a therapy task more than once, there can be more than one rating for a given ⟨user, task⟩ pair in the training data[5]. Given that algorithms (2-6) expect a single rating for each ⟨user, task⟩ pair, the training ratings are averaged by each ⟨user, task⟩ pair. Algorithms (7-8) do not require this adjustment to the training data. Hyperparameter tuning was performed using the error on the held out validation set to choose the optimal hyperparameters. The value of the nearest neighbours was tuned in *KNN* and *KNN-Baseline* ($k_{nn}$ with range: [2, 500]) using a grid search. Using a *mean-squared difference* or *Pearson* (with shrinkage) similarity metric, as well as a *user-based* or *item-based* approach to similarity, were also dimensions considered in the *KNN* and *KNN-Baseline* grid. The value of the number of latent factors ($k_{latent}$ in the range [2, 256][6]), the regularisation parameter ($\lambda$ in [0, 1]), the learning rate ($\gamma$ in [0.0001, 0.2]), and the number of training epochs ($e$ in [2, 200]) were tuned in the *SVD*, *SVD++*, *FM-Context* and *FFM-Context* algorithms using *bayesian optimisation* with 200 optimisation steps. Optimal hyperparameters were selected on the basis of the lowest MAE instead of the lowest RMSE. *SGD* was used for algorithms (5-6) and *Adagrad* for (7-8). *FM-Context* and *FFM-Context* were implemented using the *xLearn* C++ library [2] and all other models used the Python *Surprise* library [16]. Bayesian optimisation was performed with *scikit-optimize* [1].

## 4 EXPERIMENTAL RESULTS

Table 2 summarises the experimental results. It can be seen that the *BaselineOnly* model is the better performing baseline algorithm and that all of the collaborative filtering algorithms considerably outperform this baseline on MAE (by 6.5-8.8%), and also perform well on RMSE with all but one beating the baseline (by 0.5-3.0%). The notably larger improvement in MAE compared to the improvement in RMSE might be attributable to the fact that the RMSE is more sensitive to large deviations which often result from outliers (e.g., a rating that is many standard deviations from the mean rating for a user-item pair). This might suggest that further work could be done to make these RS algorithms robust to outliers, but nonetheless the improvement in accuracy per both metrics is encouraging. Table 2 further suggests that higher accuracy can be achieved when the contextual information – e.g., when a task was chosen and the characteristics of this task – is incorporated into the RS algorithm. This is evidenced by the improvement in accuracy of a further 50 basis points in Δ%MAE between the best performing non-context-aware RS algorithm, *KNN-wBaseline* (8.3%), and the best performing context-aware algorithm, *FM-Context* (8.8%).

## 5 DISCUSSION

The purpose of this study was to determine if it is feasible to use collaborative filtering RS algorithms to recommend mental health therapy tasks, given their benefits in many domains (e.g., achieving state-of-the-art accuracies) yet their apparent sparsity in the contemporary RS for digital mental health literature (which has instead focused on e.g., content-based and contextual bandit paradigms). The results from the offline evaluation show that these algorithms are consistently more accurate than simpler random or user-item baseline models when predicting the ratings of new therapy tasks for a client. Given the nature of CF algorithms is to use item ratings

[5]Note the exclusions already applied to testing & validation sets in Section 3.2.1.
[6]This is interval notation where square brackets indicate the interval is inclusive.

**Table 2: Errors of the different recommender system algorithms on the held-out testing set ratings using Dataset II from Table 1. MAE is mean absolute error and RMSE is root mean squared error; Δ and Δ% are the absolute and percentage change in the error relative to the *BaselineOnly* model. All models are defined in Section 3.2.2. Hyperparameter tuning was performed and the results are averaged over 10 runs of the experiment to control for bias from the initialisation of model parameters.**

| Model | MAE | ΔMAE | Δ%MAE | RMSE | ΔRMSE | Δ%RMSE |
|---|---|---|---|---|---|---|
| Random | 0.888 | 0.223 | 33.5% | 1.134 | 0.280 | 32.8% |
| BaselineOnly | 0.666 | 0.000 | 0.0% | 0.854 | 0.000 | 0.0% |
| KNN | 0.616 | -0.049 | -7.4% | 0.857 | 0.003 | 0.3% |
| KNN-wBaseline | 0.611 | -0.055 | -8.3% | 0.850 | -0.004 | -0.5% |
| SVD | 0.622 | -0.043 | -6.5% | 0.839 | -0.015 | -1.7% |
| SVD++ | 0.621 | -0.044 | -6.7% | 0.839 | -0.015 | -1.7% |
| FM-Context | **0.607** | **-0.059** | **-8.8%** | **0.829** | **-0.025** | **-3.0%** |
| FFM-Context | 0.614 | -0.052 | -7.8% | 0.830 | -0.025 | -2.9% |

from similar users to predict rating values for a specific user, this finding suggests that there is value in using the therapy task ratings from a group of clients when making recommendations for a specific client. For example, a CF recommender system might generate a recommendation of "clients with similar preferences to you found this task you haven't tried before really effective at improving their mood, would you like to try it now?". Given their accuracy benefits – as well as their other strengths over other RS paradigms (e.g., more diverse recommendations versus content-based models) – we thus propose that CF algorithms should be considered in the solution space for digital mental health recommender systems, alongside the alternative RS approaches assessed elsewhere.

It is further notable that incorporating context (e.g., time of day, day of week) and item attributes (e.g., effort level, category) into the CF algorithms further improved their accuracy, suggesting that this ancillary information might be useful in understanding if a client will enjoy a certain task at a certain moment. However, we note that the difference is only slight – 50 basis points between the best performing algorithms – and thus we are cautious about strongly concluding on the utility of this contextual data at this stage in the investigation (e.g., as slight differences in algorithmic performance can result due to specifics of the sample data and/or stochasticity in the training process, though the protocol used here tries to mitigate for the latter by repeating the experiment 10 times). Nonetheless, future work will incorporate additional data about context, item attributes, and anonymised user attributes – e.g., features derived from a user's behaviour during the first 1-2 weeks of using the app – into the algorithms to see if this further improves their accuracy.

Given the analysis presented here follows the offline assessment paradigm, it is important to discuss how these findings might be translated into the design of a live system. A common development pattern would be to take the algorithm trained on offline data and extend it to generate a ranked list of items that a user might enjoy given their context and past ratings. The top-K items from this list could then be filtered into a short-list which is presented to the user through an interface, where this short-list could, for example, consist of the top-K recommended items a user has not tried before. With this algorithmic protocol in place, the recommender system could then be evaluated in online user studies that assess its impact on the desired outcomes (e.g., user well-being), as well as understanding the user experience of receiving recommendations (e.g., do users want more variety or novelty in the shortlist of items recommended to them?). This feedback could then be incorporated into future versions of the underlying algorithm. The cadence of receiving recommendations (e.g., weekly, daily, instantaneously, etc.) would also need to be decided and should be based on both the context of use and user preferences (e.g., would hourly recommendations frustrate or demotivate a user?). Finally, two modes of operation might be considered. Firstly, a "direct-to-client" system, where clients receive a list of suggestions directly from the algorithm. Secondly, a "Stitch Fix" model, where a recommender system suggests a list of items to a human therapist, who then sense checks them versus their understanding of their client's needs and modifies them if required. This latter mode might be preferable in sensitive mental health contexts, for example where clients are patients with severe mental health symptoms such as suicidal behaviour.

## 6 FUTURE WORK

This work can be extended in various ways. Firstly, additional data scenarios (e.g., cold-start) and feature ablations should be assessed. Secondly, future work should take the most accurate RS algorithms identified from offline analyses like the one presented here and assess them in a user study. Furthermore, extensions to the algorithms might result in higher accuracy and/or other desirable properties (e.g., ability to generate a shortlist of recommendations with more variety). For example, graph neural networks have recently achieved state-of-the-art accuracy on RS benchmark datasets [44] and can be extended to take into account context [45]. Moreover, the objective functions of RS training procedures can be modified to change the behaviour of the algorithm. For example, a pairwise "learning to rank" loss [3, 37] could be used instead of the pointwise "mean squared error" losses used in the *SVD*, *SVD++*, *FM-Context*, and *FFM-Context* algorithms of this paper. Furthermore, *multicriteria* functions may be very relevant in this context [28], as therapy tasks may be rated along multiple dimensions (e.g., mood improvement, educational value, etc.). *Online learning* with contextual bandits is another important paradigm to consider for mHealth recommender systems [4, 35, 41], though the dataset here may not be amenable to an offline evaluation of these algorithms (e.g., via

*replay sampling* [23]). Finally, explainability of recommendations is often desirable and various recent methods could address this [47].

## 7 CONCLUSION

This work evidences how collaborative filtering algorithms improve the accuracy of rating predictions for new behavioural activation therapy tasks, using data from 973 users of a mental health therapy app. Future work will design a complete recommender system around the most accurate algorithms from this analysis and assess its benefits in a user study.

## REFERENCES


[1] 2022. *Scikit-Optimize.* https://scikit-optimize.github.io/

[2] 2022. *xLearn.* https://github.com/aksnzhy/xlearn

[3] Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook.* Springer.

[4] Mawulolo K. Ameko, Miranda L. Beltzer, Lihua Cai, Mehdi Boukhechba, Bethany A. Teachman, and Laura E. Barnes. 2020. *Offline Contextual Multi-Armed Bandits for Mobile Health Interventions: A Case Study on Emotion Regulation.* Association for Computing Machinery, New York, NY, USA, 249–258. https://doi.org/10.1145/3383313.3412244

[5] Amit Baumel, Theresa Fleming, and Stephen M. Schueller. 2020. Digital Micro Interventions for Behavioral and Mental Health Gains: Core Components and Conceptualization of Digital Micro Intervention Care. *JMIR* 22, 10 (2020). https://doi.org/10.2196/20631

[6] Amit Baumel, Frederick Muench, Stav Edan, and John M. Kane. 2019. Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *JMIR* 21, 9 (2019), 1–15. https://doi.org/10.2196/14567

[7] Angela Beck, Ronald Manderscheid, and Peter Buerhaus. 2018. The Future of the Behavioral Health Workforce: Optimism and Opportunity. *American Journal of Preventive Medicine* 54 (06 2018), S187–S189. https://doi.org/10.1016/j.amepre.2018.03.004

[8] Ken Cheung, Wodan Ling, Chris Karr, Kenneth Weingardt, Stephen Schueller, and David Mohr. 2018. Evaluation of a recommender app for apps for the treatment of depression and anxiety: An analysis of longitudinal user engagement. *Journal of the American Medical Informatics Association : JAMIA* 25 (04 2018). https://doi.org/10.1093/jamia/ocy023

[9] Pim Cuijpers, Annemieke van Straten, and Lisanne Warmerdam. 2007. Behavioral activation treatments of depression: A meta-analysis. *Clinical Psychology Review* 27, 3 (2007), 318–326. https://doi.org/10.1016/j.cpr.2006.11.001

[10] Robin De Croon, Leen Van Houdt, Nyi Nyi Htun, Gregor Štiglic, Vero Vanden Abeele, and Katrien Verbert. 2021. Health Recommender Systems: Systematic Review. *J Med Internet Res* 23, 6 (29 Jun 2021), e18035. https://doi.org/10.2196/18035

[11] Christopher R. DeCou, Katherine Anne Comtois, and Sara J. Landes. 2019. Dialectical Behavior Therapy Is Effective for the Treatment of Suicidal Behavior: A Meta-Analysis. *Behavior Therapy* 50, 1 (2019), 60–72. https://doi.org/10.1016/j.beth.2018.03.009

[12] Craig Ferguson, Robert Lewis, Chelsey Wilks, and Rosalind Picard. 2021. The Guardians: Designing a Game for Long-term Engagement with Mental Health Therapy. In *2021 IEEE Conference on Games (CoG).* 1–8. https://doi.org/10.1109/CoG52621.2021.9619026

[13] Simon Funk. 2006. *Netflix Update: Try This at Home.* https://sifter.org/~simon/journal/20061211.html

[14] Carlos A. Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (dec 2016), 19 pages. https://doi.org/10.1145/2843948

[15] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web.* 173–182.

[16] Nicolas Hug. 2020. Surprise: A Python library for recommender systems. *Journal of Open Source Software* 5, 52 (2020), 2174. https://doi.org/10.21105/joss.02174

[17] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-Aware Factorization Machines for CTR Prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16).* Association for Computing Machinery, New York, NY, USA, 43–50. https://doi.org/10.1145/2959100.2959134

[18] Yehuda Koren. 2008. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08).* Association for Computing Machinery, New York, NY, USA, 426–434. https://doi.org/10.1145/1401890.1401944

[19] Yehuda Koren. 2010. Factor in the Neighbors: Scalable and Accurate Collaborative Filtering. *ACM Trans. Knowl. Discov. Data* 4, 1, Article 1 (jan 2010), 24 pages. https://doi.org/10.1145/1644873.1644874

[20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[21] C W. Lejuez, Derek R.; Hopko, and Sandra D Hopko. 2001. A Brief Behavioral Activation. *Behavior Modification* 25, 2 (2001), 255–286. https://doi.org/10.1177/0145445501252005

[22] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web.* 661–670.

[23] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased Offline Evaluation of Contextual-Bandit-Based News Article Recommendation Algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11).* Association for Computing Machinery, New York, NY, USA, 297–306. https://doi.org/10.1145/1935826.1935878

[24] Marsha M. Linehan. 2014. *DBT Skills Training Manual, Second Edition.* Guilford Publications.

[25] Jessica Lipschitz, Christopher J Miller, Timothy P Hogan, Katherine E Burdick, Rachel Lippin-Foster, Steven R Simon, and James Burgess. 2019. Adoption of Mobile Apps for Depression and Anxiety: Cross-Sectional Survey Study on Patient Interest and Barriers to Engagement. *JMIR Ment Health* 6, 1 (25 Jan 2019), e11334. https://doi.org/10.2196/11334

[26] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2020. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In *Fourteenth ACM Conference on Recommender Systems.* Association for Computing Machinery, New York, NY, USA, 681–686. https://doi.org/10.1145/3383313.3418479

[27] David Mohr, Kathryn Tomasino, Emily Lattie, Hanah Palac, Mary Kwansy, Kenneth Weingardt, Chris Karr, Susan Kaiser, Rebecca Rossum, Leland Bardsley, Lauren Caccamo, Colleen Stiles-Shields, and Stephen Schueller. 2017. IntelliCare: An Eclectic, Skills-Based App Suite for the Treatment of Depression and Anxiety. *Journal of Medical Internet Research* 19 (01 2017). https://doi.org/10.2196/jmir.6645

[28] Diego Monti and Giuseppe Rizzo. 2021. A systematic literature review of multicriteria recommender systems. *Artificial Intelligence Review* 54 (01 2021). https://doi.org/10.1007/s10462-020-09851-4

[29] Ricardo Muñoz, Denise Chavira, Joseph Himle, Kelly Koerner, Jordana Muroff, Julia Reynolds, Raphael Rose, Josef Ruzek, Bethany Teachman, and Stephen Schueller. 2018. Digital apothecaries: a vision for making health care interventions accessible worldwide. *mHealth* 4 (06 2018). https://doi.org/10.21037/mhealth.2018.05.04

[30] Inbal Nahum-Shani, Shawna Smith, Bonnie Spring, Linda Collins, Katie Witkiewitz, Ambuj Tewari, and Susan Murphy. 2016. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine* 52 (09 2016). https://doi.org/10.1007/s12160-016-9830-8

[31] Patrick Panos, John Jackson, Omar Hasan, and Angelea Panos. 2013. Meta-Analysis and Systematic Review Assessing the Efficacy of Dialectical Behavior Therapy (DBT). *Research on Social Work Practice* 24 (02 2013), 213–223. https://doi.org/10.1177/1049731513503047

[32] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with Stress through Pop-Culture. *Proceedings - PERVASIVEHEALTH 2014: 8th International Conference on Pervasive Computing Technologies for Healthcare.* https://doi.org/10.4108/icst.pervasivehealth.2014.255070

[33] Olga Perski, Ann Blandford, Robert West, and Susan Michie. 2017. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Transl. Behav. Med.* 7, 2 (2017), 254–267. https://doi.org/10.1007/s13142-016-0453-1

[34] Chengcheng Qu, Corina Sas, Claudia Daudén Roquet, and Gavin Doherty. 2020. Functionality of Top-Rated Mobile Apps for Depression: Systematic Search and Evaluation. *JMIR Ment Health* 7, 1 (24 Jan 2020), e15321. https://doi.org/10.2196/15321

[35] Mashfiqui Rabbi, Predrag V. Klasnja, Tanzeem Choudhury, Ambuj Tewari, and Susan A. Murphy. 2019. Optimizing mHealth Interventions with a Bandit. *Studies in Neuroscience, Psychology and Behavioral Economics* (2019).

[36] Steffen Rendle. 2010. Factorization Machines. In *2010 IEEE International Conference on Data Mining.* 995–1000. https://doi.org/10.1109/ICDM.2010.127

[37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09).* AUAI Press, Arlington, Virginia, USA, 452–461.

[38] Darius A. Rohani, Andrea Quemada Lopategui, Nanna Tuxen, Maria Faurholt-Jepsen, Lars V. Kessing, and Jakob E. Bardram. 2020. *MUBS: A Personalized Recommender System for Behavioral Activation in Mental Health.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376879

[39] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *Proceedings*

*of the Sixteenth ACM Conference on Economics and Computation* (Portland, Oregon, USA) *(EC '15)*. Association for Computing Machinery, New York, NY, USA, 453–470. https://doi.org/10.1145/2764468.2764488

[40] Brent Smith and Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing* 21 (05 2017), 12–18. https://doi.org/10.1109/MIC.2017.72

[41] Ambuj Tewari and Susan A. Murphy. 2017. From Ads to Interventions: Contextual Bandits in Mobile Health. In *Mobile Health - Sensors, Analytic Methods, and Applications*.

[42] Lee Valentine, Simon D'Alfonso, and Reeva Lederman. 2022. Recommender systems for mental health apps: advantages and ethical challenges. *AI & SOCIETY* (2022). https://doi.org/10.1007/s00146-021-01322-w

[43] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR Mhealth Uhealth* 4, 3 (21 Sep 2016), e111. https://doi.org/10.2196/mhealth.5960

[44] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 165–174. https://doi.org/10.1145/3331184.3331267

[45] Jiancan Wu, Xiangnan He, Xiang Wang, Qifan Wang, Weijian Chen, Jianxun Lian, and Xing Xie. 2020. Graph Convolution Machine for Context-aware Recommender System. *arXiv preprint arXiv:2001.11402* (2020).

[46] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.

[47] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14 (2020), 1–101.