# Data-Driven Implications for Translating Evidence-Based Psychotherapies into Technology-Delivered Interventions

Jessica Schroeder
University of Washington
jesscs@cs.washington.edu

Jina Suh
University of Washington,
Microsoft Research
jinasuh@cs.washington.edu

Chelsey Wilks
Harvard University
chelseywilks@fas.harvard.edu

Mary Czerwinski
Microsoft Research
marycz@microsoft.com

Sean A. Munson
University of Washington
smunson@uw.edu

James Fogarty
University of Washington
jfogarty@cs.washington.edu

Tim Althoff
University of Washington
althoff@cs.washington.edu

## ABSTRACT

Mobile mental health interventions have the potential to reduce barriers and increase engagement in psychotherapy. However, most current tools fail to meet evidence-based principles. In this paper, we describe data-driven design implications for translating evidence-based interventions into mobile apps. To develop these design implications, we analyzed data from a month-long field study of an app designed to support dialectical behavioral therapy, a psychotherapy that aims to teach concrete coping skills to help people better manage their mental health. We investigated whether particular skills are more or less effective in reducing distress or emotional intensity. We also characterized how an individual's disorders, characteristics, and preferences may correlate with skill effectiveness, as well as how skill-level improvements correlate with study-wide changes in depressive symptoms. We then developed a model to predict skill effectiveness. Based on our findings, we present design implications that emphasize the importance of considering different environmental, emotional, and personal contexts. Finally, we discuss promising future opportunities for mobile apps to better support evidence-based psychotherapies, including using machine learning algorithms to develop personalized and context-aware skill recommendations.

## CCS CONCEPTS

• **Human-Centered Computing** → **Human Computer Interaction**.

## KEYWORDS

Mobile Health Interventions, Mental Health, Data Science, Dialectical Behavioral Therapy


*PervasiveHealth '20, May 18–20, 2020, Atlanta, GA, USA*

ACM ISBN 978-1-4503-7532-0/20/05…$15.00
https://doi.org/10.1145/3421937.3421975

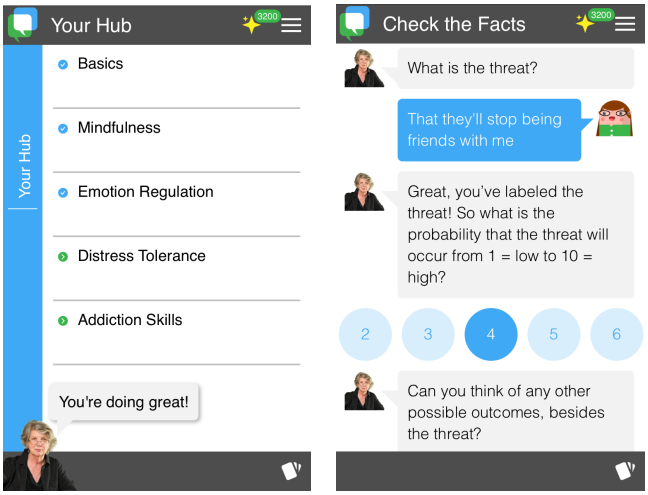
**ACM Reference Format:**
Jessica Schroeder, Jina Suh, Chelsey Wilks, Mary Czerwinski, Sean A. Munson, James Fogarty, and Tim Althoff. 2020. Data-Driven Implications for Translating Evidence-Based Psychotherapies into Technology-Delivered Interventions. In *14th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '20), May 18–20, 2020, Atlanta, GA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3421937.3421975

## 1 INTRODUCTION

Mental health disorders are a leading cause of disability and death worldwide, with approximately 18% of US adults experiencing a mental illness in a given year [39]. Evidence-based psychotherapy is often effective for treating mental health conditions [56], as it can help people develop positive coping mechanisms to better manage their conditions. Unfortunately, people generally struggle to engage with psychotherapy, preventing them from managing their mental health disorders [40]. Increasing engagement with psychotherapy is therefore a top priority [4, 11, 12, 29, 30].

Technology-delivered mental health interventions have the potential to reduce barriers and increase engagement in psychotherapy, as they can reduce the financial and time burdens associated with attending in-person therapies [13] and increase comfort in providing honest disclosure [14, 15, 27]. Unfortunately, despite high interest in technology-delivered mental health interventions [47], commercially-available mental health apps often fall short of expectations; only 10% of apps aimed at supporting therapies for depression meet evidence-based principles [19], and many people view current digital psychotherapies as ineffective [44]. These challenges suggest a need for improved understanding of how to effectively translate evidence-based psychotherapies into technology-delivered interventions.

In this paper, we present data-driven design implications for such translations based on our examination of how people use Pocket Skills (Figure 1), a mobile web app designed to provide holistic support for Dialectical Behavioral Therapy (DBT). DBT is a skills-based therapy designed to support people with complex, difficult-to-treat disorders in developing concrete skills to help them solve problems, maintain positive relationships, and navigate negative events and emotions [24, 25]. Although research has shown that DBT can help people successfully manage a wide range of disorders [10, 26, 36], it has generally been difficult or impossible to quantitatively analyze the effectiveness of *particular*




**(a) Pocket Skills includes four modules, each focusing on different types of skills.**

**(b) Each skill walks people through DBT content via a conversational interface.**

**Figure 1: Pocket Skills helps people learn and practice skills.**

*skills* in *real-world contexts*. The translation of evidence-based interventions into mobile apps—and the collection of usage data within those apps—provides new opportunities to conduct such analyses, informing designs that help ensure people receive the best possible support. Similarly, comparing the effectiveness of different skills for different subgroups of people (e.g., those with particular disorders, demographics, or other characteristics) could reveal individuals that may need additional support.

In this work, we therefore examined the effectiveness of individual DBT skills when translated into a mobile app, both overall and for different subgroups of people. The original month-long field study of Pocket Skills [51] collected: 1) survey data, including validated scales to assess progress, and 2) app usage data, including participant-reported skill effectiveness in terms of increasing mindfulness and reducing emotional distress. We analyzed this survey and usage data to investigate the following questions:

RQ1: When did study participants use the skills?

RQ2: Were particular skills more or less effective (e.g., in reducing emotional intensity and distress, in fostering mindfulness)?

RQ3: Were skills more or less effective for different subgroups of people (e.g., those with or without certain conditions)?

RQ4: Did skill-level effectiveness influence overall depression, anxiety, or skill use improvement throughout the study?

RQ5: Can we *predict* a particular skill's effectiveness, given participant and skill characteristics?

We found that skills designed to help people regulate their emotions tended to be particularly helpful, while skills designed to help people manage distress were less effective (Section 4.1 and Section 4.2). We also found skill effectiveness differed among subgroups of people (Section 4.3), and that participants who reported higher skill effectiveness tended to report more improvement in depressive symptoms after the study (Section 4.4). We then developed a model to predict the effectiveness of specific skills (Section 4.5). Based on these findings, we developed data-driven design implications for translating evidence-based therapies into mobile applications (Section 5.1). Our design implications emphasize the importance of considering environmental contexts, emotional contexts, and personal contexts. Finally, we discuss important areas for future work, including opportunities for personalized and context-aware skill suggestions (Section 5.2).

## 2 BACKGROUND

In this section, we provide additional background on Dialectical Behavioral Therapy (DBT). We also describe prior work investigating the use of technology to monitor and predict mental health symptoms and to support positive coping skill use.

### 2.1 Dialectical Behavioral Therapy (DBT)

Dialectical Behavioral Therapy (DBT) was developed to treat complex behaviors associated with high emotional dysregulation [24]. DBT was originally designed as a treatment for borderline personality disorder (BPD) [26], but it has also been successfully applied to people with addictive behavior, eating disorders, and mood disorders [10]. The use of DBT skills as a whole has been shown to improve suicidal and self-injurious behavior, expressions of anger, and interpersonal problems [36].

DBT skills are traditionally separated into ***modules***. Each module helps people learn specific types of ***skills***. For example, the *Mindfulness* module contains skills dedicated to teaching people to accept the moment without judgment (e.g., by observing their breathing, by describing their thoughts). The *Emotion Regulation* module is designed to help people understand, manage, and adjust their emotional states. The *Distress Tolerance* module gives people specific healthy alternatives to unhealthy behaviors (e.g., instead of self-harming, people can take a cold shower or gently snap a rubber band on their wrist). By working through the modules and learning specific skills that can replace unhealthy behavior, people can start to apply those skills in their lives when they need them, helping them better navigate negative events and emotions as part of managing their mental illnesses [36].

### 2.2 Technological Support of Mental Health

Researchers have investigated technological support of mental health from a variety of perspectives. For example, research has investigated monitoring and predicting mental health symptoms through mobile phone sensors (e.g., [5, 57]) and social media posts (e.g., [2, 8, 9, 28, 58]). A large body of work has examined how machine learning techniques can support detection, diagnosis, and treatment of a myriad of mental health conditions (e.g., [52]). Researchers have also investigated personalizing mental health interventions (e.g., by recommending activities to manage stress [48] or prevent negative moods [18] based on an individual's past sleep, diet, and activity data; by recommending interventions based on an individual's personal characteristics and context [41]). Other work has examined evaluating counseling sessions through natural language processing and machine learning approaches to differentiate high-quality and low-quality counselors or counseling

sessions (e.g, [1, 43, 54]). Similar to this prior work, we aim to quantitatively assess whether particular DBT skills may be particularly effective for different individuals.

Prior work has also examined how technology could support skills practice for a range of skill-based psychotherapies. For example, a suite of skills-based apps has been shown to reduce depression and anxiety [32]. The DBT Coach is designed to provide constantly-available, interactive walkthroughs of DBT skills [45, 46]. Similarly, the Virtual Hope Box includes analogous skills to the mindfulness and distress tolerance skills found in DBT, emphasizing support, comfort, distraction, and relaxation [7]. Participants in the Pocket Skills feasibility study reported improvements in depression, anxiety, and DBT skills use, describing increased engagement with DBT that helped them learn skills and ultimately apply them to their daily lives [51]. The concept of suggesting specific interventions when people need them, known as "just in time interventions" [34], has been pursued in other health contexts (e.g., promoting physical activity [17], stress management [20], and weight management [53]), with recent interest in applying them to positive coping skill use [21]. We build on this prior work and advance this opportunity by quantitatively determining *which* skills may be more or less effective, and whether skill effectiveness varies for different subgroups of people.

## 3 DATASET

To examine our research questions, we reanalyzed data from a previous field deployment of Pocket Skills [51]. These data required additional processing and have associated limitations.

### 3.1 Original Data Collection

The Pocket Skills deployment consisted of a 4-week field study [51]. 100 people were recruited; 27 dropped out over the course of the study, resulting in 73 total participants. The app included modules for *Mindfulness*, *Emotion Regulation*, *Distress Tolerance*, and *Addiction Skills* adapted from the DBT Skills Training Manual and Workbooks [25]. Each *module* contained module-specific *skills* presented via a conversational interface (see Figure 1). Some of these *skills* contained *subskills*, or different options for completing the skill (e.g., the *Mindfulness* skill of *Observing* included subskills for observing *breathing*, *sounds*, *visuals*, and *everyday life*).

Throughout the study, Schroeder et al. collected data on participant characteristics via surveys, including demographic information (e.g., gender, age); their anxiety, depression, and coping skill use (measured with the OASIS [38], PHQ-9 [23], and DBT Ways of Coping Checklist [37]); and other characteristics (e.g., what disorders they have been diagnosed with, whether they take medication, what modules they preferred). They also collected app usage data throughout the study, including app navigation and participant inputs for the skills. Many skills included Likert-scale ratings of how a participant felt before and/or after completing the skill, which we refer to in this paper as *pre- and post-ratings*. For example, *Mindfulness* skills often asked people to rate how mindful they felt after completing the skill; *Emotion Regulation* skills often asked people to rate their emotional intensity before and after completing the skill; and *Distress Tolerance* skills asked people to rate their level of distress before and/or after completing the skill.

### 3.2 Data Processing

To process the Pocket Skills study data, we first reviewed the app content to examine the skills themselves. We identified 11 skills that had pre- and post-ratings and 26 skills that had only post-ratings: 18 distinct *Mindfulness* skills (all with post-ratings only); 5 distinct *Emotion Regulation* skills (all with post- and pre-ratings); and 14 distinct *Distress Tolerance* skills (9 with post-ratings only and 5 with post- and pre-ratings). We excluded from our analyses any skill that did not include any Likert-scale ratings. No participant completed the *Emotion Regulation* skill *Cope Ahead*; all other skills were practiced at least once by at least one participant. No *Addiction* skills had Likert-scale ratings, so the module was excluded. Supplementary Table 1 shows our final list of *modules*, *skills*, and *subskills*. Four participants did not complete any of the included skills over the course of the study, and were therefore excluded from the dataset.

After we identified the skills to include, we extracted each distinct use of a specific skill from the usage log data and derived metadata (e.g., the ratings, the first use of a skill, the total order of practiced skills). Each interaction with the app (e.g., enter/exit skill screen, rate mindfulness) was logged with a local timestamp from each participant's device, from which we derived temporal metadata (e.g., skill rating time of day, day of week). Because participants were able to access the app even after the study was completed, we filtered the data to only include skills practiced between the study intake (July 19th, 2017) and the last submitted exit survey (August 23rd, 2017). We then standardized the ratings: some had Likert-scales from 1-10, some from 1-5, and some binary (yes/no). *Mindfulness* scores were better if higher (i.e., a higher score indicated more mindfulness), but lower scores were better for all other modules (e.g., lower *Distress Tolerance* scores indicated lower distress). We therefore shifted everything to a 5-point scale, and reversed *Mindfulness* ratings so lower ratings were better (i.e., higher mindfulness). Finally, we computed the difference between pre- and post-ratings to characterize improvement before and after completing a skill, referred throughout the paper as *skill improvement*. Because lower numbers indicate better ratings, negative numbers indicate more improvement.

We grouped participants within categories for analysis based on their survey responses (see Table 1 for the categories and distributions). Categories included education; age; gender; intake survey scores on the PHQ-9 and OASIS; the number of family members living close (within a 50-mile radius); whether they were on any mental health related medication; and any mental health disorders they were diagnosed with. For age, close family, and education, we defined buckets to better balance the groups. Because participants had been diagnosed with a wide range of disorders, we grouped disorders by category based on the Diagnostic and Statistical Manual of Mental Disorders (DSM) [3]. Categories included neurodevelopmental disorders, mood disorders, anxiety disorders, personality disorders, eating disorders, and no diagnosed disorder. Each participant was then categorized as having or not having each disorder type, based on their reported disorders. Diagnosed disorders were all self-reported; the original study did not include clinical diagnostic interviews.

### 3.3 Limitations

Our dataset has a several associated limitations. First, we found evidence of incomplete data logging: one datapoint had a post-skill rating without a pre-skill rating, which should not have been possible given the app design. We excluded that data point, but its existence proves that our skill practice data is not complete. Second, we cannot say how many skills people practiced or applied in their daily lives without explicitly using the app. Participants reported that having the app reminded them to use skills in the moment, sometimes without the app; although the app included self-tracking of moods and behaviors, it did not allow users to log skills practiced outside of the app. Third, for many skills, participants may not have actually completed the skill at the time it was suggested. For example, some *Distress Tolerance* skills suggest activities such as taking a walk or a cold shower; the app did not support indication of whether or when participants actually completed those activities. Future tools designed to support DBT could include ways to report these aspects of DBT skill practice, which would allow further research and development of intelligent support for skill practice.

The study methods also introduced some biases. For ethical reasons, all participants were enrolled in therapy at the time of the study; we therefore cannot characterize what positive effects may be due to their therapy, rather than the app itself, beyond self-report by the participants. We also can only comment on the usefulness of the skills as they were translated in the app, which may be different than the usefulness of skills more traditionally taught and practiced. The majority of the study participants were female, possibly because women are more likely to be diagnosed with borderline personality disorder (BPD) (although prevalence is thought to be approximately equal [49]) and are more likely to seek therapy for BPD [16] and in general [55]. The role of gender on skill effectiveness should therefore be examined in future studies.

Finally, the clinical psychologists on the team ordered the modules based on traditional DBT practices, but that ordering may have encouraged more use of skills with only post-ratings. Additionally, at the beginning of the study, participants could not access the *Emotion Regulation* module until they had gone through the entire *Mindfulness* module. Based on participant feedback, all modules were unlocked after the second week; however, this initial limitation encouraged more use of *Mindfulness* skills, which lacked the pre-ratings needed to observe skill improvements. The *Distress Tolerance: Self-Soothe* skills were also added halfway through the study, limiting the amount of time participants were able to use them. Future studies investigating relative skill effectiveness should consider designs that collect pre- and post-ratings for every skill and should examine possible ordering effects.

## 4 METHODS AND RESULTS

In this section, we describe our methods and results from the statistical analyses and modeling we conducted to answer our research questions. This paper focuses on our quantitative analyses of skill usage and participant characteristics; a previous publication presents qualitative and quantitative descriptions of participant experiences while using Pocket Skills [51].

| | |
|---|---|
| Gender | 62 female, 6 male,<br>1 Genderqueer/androgynous |
| Age | 18-63 ($\bar{X}$=37.3) |
| Age Buckets | <25 (12), 25-35 (17), 35-45 (21), >45 (19) |
| Education | High school (7), Some college (16),<br>2-year degree (5), 4-year degree (21),<br>Professional degree (2),<br>Master's degree (14), Doctorate (4) |
| Education Buckets | <4 year degree (28), 4 year degree (21),<br>>4 year degree (20) |
| Depression at Intake (Via PHQ-9) | Minimal or None (6), Mild (17),<br>Moderate (18), Moderately Severe (16),<br>Severe (12) |
| Anxiety at Intake (Via OASIS) | No anxiety disorder (12),<br>Anxiety disorder (57) |
| On Related Medication | 12 no, 54 yes, 3 declined to answer |
| Types of Disorders | Mood (42), Anxiety (43), Personality (27),<br>Neurodevelopmental (8), Eating (3) |
| Close Family Member Buckets | 0 (13), 1-2 (24), 3-4 (20), >4 (12) |

**Table 1: Participant categories. Many categories were biased towards particular subgroups, necessitating the bucketing.**

We address RQ1–RQ4 by examining the data along different individual dimensions through stratification and statistical tests. In Section 4.1 and Section 4.2, we describe overall skill use throughout the study, as well as skill effectiveness (e.g., how much participant emotional intensity or distress improved after using the skill, how mindful participants felt after using the skill) both in general and between different subgroups of participants. We examine how these participant and skill-level characteristics correlated with mental health improvements over the course of the study (i.e., improvements in anxiety, depression, and skill use) in Section 4.4. We used Python to clean and process the data and calculate group averages, then used R for our statistical analyses.

Finally, we address RQ5 by consolidating our findings and explorations in a machine-learning (ML) model that jointly uses the variables and leverages interactions between variables. In Section 4.5, we developed a model to predict skill improvement given participant and skill characteristics and to demonstrate the incremental value of additional information. We used scikit-learn, a Python machine learning package [42], to develop these models.

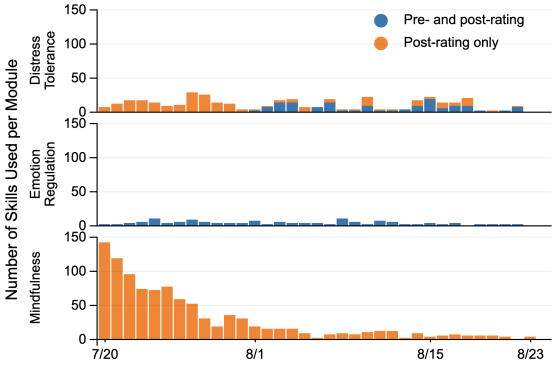



**Figure 2: Distribution of the total number of skills practiced over the course of the Pocket Skills feasibility study, showing skills with pre- and post-ratings and skills with only post-ratings.**

## 4.1 RQ1: Skill Use

We first examined overall skill use throughout the study.

*4.1.1 Methods.* To investigate overall skill use, we visualized the data and computed basic sums and averages.

*4.1.2 Results.* We found 1483 total skills practiced by 69 participants: 974 *Mindfulness* skills (all with post-only ratings); 127 *Emotion Regulation* skills (all with pre- and post-ratings); and 382 *Distress Tolerance* skills (131 with pre- and post-ratings and 258 with post-ratings only). *Mindfulness* skills were most often used at the beginning of the study, with *Emotion Regulation* and *Distress Tolerance* skills more evenly distributed throughout (see Figure 2). Between July and August, *Mindfulness* skill use went from 78% of total usage to 37%, while *Distress Tolerance* went from 16% to 47%, perhaps because *Distress Tolerance: Self-Soothe* skills were added on August 1st. In the same time period, *Emotion Regulation* skill usage increased from 5% to 16%, perhaps because the *Emotion Regulation* module was universally unlocked (i.e., no longer depended on the completion of the *Mindfulness* module) on the same date. Usage patterns may also be partially due to an ordering effect (i.e., because the *Mindfulness* module was presented first in the app). Total usage of the skills dropped by more than half between July and August, indicating that skills were used less overall as the study progressed.

Each participant completed an average of 21.5 skills throughout the study (min=1, max=93, stdev=20.57). On average, participants completed 14.1 *Mindfulness* skills (min=0, max=70, stdev=15.6), 1.84 *Emotion Regulation* skills (min=0, max=10, stdev=2.15), and 5.54 *Distress Tolerance* skills (min=0, max=42, stdev=7.00). Of the 69 participants, 51 practiced at least one skill more than once. Of the skills practiced with pre- and post-ratings, 213 (82.6%) had pre-ratings of $\geq 3$ (i.e., indicated high levels of emotional intensity or distress before completing the skill).

## 4.2 RQ2: Overall Skill Effectiveness

We next compared overall skill effectiveness for the different modules and skills across all participants.

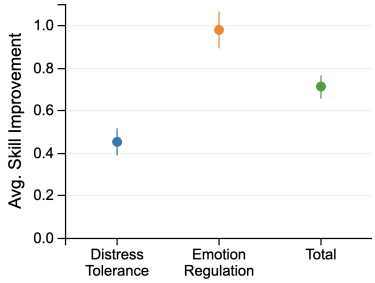



**Figure 3: Average skill improvement on the 5-point scales of all skill uses (purple), *Emotion Regulation* skills (blue), and *Distress Tolerance* skills (red), with standard error bars. People improved more using *Emotion Regulation* skills than *Distress Tolerance* skills.**

*4.2.1 Methods.* After standardizing skill ratings (see Section 3.2), our data included skills with only pre-ratings, skills with only post-ratings, and skills with both pre- and post-ratings. To investigate improvement trends, we examined skills with both pre- and post-ratings; similarly, to investigate post-rating trends, we examined skills with only post-ratings. For overall skill use, we examined skills at both the *module* and *subskill* granularities (see Supplementary Table 1). We used t-tests to analyze effectiveness by module: one to examine differences in skill improvement between the *Emotion Regulation* and *Distress Tolerance* skills (the only modules with skills that had pre- and post-ratings) and another to examine differences in post-ratings between *Distress Tolerance* and *Mindfulness* (the only modules with skills that had only post-ratings).

For our *subskill*-level analyses, we used one-way analysis of variance (ANOVA) tests to examine differences in 1) skill improvement and 2) post-ratings by *subskill*. When we found significant results, we then investigated pairwise differences, employing Tukey's HSD procedure to correct for the increased risk of Type I error due to unplanned comparisons.

*4.2.2 Results.* For skills with pre- and post-ratings, we found a significant effect of *module* on skill improvement ($t(233.16) = 4.956$, $p<0.001$). *Emotion Regulation* skills correlated with more than half a point more improvement on our 5-point scale ($\bar{X}$=-0.98) than *Distress Tolerance* skills ($\bar{X}$=-0.45). Figure 3 illustrates the overall average skill improvements across different modules. We also found a significant effect of the specific *subskills* ($F(9, 248) = 3.901$, $p<0.001$). Our Tukey HSD test accounting for unplanned comparisons revealed *Emotion Regulation* skills, particularly the *Problem Solve* skill, generally correlated with significantly greater improvement than *Distress Tolerance* skills of *Self-Soothe* (see Supplementary Section S.2).

For skills with only post-ratings, we found no significant difference between *modules*. We did find a significant effect of the *subskills* ($F(25, 1199) = 4.461$, $p<0.001$). Our Tukey HSD test accounting for unplanned comparisons revealed 14 significant pairwise differences (see Supplementary Table 2). Generally, *Mindfulness* skills of *Observing*, *Describing*, and *Participating* correlated with better post-ratings than *Mindfulness* skills of *Non-Judgementality*. *Distress Tolerance* skills of *Distracting* also correlated with better post-ratings than *Mindfulness* skills of *Non-Judgementality*.

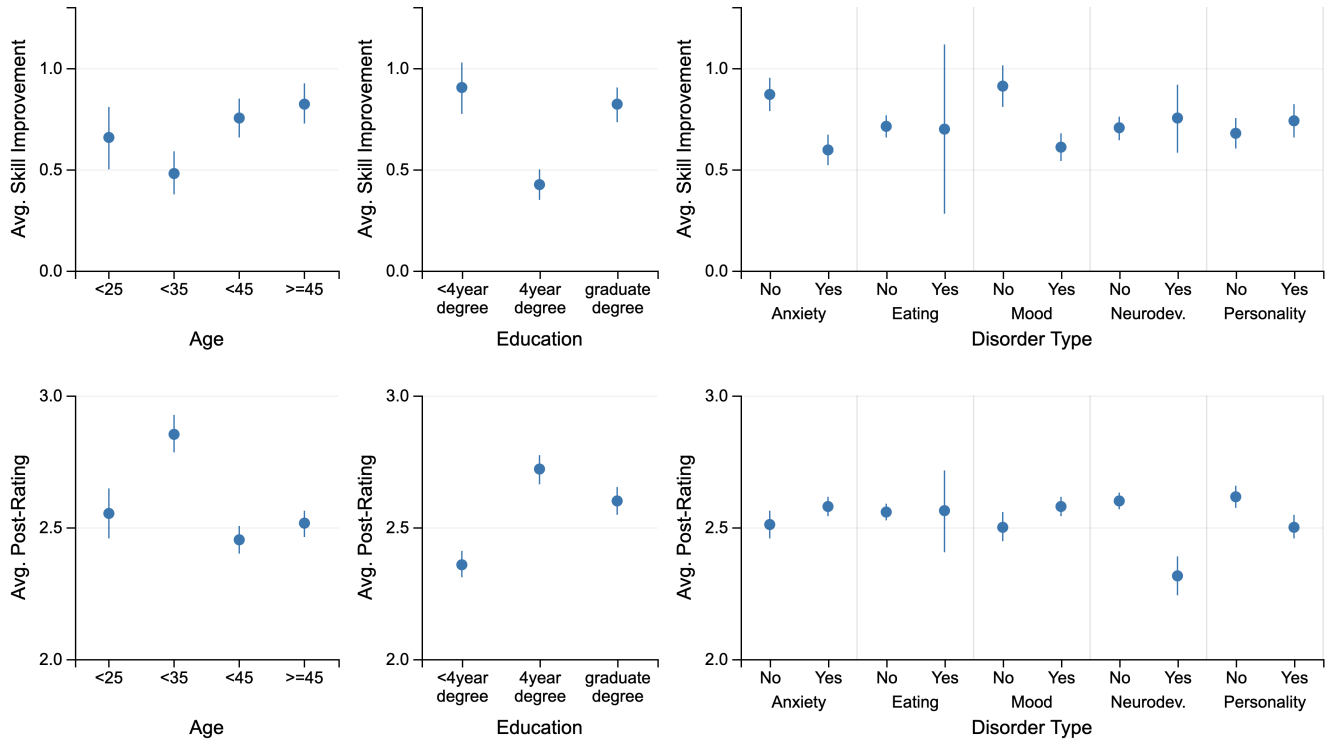



**Figure 4: Average skill improvement (top) and post-ratings (bottom) across age subgroups (left), education subgroups (middle), and disorder types (right), with standard error bars. *Higher* improvement indicates more improvement (i.e., is better), while *lower* post-ratings indicate more positive ratings (i.e., is better). Some subgroups varied more, on average, than others. Section 4.3.2 discusses significant differences between subgroups.**

## 4.3 RQ3: Skill Effectiveness Across Subgroups

After examining overall skill effectiveness, we investigated whether different subgroups of people reported different skill effectiveness.

*4.3.1 Methods.* We first separated participants into the subgroups described in Section 3.2 and Table 1. We ran t-tests on categories with two groups (i.e., whether or not the participant completing the skill: took medications to manage their mental health; had anxiety disorder at the intake survey based on their OASIS results; had a mood, anxiety, eating, personality, neurodevelopmental, or no disorder diagnosed) to investigate differences between subgroups for each type of skill effectiveness (i.e., skill improvement, post-ratings). For each type of skill effectiveness, we then used the Benjamini-Hochberg procedure [6] on the t-test results to correct for multiple comparisons. Our results report these adjusted p-values.

For groups with three or more subgroups (i.e., initial PHQ-9 class; gender; close family, age, and education buckets), we ran one-way ANOVA tests to investigate subgroup differences in each type of skill effectiveness measure. When we found significant results, we again investigated pairwise differences using Tukey's HSD procedure to account for unplanned comparisons.

*4.3.2 Results.* Figure 4 illustrates the differences in average skill improvements and post-ratings across various participant subgroups. For skill improvement, *Education* had a significant effect ($F(2, 255)=7.068$, $p<0.01$), with having <4 year degree correlating with 0.48 points more improvement than having a 4-year degree ($t=3.404$, $p<0.01$) and having a graduate degree correlating with 0.4 points more improvement than having a 4-year degree ($t=-3.081$, $p<0.01$). We found no significant differences in skill improvements between subgroups in any other categories.

For skill post-ratings, we found a significant difference depending on whether people had a *personality disorder* ($t(1220.7)=2.533$, $p<0.05$): people who had a *personality disorder* rated skills slightly more positively ($\bar{X}=2.40$) than those who did not ($\bar{X}=2.56$). We similarly found a significant effect of *neurodevelopmental disorder* ($t(252.68)=2.580$, $p<0.05$), with people who had a neurodevelopmental disorder rating skills slightly more positively ($\bar{X}=2.28$) than those who did not ($\bar{X}=2.52$). We also found a significant difference in post-rating given *medication use* ($t(214.39)=3.1043$, $p<0.05$), with people who took medication rating skills slightly more positively ($\bar{X}=2.43$) than people who did not take medication ($\bar{X}=2.75$). *Education* also significantly correlated with post-rating ($F(2, 1222)=6.27$, $p<0.01$), where a <4 year degree correlated with slightly better post-ratings than having a 4-year degree (by 0.28 points; $t=3.384$, $p<0.01$) or a graduate degree (by 0.20 points; $t=2.561$, $p<0.05$). *Number of family members close* also had a significant effect ($F(3, 1221)=3.36$, $p<0.05$), with having no family close correlating with less positive ratings than having 1-2 (by 0.38 points; $t=-2.90$, $p<0.05$) or 3-4 (by 0.42 points; $t=-3.10$, $p<0.05$). Finally, we found a significant effect of *age* ($F(3, 1221)=4.19$, $p<0.01$); being 25-35 years old correlated with more negative ratings than either being between

35-45 years old (by 0.34 points; $t$=-3.502, $p$<0.01) or being greater than 45 years old (by 0.25 points; $t$=-2.651, $p$<0.05). We found no significant differences in post-ratings between any other subgroups.

## 4.4 RQ4: Skill vs. Validated Scale Improvement

We next investigated whether participant and skill characteristics correlated with overall improvements in the clinically validated scales that participants in the feasibility study [51] completed for anxiety (OASIS [38]), depression (PHQ-9 [23]), and skill use (DBT Ways of Coping Checklist [37]) throughout the study.

*4.4.1 Methods.* We first calculated score differences between the intake and exit surveys for the PHQ-9, OASIS, and DBT Ways of Coping Checklist scales. We then extracted additional participant characteristics, including: their favorite and least favorite module; how much they felt Pocket Skills helped their goals and skill use; and whether they felt they practiced more skills with Pocket Skills than they would have practiced without it. We also examined participant skill use patterns, including total skills practiced in each module; whether they repeated any skills; their average, best, and worse skill improvement and post-rating; and the proportion of unique skills they practiced (i.e., an individual's number of unique skills over the number of total skills practiced).

We performed mixed model analyses of variance for each scale, treating the characteristics described above as well as medication use and demographics (i.e., age, education, and close family member buckets) as fixed effects and the *specific disorder types* (i.e., whether they reported any disorders within each disorder category) as random effects to account for any heterogeneity within the overarching disorder types. We again investigated pairwise differences using Tukey's HSD procedure to adjust for repeated testing.

*4.4.2 Results.* For depression improvement, we found significant main effects of *age bucket* ($F$(3, 9.3811)=4.5852, $p$<0.05); *family bucket* ($F$(3, 9.6203)=4.3922, $p$<0.05); *education bucket* ($F$(2, 9.5581)=5.0718, $p$<0.05); *favorite module* ($F$(3, 9.1457)=4.5373, $p$<0.05); *best skill improvement* ($F$(1, 9.3973)=5.3843, $p$<0.05); and *best skill post-rating* ($F$(1, 9.6632)=8.5388, $p$<0.05). Participants with larger best skill improvements and more positive best post-skill ratings tended to improve more on their PHQ-9 score. Pairwise analyses revealed that being 35 or younger generally correlated with more improvement than being older than 35, and that having zero family members close correlated with more improvement than having any number greater than zero. Having a <4 year degree correlated with more improvement than having a graduate degree ($z$=-3.161, $p$<0.01). Finally, people who preferred the *Addiction Skills* module improved more than those who preferred any other module. See Supplementary Section S.3 for Tukey test details.

For anxiety improvement, we found a significant main effect of *age bucket*, with participants who were older than 35 again improving less than those who were 25-35 (see Supplementary Section S.3). The skill use model yielded no significant results.

## 4.5 RQ5: Predictability of Skill Effectiveness

Finally, we examined the feasibility of predicting skill effectiveness for particular participants and skills. Given participant and skill use characteristics and each participant's historical skill usage and rating data, we built four different machine learning classifiers to predict whether a specific skill would result in a skill improvement for that participant (i.e., lower emotional intensity or distress after skill use).

*4.5.1 Methods.* To predict whether using a skill would lead to a skill-level improvement for a participant, we used the subset of data (N=258) that included pre- and post-ratings to train binary classification models with positive prediction corresponding to skill improvement. We used participant and skill use characteristics as features, including contextual information (e.g., time of day, day of week, emotional intensity or distress rating prior to skill use); usage patterns (e.g., consecutive use of a skill within 15 minutes); historical skill improvement (e.g., cumulative average of skill improvements, skill improvement from the last skill use); and preferences (e.g., favorite module). All categorical features were one-hot encoded to transform each category into a binary representation necessary for model construction.

We trained binary classifiers using four different learning algorithms (i.e., decision tree, gradient boosting, random forest, logistic regression). We trained and evaluated each model using leave-one-participant-out cross validation. We used the average model accuracy and AUC on the hold-out participants as our metric to prevent overfitting and tune hyper-parameters of the learning algorithm (e.g., depth of the tree, max feature count, number of estimators, minimum samples for splitting nodes and leaf nodes, regularization strength, penalty). Because results from both test accuracy and test AUC were qualitatively similar, we report only test accuracy (see Supplementary Section S.4 for AUC results).

Finally, we performed a feature ablation study by training independent classifiers using all except one feature in order to understand the informativeness of each feature. We specifically focused on the set of features or variables that we found to have significant effect on the skill improvement in our statistical analyses (see Supplementary Section S.4).

*4.5.2 Results.* Given all features, the decision tree classifier yielded the best test accuracy (72.1%). All classifiers performed better than the base rate of the majority class (57.4%, or 148 out of 258 skill uses, were reported to have improved ratings). Our feature ablation study revealed that *skill ID* was the most impactful feature for the decision tree, gradient boosting, and logistic regression classifiers, leading to a drop of 17.1%, 13.4%, and 4.9% respectively in accuracy when removed. On the other hand, *mood disorder* was the most impactful feature for the random forest classifier, leading to a drop of 1.1% in accuracy when removed.

We further found that different classifiers perform better or worse on individual skills. For example, Figure 5 shows that the accuracy ranges from 20% for a random forest classifier to 80% for gradient boosting and decision tree classifiers for the *Brief Check the Facts* skill. Similarly, accuracy varied dramatically within specific classifiers for different skills. For example, for the logistic regression classifier, the accuracy for *Distress Tolerance* skills (54.2%) is lower than the accuracy for *Emotion Regulation* skills (68.8%).

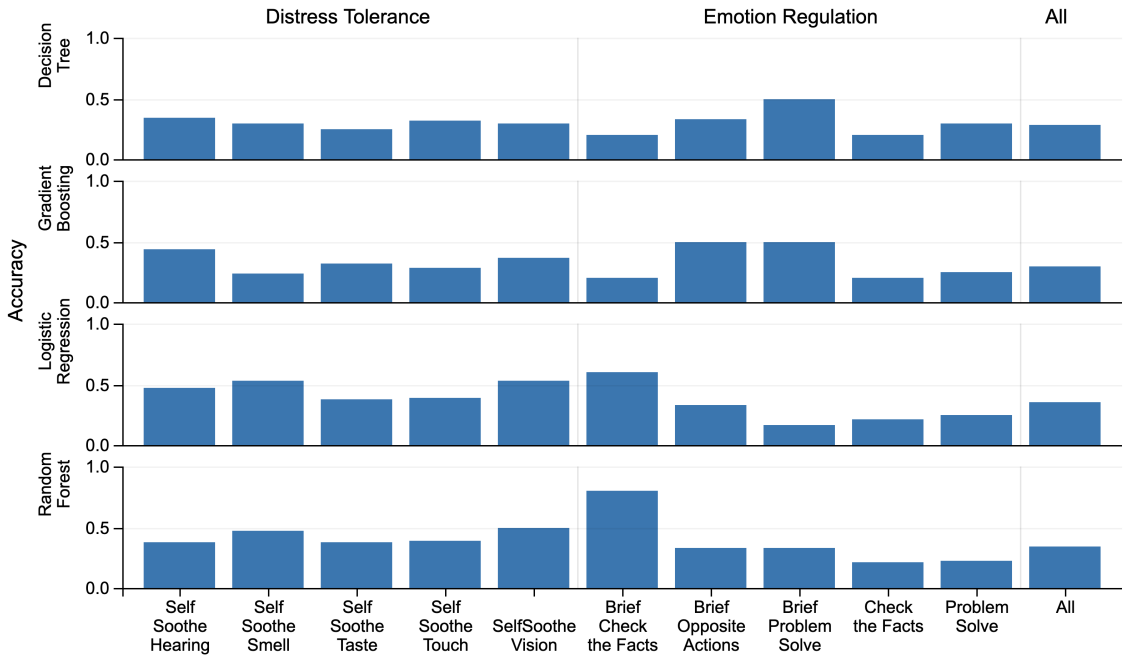



**Figure 5: Accuracy of model prediction across individual skills, with accuracy of a single model across all skills displayed on the right. The varying accuracy of different models for different skills, together with an ablation study highlighting the importance of *skill ID*, reveals a need for skill-specific models that can account for different context relevant to each skill.**

## 5 DISCUSSION

In this paper, we analyzed data from a month-long field study of Pocket Skills, a mobile web application designed to provide holistic support for dialectical behavioral therapy (DBT). Based on our findings, we present design implications for future translations of evidence-based therapies into mobile apps, including the importance of designing for environmental context, emotional context, and personal context. We also discuss the promise for future work in personalized and context-aware recommendations for skill-based mental health interventions. These design implications and opportunities can inform future designs for the increasingly-prevalent mobile app interventions aimed to help people better manage their mental health.

### 5.1 Data-Driven Design Implications

Our findings highlight a range of contextual information that can influence skill effectiveness for an individual using a mobile app to identify and implement a positive coping skill.

*5.1.1 Designing for Environmental Context.* Overall, we found that participants consistently reported specific skills to be more effective than others. In Section 4.2, we described how participants generally improved more using *Emotion Regulation* skills and less using *Distress Tolerance: Self-Soothe* skills. This difference may be due to the immediate feasibility of the activities suggested in the respective skills. *Emotion Regulation* skills generally walk people through considering and solving the problem they have, and can therefore be completed in most contexts. In contrast, the included *Distress Tolerance: Self-Soothe* skills suggest specific methods to distract people from distress that may be difficult to complete at certain times (e.g., eating something spicy, taking a cold shower). These activities are helpful when one has the ability to complete them, and are therefore useful in certain environmental contexts. However, they often cannot be immediately performed (e.g., when the person using the app is not at home). The post-ratings we collected therefore may not always reflect how people felt after completing the activity, as they may have been unable to do so when the post-ratings were solicited. Skills that suggest activities that may be infeasible in certain contexts should therefore be designed differently than those that can be done anywhere at anytime, perhaps by asking the person if they are able to immediately complete the skill and enabling them to explicitly schedule the skill for a future time if they cannot.

*5.1.2 Designing for Emotional Context.* In addition to enabling people to indicate later uses of a skill, apps designed to support skill-based interventions should recognize emotional circumstances in which it may be inappropriate to suggest skills that cannot be immediately completed. As we reported in Section 4.2, the majority of people using skills with pre- and post-ratings reported high distress or emotional intensity in the pre-rating of the skill: they were using the app *in the moment*, rather than practicing the skill so they could use it during *future* times of distress. Although all skills were directly adapted from the DBT Skills Training and Workbooks [51], Pocket Skills is a constantly-available resource and may therefore be used much differently than a traditional DBT skills worksheet. Instead of directly translating content, designers of

tools to support skill-based interventions must consider how to best adapt skill-related materials to consider these different use cases. For example, a tool could differentiate between people practicing skills versus people who are currently in distress and attempting to use a skill *in the moment*. When a person indicates they are in distress, such a tool could focus on guiding them towards skills that are most likely to be feasible and useful in their current context. A tool could even *sense* emotional state (e.g., using sensing techniques such as those described in [22, 33]) and send push notifications during times of distress to directly suggest an appropriate skill. Such just-in-time interventions have been shown to support health behavior change (e.g., in promoting physical activity [17], in stress management [20], in weight management [53], in smoking cessation [35]). Our results both support emerging interest in applying such techniques to positive coping skill use (e.g., [21]) and differentiate the additional need to support skill development when people are *not* distressed. For example, a tool could then focus on supporting discovery of new skills, so people can continue to expand their positive coping skill "toolbox" and identify the practices that work best for them.

*5.1.3 Designing for Personal Context.* In addition to the trends we found in overall skill use, we also found that individual characteristics were often correlated with different levels of effectiveness, both in terms of individual skills and overall improvement throughout the study. As we discussed in Section 4.3, the type of disorder an individual had sometimes correlated with different levels of skill effectiveness, as did their education level, age, medication use, and the number of physically close family members. Section 4.4 revealed that many of those characteristics also correlated with differences in depressive symptom improvement throughout the study. Our investigations in Section 4.4 additionally revealed that people who had larger best improvements and post-ratings in individual skills tended to improve more in depressive symptoms, so helping people find skills that are effective for them could also help them improve their mental health overall.

Our preliminary results also indicate that individual preferences may influence effectiveness: as we detailed in Section 4.4, participants who preferred the *Addiction Skills* module reported higher improvements in depression after the study that those who preferred other modules. However, we cannot confirm whether practicing those skills leads to better skill-level improvements because the module did not include any Likert-scale ratings.

Given these differences between subgroups of people, future designs should account for individual characteristics and preferences in intervention activities. Future studies should also investigate how to better understand, acknowledge, and counteract any disparities or detrimental effects that could result from an individual's characteristics and preferences.

## 5.2 Opportunities for Intelligent Support

Researchers have investigated how technology could use predictive models to identify and support people with mental health conditions (see Section 2.2). Machine learning has also been used to match stress relief interventions to particular individuals and contexts (e.g., [41, 48]). Mohr et al. recently examined a recommender system that identified skills an individual was particularly likely to use, finding that such recommendations resulted in improved depressive symptoms [31]. Preliminary evidence therefore suggests that incorporating such models into technology-delivered mental health interventions could better support people in engaging with psychotherapy.

People using Pocket Skills must currently discover which skills are most helpful to them on their own and must remember to use those skills during times of distress. However, our preliminary modeling results indicate promise for predicting whether a skill would yield improvement based on participant and skill use characteristics, indicating a potential to intelligently identify effective skills (see Section 4.5). Future studies with more participants, as well as skills designed to consistently include and elicit appropriate pre- and post-ratings, could generate data for more advanced modeling and prediction of skill effectiveness.

Predictive models would enable apps to generate skill recommendations, better supporting people in discovering and using effective skills. Our analyses suggested that some skills may be more effective than others, and our feature ablation study similarly revealed that the skill itself was consistently an important feature for predicting skill improvement. As we describe in Section 5.1.1, these differences in effectiveness may be due to the translation of the skill into app content rather than the skill itself. However, if certain skills do tend to be more effective than others, an app could suggest that people generally focus on those skills. Our results suggested that without such recommendations, people tend to go through the modules and skills in the order the app presents (see Section 4.1). Even with advanced predictive modeling techniques, people who are new to an app would lack necessary data for personalized recommendations. This common "cold-start problem" of recommender systems [50] can be alleviated through general recommendations based on someone's demographics or other static features. Expert advice could also inform recommendations based on specific conditions and characteristics (e.g., by working with a psychologist to develop general recommendations). Such recommendations would enable even new users to discover skills that are more likely to help them in a given moment.

In this paper, we explored a one-size-fits-all predictive model, in which a single model is used to generate all predictions (see Section 4.5). We found that models trained on our current set of features can predict individual improvements more successfully (73.1%) compared to the base rate of the majority class (57.4%). In addition to demonstrating this potential for modeling approaches, we also revealed opportunities for future work to improve predictions. For example, we found that certain learning algorithms performed better at predicting skill improvements for different individual skills (see Figure 5). Future approaches could include an ensemble of models to further improve skill recommendations, leveraging: 1) different learning algorithms, to account for any differences in model performance for different skill or participant characteristics, and 2) different models, to account for any differences in data needs for individual skills (e.g., a model for *Distress Tolerance: Self-Soothe* skill could leverage location data).

Our modeling approach was also limited by a lack of richer contextual information. Given more complete data, a predictive model could allow personalized recommendations based on the contexts described above (e.g., in-the-moment environment and emotions, preferences and goals, personal characteristics). For example, during times of emotional distress, a recommender

system could examine the individual's preferences, current context, and historical app use to suggest skills that are most likely to result in improvements for that person. The same system could help people diversify their skill portfolio by recommending they practice new skills that were helpful to similar individuals. Recommendations could also depend on the individual's environmental context, suggesting different skills when a person is at home, in transit, or on the bus. Such recommendations could further support people in identifying positive coping skills that work for them and implementing those skills in their lives when they need them.

## 6 CONCLUSION

Mobile mental health interventions are becoming increasingly ubiquitous, prompting a need for improved understanding to help designers more consistently base these interventions on evidence-based principles. We analyzed data from a month-long field study of Pocket Skills, a mobile web app designed to provide holistic support for dialectical behavioral therapy (DBT) to help people develop positive coping skills. We identified several factors that contribute to skill effectiveness, including the skill itself and participant characteristics and preferences. We also developed machine learning models to predict skill-based improvements. Based on our findings, we presented design implications for translating evidence-based psychotherapies into application content, including the need to consider different environmental, emotional, and personal contexts. Finally, we discussed opportunities to use machine learning techniques for better mental health support by producing personalized and context-aware skill recommendations based on an individual's personal characteristics, preferences, past application use, and in-the-moment emotions and environmental factors.

## ACKNOWLEDGMENTS

The authors would like to acknowledge Ran Gilad-Bachrach, Javier Hernandez, Daniel McDuff, and Kael Rowan for their advice on data analysis and their insights and feedback on this work as well as Microsoft Research for providing the data used in this research. Research was supported in part by the National Science Foundation under awards DGE-1256082, IIS-1553167, IIS-1813675, and IIS-1901386, the National Institutes of Health under awards R01LM012810 and P50MH115837, an Adobe Data Science Research Award, Supportiv, and the Allen Institute for Artificial Intelligence (AI2). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

## REFERENCES

[1] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-Scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476.

[2] Nazanin Andalibi, Pinar Oxturk, and Andrea Forte. 2017. Sensitive Self-Disclosures, Responses, and Social Support On Instagram: The Case of #Depression. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. https://doi.org/10.1145/2998181.2998243

[3] America Psychiatric Association. 1994. Diagnostic and Statistical Manual of Mental Disorders.

[4] Jakob E Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. 2013. Designing Mobile Health Technology for Bipolar Disorder: A Field Trial of the Monarca System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2627–2636. https://doi.org/10.1145/2470654.2481364

[5] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health. *Psychiatric Rehabilitation Journal* 38, 4 (2015), 313–313. https://doi.org/10.1037/prj0000130

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[7] Nigel E Bush, Steven K Dobscha, Rosa Crumpton, Lauren M Denneson, Julia E Hoffman, Aysha Crain, Risa Cromer, and Julie T Kinn. 2015. A Virtual Hope Box Smartphone App as an Accessory to Therapy: Proof-of-Concept in a Clinical Sample of Veterans. *Suicide & Life-Threatening Behavior* 45, 1 (2015), 1–9. http://doi.org/10.1111/sltb.12103

[8] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation From Mental Health Content in Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2858036.2858207

[9] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression Via Social Media. *ICWSM* 13 (2013), 1–10.

[10] Linda Dimeff and Marsha M Linehan. 2001. Dialectical Behavior Therapy in a Nutshell. *The California Psychologist* 34, 3 (2001), 10–13.

[11] Gavin Doherty, David Coyle, and John Sharry. 2012. Engagement With Online Mental Health Interventions: An Exploratory Clinical Study of a Treatment for Depression. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2207676.2208602

[12] Stefan Rennick Egglestone, Sarah Knowles, Gill Toms, Penny Bee, Karina Lovell, and Peter Bower. 2016. Health Technologies 'In the Wild': Experiences of Engagement With Computerised CBT. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2858036.2858128

[13] National Institute for Clinical Excellence and Great Britain. 2002. *Guidance on the Use of Computerised Cognitive Behavioural Therapy for Anxiety and Depression*. National Institute for Clinical Excellence.

[14] Lina Gega, Isaac Marks, and David Mataix-Cols. 2004. Computer-Aided CBT Self-help for Anxiety and Depressive Disorders: Experience of a London Clinic and Future Directions. *Journal of Clinical Psychology* 60, 2 (2004), 147–57. https://doi.org/10.1002/jclp.10241

[15] Itzhak Gilat, Yishai Tobin, and Golan Shahar. 2011. Offering Support to Suicidal Individuals in An Online Support Group. *Archives of Suicide Research : Official Journal of the International Academy for Suicide Research* 15, 3 (2011), 195–206. https://doi.org/10.1080/13811118.2011.589675

[16] Marianne Goodman, Uday Patil, Lauren Steffel, Jennifer Avedon, Scott Sasso, Joseph Triebwasser, and Barbara Stanley. 2010. Treatment Utilization By Gender in Patients With Borderline Personality Disorder. *Journal of Psychiatric Practice* 16, 3 (2010), 155–63. https://doi.org/10.1097/01.pra.0000375711.47337.27

[17] Wendy Hardeman, Julie Houghton, Kathleen Lane, Andy Jones, and Felix Naughton. 2019. A Systematic Review of Just-In-Time Adaptive Interventions (JITAIs) to Promote Physical Activity. *International Journal of Behavioral Nutrition and Physical Activity* 16, 1 (2019), 31. https://doi.org/10.1186/s12966-019-0792-7

[18] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What Does All This Data Mean for My Future Mood? Actionable Analytics and Targeted Reflection for Emotional Well-Being. *Human–Computer Interaction* 32, 5-6 (2017), 208–267. https://doi.org/10.1080/07370024.2016.1277724

[19] Anna Huguet, Sanjay Rao, Patrick J Mcgrath, Lori Wozney, Mike Wheaton, Jill Conrod, and Sharlene Rozario. 2016. A Systematic Review of Cognitive Behavioral Therapy and Behavioral Activation Apps for Depression. *PLOS One* 11, 5 (2016), E0154248. https://doi.org/10.1371/journal.pone.0154248

[20] Luis G Jaimes, Martin Llofriu, and Andrew Raij. 2015. Preventer, A Selection Mechanism for Just-In-Time Preventive Interventions. *IEEE Transactions on Affective Computing* 7, 3 (2015), 243–257. https://doi.org/10.1109/TAFFC.2015.2490062

[21] Adrienne S Juarascio, Megan N Parker, Madeline A Lagacey, and Kathryn M Godfrey. 2018. Just-In-Time Adaptive Interventions: A novel Approach for Enhancing Skill Utilization and Acquisition in Cognitive Behavioral Therapy for Eating Disorders. *International Journal of Eating Disorders* 51, 8 (2018), 826–830.

[22] Eiman Kanjo, Luluah Al-Husain, and Alan Chamberlain. 2015. Emotions in Context: Examining Pervasive Affective Sensing Systems, Applications, and Analyses. *Personal and Ubiquitous Computing* 19, 7 (2015), 1197–1212.

[23] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine* 16, 9 (2001), 606–13. https://www.ncbi.nlm.nih.gov/pubmed/11556941

[24] Marsha M Linehan. 1993. *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. Guilford Press.

[25] Marsha M Linehan. 2014. *DBT Skills Training Manual*. Guilford Publications.

[26] Marsha M Linehan, Heidi L Heard, and Hubert E Armstrong. 1993. Naturalistic Follow-Up of a Behavioral Treatment for Chronically Parasuicidal Borderline Patients. *Archives of General Psychiatry* 50, 12 (1993), 971–4. https://www.ncbi.nlm.nih.gov/pubmed/8250683

[27] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's Only A Computer: Virtual Humans Increase Willingness to Disclose. *Computers in Human Behavior* 37 (2014), 94–100. https://doi.org/10.1016/j.chb.2014.04.043

[28] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025932

[29] Mark Matthews, Gavin Doherty, David Coyle, and John Sharry. 2008. Designing Mobile Applications to Support Mental Health Interventions. In *Handbook of research on user interface design and evaluation for mobile technology*. IGI Global, 635–656. http://doi.org/10.4018/978-1-59904-871-0.ch038

[30] David C Mohr, Michelle Nicole Burns, Stephen M Schueller, Gregory Clarke, and Michael Klinkman. 2013. Behavioral Intervention Technologies: Evidence Review and Recommendations for Future Research in Mental Health. *General Hospital Psychiatry* 35, 4 (2013), 332–338. https://doi.org/10.1016/j.genhosppsych.2013.03.008

[31] David C Mohr, Stephen M Schueller, Kathryn Noth Tomasino, Susan M Kaiser, Nameyeh Alam, Chris Karr, Jessica L Vergara, Elizabeth L Gray, Mary J Kwasny, and Emily G Lattie. 2019. Comparison of the Effects of Coaching and Receipt of App Recommendations on Depression, Anxiety, and Engagement in the IntelliCare Platform: Factorial Randomized Controlled Trial. *Journal of Medical Internet Research (JMIR)* 21, 8 (2019), e13609. https://doi.org/10.2196/13609

[32] David C Mohr, Kathryn Noth Tomasino, Emily G Lattie, Hannah L Palac, Mary J Kwasny, Kenneth Weingardt, Chris J Karr, Susan M Kaiser, Rebecca C Rossom, Leland R Bardsley, Lauren Caccamo, Colleen Stiles-Shields, and Stephen M Schueller. 2017. IntelliCare: An Eclectic, Skills-Based App Suite for The Treatment of Depression and Anxiety. *Journal of Medical Internet Research* 19, 1 (2017). https://dx.doi.org/10.2196%2Fjmir.6645

[33] Margaret E Morris and Adrian Aguilera. 2012. Mobile, Social, and Wearable Computing and the Evolution of Psychological Practice. *Professional Psychology: Research and Practice* 43, 6 (2012), 622.

[34] Inbal Nahum-Shani, Shawna N Smith, Ambuj Tewari, Katie Witkiewitz, Linda M Collins, Bonnie Spring, and S Murphy. 2014. Just In Time Adaptive Interventions (JITAIs): An Organizing Framework for Ongoing Health Behavior Support. *Methodology Center technical report* 2014 (2014), 14–126. https://doi.org/10.1007/s12160-016-9830-8

[35] Felix Naughton. 2017. Delivering "Just-In-Time" Smoking Cessation Support via Mobile Phones: Current Knowledge and Future Directions. *Nicotine & Tobacco Research* 19, 3 (2017), 379–383. https://doi.org/10.1093/ntr/ntw143

[36] Andrada D Neacsiu, Shireen L Rizvi, and Marsha M Linehan. 2010. Dialectical Behavior Therapy Skills Use as a Mediator and Outcome of Treatment for Borderline Personality Disorder. *Behaviour Research and Therapy* 48, 9 (2010), 832–9. https://doi.org/10.1016/j.brat.2010.05.017

[37] Andrada D Neacsiu, Shireen L Rizvi, Peter P Vitaliano, Thomas R Lynch, and Marsha M Linehan. 2010. The Dialectical Behavior Therapy Ways of Coping Checklist: Development and Psychometric Properties. *Journal of Clinical Psychology* 66, 6 (2010), 563–582. https://doi.org/10.1002/jclp.20685

[38] Sonya B Norman, Shadha Hami Cissell, Adrienne J Means-Christensen, and Murray B Stein. 2006. Development and Validation of an Overall Anxiety Severity and Impairment Scale (OASIS). *Depression and Anxiety* 23, 4 (2006), 245–9. https://www.ncbi.nlm.nih.gov/pubmed/16688739

[39] National Institutes of Health: National Institute of Mental Health. 2015. *Any Mental Illness (AMI) Among U.S. Adults*. https://www.nimh.nih.gov/health/statistics/prevalence/any-mental-illness-ami-among-us-adults.shtml

[40] Mark Olfson, Ramin Mojtabai, Nancy A Sampson, Irving Hwang, Benjamin Druss, Philip S Wang, Kenneth B Wells, Harold Alan Pincus, and Ronald C Kessler. 2009. Dropout from Outpatient Mental Health Care in the United States. *Psychiatric Services* 60, 7 (2009), 898–907. https://doi.org/10.1176/ps.2009.60.7.898

[41] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with Stress Through Pop-culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare* (Oldenburg, Germany) *(PervasiveHealth '14)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 109–117. https://doi.org/10.4108/icst.pervasivehealth.2014.255070

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[43] Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-quality Counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[44] Brenna N Renn, Theresa J Hoeft, Heather Sophia Lee, Amy M Bauer, and Patricia A Areán. 2019. Preference for In-Person Psychotherapy versus Digital Psychotherapy Options for Depression: Survey of Adults in the US. *npj Digital Medicine* 2, 1 (2019), 6. https://doi.org/10.1038/s41746-019-0077-1

[45] Shireen L Rizvi, Linda A Dimeff, Julie Skutch, David Carroll, and Marsha M Linehan. 2011. A Pilot Study of the DBT Coach: An Interactive Mobile Phone Application for Individuals With Borderline Personality Disorder and Substance Use Disorder. *Behavior Therapy* 42, 4 (2011), 589–600. https://doi.org/10.1016/j.beth.2011.01.003

[46] Shireen L Rizvi, Christopher D Hughes, and Marget C Thomas. 2016. The DBT Coach Mobile Application as an Adjunct to Treatment for Suicidal and Self-Injuring Individuals With Borderline Personality Disorder: A Preliminary Evaluation and Challenges to Client Utilization. *Psychological Services* 13, 4 (2016), 380–388. https://doi.org/10.1037/ser0000100

[47] Caryn Kseniya Rubanovich, David C Mohr, and Stephen M Schueller. 2017. Health App Use Among Individuals With Symptoms of Depression and Anxiety: A Survey Study With Thematic Coding. In *JMIR Mental Health*. https://doi.org/10.2196/mental.7603

[48] Akane Sano, Paul Johns, and Mary Czerwinski. 2015. HealthAware: An Advice System for Stress, Sleep, Diet and Exercise. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 546–552. https://doi.org/10.1109/ACII.2015.7344623

[49] Randy A Sansone and Lori A Sansone. 2011. Gender Patterns in Borderline Personality Disorder. *Innovations in Clinical Neuroscience* 8, 5 (2011), 16–20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3115767/

[50] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 253–260. https://doi.org/10.1145/564376.564421

[51] Jessica Schroeder, Chelsey Wilks, Kael Rowan, Arturo Toledo, Ann Paradiso, Mary Czerwinski, Gloria Mark, and Marsha M Linehan. 2018. Pocket Skills: A Conversational Mobile Web App To Support Dialectical Behavioral Therapy. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 398. https://doi.org/10.1145/3173574.3173972

[52] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. 2019. Machine Learning in Mental Health: A Scoping Review of Methods and Applications. *Psychological Medicine* 49, 9 (2019), 1426–1448. https://doi.org/10.1017/S0033291719000151

[53] Donna Spruijt-Metz, Cheng KF Wen, Gillian OâĂŹReilly, Ming Li, Sangwon Lee, BA Emken, Urbashi Mitra, Murali Annavaram, Gisele Ragusa, and Shrikanth Narayanan. 2015. Innovations in the Use of Interactive Technology to Support Weight Management. *Current Obesity Reports* 4, 4 (2015), 510–519. https://doi.org/10.1007%2Fs13679-015-0183-6

[54] Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A Comparison of Natural Language Processing Methods For Automated Coding Of Motivational Interviewing. *Journal of Substance Abuse Treatment* 65 (2016), 43–50.

[55] John T Vessey and Kenneth I Howard. 1993. Who Seeks Psychotherapy? *Psychotherapy: Theory, Research, Practice, Training* 30, 4 (1993), 546. http://doi.org/10.1037/0033-3204.30.4.546

[56] Bruce E Wampold and Zac E Imel. 2015. *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work*. Routledge.

[57] Rui Wang, Andrew T Campbell, and Xia Zhou. 2015. Using Opportunistic Face Logging from Smartphone to Infer Mental Health: Challenges and Future Directions. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 683–692. https://doi.org/10.1145/2800835.2804391

[58] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

# SUPPLEMENTARY INFORMATION

In this document, we further describe the analyzed skills and include more details on the results of our statistical analyses.

## S.1 Modules and Skills

Supplementary Table 1 includes details on the modules and skills examined.

| Module | Skill | Subskill |
|---|---|---|
| Mindfulness | Observe[b] | Breathing |
| | | Sounds |
| | | Visuals |
| | | Everyday Life |
| | Describe[b] | Visuals |
| | | Thoughts |
| | | Expressions |
| | | Everyday Life |
| | Participate[b] | Counting |
| | | Jumping |
| | | Laugh Club |
| | | Saying Words |
| | | Walking |
| | | Everyday Life |
| | NonJudgementally[b] | Observe Judgments |
| | | Rephrase Judgmental Statements |
| | One Mind[b] | – |
| Emotion Regulation | Check the Facts[a] | – |
| | /Brief CtF | – |
| | Opposite Action[a] | – |
| | /Brief OA | – |
| | Problem Solve[a] | – |
| | /Brief PS | – |
| Distress Tolerance | Distract[b] | Comparisons |
| | | Emotions |
| | | Pushing Away |
| | | Sensations |
| | | Thoughts |
| | Self-Soothe[a] | Hearing |
| | | Smell |
| | | Taste |
| | | Touch |
| | | Vision |
| | TIP[b] | Intense Exercise |
| | | Paced Breathing |
| | | Progressive Muscle Relaxation |
| | | Temperature |

**[a]Skills with both pre- and post-ratings (with which we could calculate skill improvement).**
**[b]Skills with post-ratings only.**
**Supplementary Table 1: Modues, skills, and subskills examined in our analyses.**

| Finding | Estimate | *p* Value |
|---|---|---|
| DT Distract via Pushing Away *vs* DT Distract via Comparisons | -1.20 | 0.0227 |
| DT Distract via Sensations *vs* DT Distract via Comparisons | -1.43 | 0.0171 |
| M Participate via Saying Words *vs* DT Distract via Comparisons | -1.04 | 0.0104 |
| M Non-Judgementally Observe Judgments *vs* DT Distract via Pushing Away | 0.992 | 0.0307 |
| M Non-Judgementally Rephrase Judgmental Statements *vs* DT Distract via Pushing Away | 1.25 | 0.0000617 |
| M Non-Judgementally Observe Judgments *vs* DT Distract via Sensations | 1.23 | 0.0302 |
| M Non-Judgementally Rephrase Judgmental Statements *vs* Distract via Sensations | 1.49 | 0.000351 |
| M Non-Judgementally Rephrase Judgmental Statements *vs* Distract via Thoughts | 0.901 | 0.0472 |
| M Participate via Saying Words *vs* M Non-Judgementally Observe Judgments | -0.836 | 0.00468 |
| M Focus On One Thing at a Time *vs* M Non-Judgementally Rephrase Judgmental Statements | -0.710 | 0.00307 |
| M Describe Visuals *vs* M Non-Judgementally Rephrase Judgmental Statements | -0.749 | 0.00234 |
| M Observe Breathing *vs* M Non-Judgementally Rephrase Judgmental Statements | -0.819 | 0.00000986 |
| M Observe Sounds *vs* M Non-Judgementally Rephrase Judgmental Statements | -0.828 | 0.0000127 |
| M Participate via Saying Words *vs* M Non-Judgementally Rephrase Judgmental Statements | -1.10 | 0.0000000218 |

**Supplementary Table 2: Skill effectiveness: examining pairwise differences in skills for post-ratings. Positive deltas indicate the first skill results in *worse* post-ratings than the second skill. M=*Mindfulness*, ER=*Emotion Regulation*, DT=*Distress Tolerance*.**

## S.2 Overall Skill Use

In our Tukey test examining overall skill improvement between subskills, the *Emotion Regulation* skill of *Problem Solve* resulted in more improvement than the *Distress Tolerance* skills of *Self-Soothe via Hearing* (by 0.64 points; $t$=-3.252, $p$<0.05), *Self-Soothe via Touch* (by 0.90 points; $t$=-4.388, $p$<0.01), and *Self-Soothe via Vision* (by 0.63 points; $t$=-3.361, $p$<0.05). The *Emotion Regulation* skill *Check the Facts* also resulted in more improvement than the *Distress Tolerance* skill *Self-Soothe via Touch* (by 0.79 points; $t$=-4.108, $p$<0.01).

See Supplementary Table 2 for results of our Tukey test examining overall post-skill rating between subskills.

## S.3 Skill vs. Scale Improvement

For the PHQ-9 model, our Tukey test examining *age bucket* revealed that being under 25 ($z$=3.414, $p$<0.01) and 25-35 ($z$=3.167, $p$<0.01) years old correlates with more improvement than 35-45 years old. Being under 25 ($z$=2.698, $p$<0.05) and 25-35 ($z$=2.869, $p$<0.05) years old also correlates with more improvement than being greater than 45 years old. Our examination of *family bucket* revealed that having zero family members living close correlates with greater improvement than having 1 or 2 ($z$=3.618, $p$<0.01), 3 or 4 ($z$=3.021, $p$<0.05), or more than 4 ($z$=-2.781, $p$<0.05). Finally, the Tukey results for *favorite module* revealed that preferring the Addiction module correlates in more improvement than the Distress Tolerance module ($z$=3.638, $p$<0.01), the Emotion Regulation module ($z$=3.391, $p$<0.01), and the Mindfulness module ($z$=3.564, $p$<0.01).

For the OASIS model, we found that being 25-35 correlates with greater improvement than being 35-45 ($z$=2.939, $p$<0.05) or older than 45 ($z$=3.278, $p$<0.01).

Consistent with our other analyses, these models use the buckets defined in Section 3.2 for the variables of *age*, *number of family members close*, and *education*. If these variables are instead treated as continuous, the resulting models yield no significant effects. This discrepancy may be due to a nonlinear relationship to those variables or due to our relatively limited dataset.

## S.4 Models

Supplementary Table 3 presents a list of skill use characteristic features used in predictive modeling. Supplementary Table 4 presents the classifier performance across the four different learning algorithms and various feature sets.

| Feature | Description | Type |
|---|---|---|
| Skill ID | Unique skill identifier | Categorical |
| Day of week | Day of week for skill use | Categorical |
| Time of day | Time of day for skill use grouped by morning (6-12), afternoon (12-18), evening (18-24), and night (0-6). | Categorical |
| Consecutive use of any skill | Boolean indicating whether or not another skill is used within 15 minutes prior to the current skill use | Categorical |
| Consecutive use of the same skill | Boolean indicating whether or not the same skill is used within 15 minutes prior to the current skill use | Categorical |
| Cumulative average skill improvement | Cumulative average of skill improvements since the study intake | Numerical |
| Last improvement | Skill improvement of the last used skill (0 for the first skill use) | Numerical |
| Pre-rating | Pre-rating of emotional intensity or distress | Numerical |

**Supplementary Table 3: Skill use characteristic features used for training predictive models.**

| Feature set | DT | LR | GB | RF |
|---|---|---|---|---|
| All | 0.721 / 0.628 | 0.643 / 0.665 | 0.705 / 0.699 | 0.659 / 0.632 |
| All but skill | 0.62 / 0.457 | 0.636 / 0.616 | 0.64 / 0.565 | 0.663 / 0.648 |
| All but pre-rating | 0.721 / 0.628 | 0.663 / 0.69 | 0.698 / 0.708 | 0.667 / 0.657 |
| All but mood disorder | 0.721 / 0.628 | 0.655 / 0.685 | 0.709 / 0.703 | 0.64 / 0.621 |
| All but anxiety disorder | 0.721 / 0.628 | 0.64 / 0.666 | 0.698 / 0.707 | 0.671 / 0.609 |
| All but education | 0.721 / 0.628 | 0.643 / 0.671 | 0.671 / 0.703 | 0.698 / 0.677 |
| All but favorite module | 0.721 / 0.628 | 0.651 / 0.664 | 0.713 / 0.703 | 0.655 / 0.672 |
| All but least favorite module | 0.721 / 0.628 | 0.647 / 0.659 | 0.694 / 0.705 | 0.647 / 0.662 |

**Supplementary Table 4: Predictive model performances (denoted by "test accuracy / test AUC") across four different learning algorithms—decision tree (DT), logistic regression (LR), gradient boosting (GB), and random forest (RF)—and various feature sets.**