Designing a Personalization Intervention To Reduce Churn in Exercise Mobile Apps: A Research Proposal

Chris Li

Follow this and additional works at: https://scholarship.claremont.edu/cmc_theses

Part of the Cognitive Psychology Commons, Health Psychology Commons, and the Human Factors Psychology Commons

Recommended Citation
Li, Chris, "Designing a Personalization Intervention To Reduce Churn in Exercise Mobile Apps: A Research Proposal" (2024). CMC Senior Theses. 3430.
https://scholarship.claremont.edu/cmc_theses/3430

This Open Access Senior Thesis is brought to you by Scholarship@Claremont. It has been accepted for inclusion in this collection by an authorized administrator. For more information, please contact scholarship@claremont.edu.
Claremont McKenna College

Designing a Personalization Intervention To Reduce Churn in Exercise Mobile Apps: A Research Proposal

submitted to
Professor Gabriel Cook

by
Christopher Li

For Senior Thesis
Fall 2023
December 4th, 2023
# Table of Contents

Abstract 3

Introduction 4

Hypotheses 8

Methods 9

Research Setting and Participants 10

Variables 11

Personalization 11

Surveys (Perceived Ease of Use and Perceived Usefulness) 15

User Engagement (UE) 18

Procedure 20

Data Analysis 22

Binary Logistic Regression 23

Odds-Ratios 24

Probabilities 24

Likelihood Ratio Test 25

AIC/BIC Test 26

AUC-ROC Test 27

Wald Test 28

Mann-Whitney U Test 28

Combining Tests to Understand Regression Results 29

Discussion 30

Building Upon Previous Research 31

Unanswered Questions and Limitations 32

Real-world application 35

References 36
Abstract

Sports and fitness mHealth app development has rapidly increased since the COVID-19 pandemic due to changes in living situations that increased the demand for exercising remotely. App developers struggle to understand the various strategies in reducing the amount of users that churn as time progresses. This research proposal will primarily focus on the effect that personalization has on churn, seeing the increased development of mobile health apps, strong desire for curated experiences, and the recent relevance of AI. To fully understand the relationship between personalization and churn, moderator effects of perceived ease of use, perceived usefulness, and user engagement will be investigated. A usability study with physically active college students in the US is proposed to collect data for analysis, with the hopes of providing app developers with a greater understanding of implementing personalization.
Introduction

Health and fitness goals are known to be abandoned much quicker than intended, and abandonment is no different with virtual and online providings. Today, 91% of Americans own a smartphone (Marino, 2023) and spend an average of 3.5 hours per day on them (Howarth, 2023). In turn, people are using mobile health (mHealth) apps, and people tend to switch between apps for reasons such as loss of novelty or provided utility (Herrmann & Kim, 2017). Out of all digital health app users, only 20.2% retain after the first 2 days, 8.5% after a week, and 2.5%-4% after a month, showing how churn rapidly increases over time (Handa, 2022). While plenty of research has studied the intentions to use fitness apps, not many directly study the mechanisms behind churn. Understanding churn would allow for companies to develop app features that mitigate it, leading to targeted marketing and more sustainable business models.

mHealth apps are more important than ever in helping people stay healthy due to the COVID-19 pandemic forcing people to quarantine at home. As a result, people could no longer workout in gyms and had to adapt exercise habits to be suitable at home, resulting in lower exercise frequencies and durations (Gjestvang, Tangen et al., 2022). However, the health and fitness iOS market is projected to exceed expected growth by 29.9% (Kalogtra, Raja et al., 2022); their focus has shifted from dieting to sports and fitness because of the increase in demand for working out remotely (Angosto, Grimaldi-Puyana et al., 2023). The main draw of mHealth apps is in its convenience, which remove extra steps in visiting websites and commuting to gyms so that committing to health is more accessible (Amagai, Pila et al., 2022). Additionally, public mental health took a massive toll because of uncertainties in health conditions and health insurance (Cui, Lu et al., 2022). Symptoms of depression, anxiety, stress, and sleep disorders increased substantially due to limited social interaction and financial
uncertainty, and home-based exercise proved to be a key protective measure (Rajkumar, Rajan et al., 2022). Essentially, mobile health apps make exercise convenient, which helped alleviate public mental health during the COVID-19 pandemic and continues to do so today.

Understanding use cases and outcomes of using mHealth apps can help pinpoint key features to develop. Research has primarily focused on targeting inactive groups like obese individuals and individuals with chronic illnesses, rather than active groups (Larango, Ding et al., 2020; Hi-Park, Hwang et al., 2019). Athletic groups that have been investigated include runners and individuals that exercise in fitness centers. mHealth apps provide runners with positive feelings towards exercise, though only for serious runners and not for recreational runners (Dallinga, Mennes et al., 2015), potentially indicating that exercise identity is an underlying component of the relationship between exercise and mHealth usage (Barkley, Lepp et al., 2020). Using mobile health apps does not lead to improved habits or increased motivation to continue exercising in fitness centers (Torrente, Javaloyes et al., 2021), potentially signaling that commuting to fitness centers is too great of a barrier or that fitness centers are not linked to mHealth usage. Ultimately, exploring different applications of health and fitness apps can enable developers to create experiences that promote exercise.

Mobile health apps have proven to increase physical activity, motivate intense exercise, help overcome barriers to exercise, and increase self-efficacy in active users as opposed to non-users or previous users (Litman, Rosen et al., 2015). Individuals tend to continue using mHealth apps that take user characteristics and activity into consideration to provide personalized content (Rajan, 2019; Angosto, Grimaldi-Puyana et al., 2023). Personalization can be defined as the ability to provide tailored content and services to individuals by using knowledge about their preferences and behaviors (Adomavicius & Tuzhilin, 2005; Fan et al.)
(2006) defines personalization as “a process which changes the functionality, interface, content, or distinctiveness of system to increase its personal relevance”. A mobile app that considers user demographics, user characteristics, and intentions at the start of app usage can potentially increase user adoption (Damsgaard et al., 2007), while continuous updates from user-collected data can help goal attainment and support continued usage (Vaghefi, Tulu et al., 2019). Users report that apps with personalized experiences provide them with feelings of competence and relatability, subsequently aiding users in pursuing growth-oriented activities and creating a strong brand loyalty (Oulasvirta, Blom 2007). Moreover, incorporating personalization almost always involves the use of artificial intelligence (AI) in some way. Recommender systems are the main focus in mobile apps, which use machine learning algorithms to analyze user behaviors and/or behaviors of similar users to provide the most relevant recommendations (Sarker, Uddin et al., 2021). Ultimately, incorporating personalization motivates behavior change by allowing users to select and follow preferred app selections (Tong, Quiroz et al., 2022).

The TAM (Technology Acceptance Model) was created by Davis (1989) and is still commonly used to understand why people adopt and use technology. Perceived ease of use and perceived usefulness work together to determine users’ intention to use technology, which then shapes their usage behavior. Perceived ease of use is defined as “the degree to which a person believes that using a particular system would be free of effort” (Davis F.D., 1989). Improving perceived ease of use of technology encourages continued usage by creating a less labor intensive experience (Wang & Qi, 2021); incorporating personalization improves perceived ease of use by providing easier access to relevant content (Li, 2020). When users believe that an app is easy to use, useful, and are highly engaged, their likelihood to churn decreases (Guo et al., 2022). Perceived usefulness refers to “the degree to which a person believes that using a
particular system would enhance his or her job performance” (Davis F.D., 1989). Kargin and Basoglu (2006) investigated factors that influence mobile service adoption and found that personalization had a direct impact on perceived usefulness. Seeing that the provided utility of apps decreases overtime (Herrmann & Kim, 2017), improving perceived ease of use and perceived usefulness overtime should prevent user churn.

User engagement is commonly used as a metric to define success, and can be defined as “a quality of the user experience that emphasizes the positive aspects of interaction – in particular the fact of being captivated by the technology” (Attfield, Kazai et al., 2011). User engagement and user retention are frequently discussed together, as any amount of engagement means that the app is still used to some extent (Pal, Baur et al., 2020, Amagai, Pila et al., 2022). Moreover, 90% of users that use an app at least once a week are retained (AppsFlyer, 2022). Thus, personalization should provide relevant content to create a positive and captivating user experience, where improvements in user engagement overtime should prevent user churn.

A fitness app that creates a personalized experience should provide users with relevant content, causing users to feel special, cared for, and motivated to exercise. When moderated by improved measurements of perceived ease of use, perceived usefulness, and user engagement overtime, the product sustains novelty and utility, causing users to feel loyal to the product. Ultimately, the degree of churn will decrease. A fitness app that fails to create a personalized experience would likely have decreased ratings of perceived ease of use, perceived usefulness, and user engagement overtime, as it will take more effort to find desirable workouts, and the less relevant content should result in a higher degree of churn.

However, an alternative result is that there are no significant findings of personalization impacting churn, suggesting that the features are not creating a special experience that attracts
users throughout usage. Because personalization utilizes user data, poor implementation could lose user trust as a result of privacy violations; a poor personalization model could over personalize or show irrelevant contents and messages that lead to frustration (Artug, 2022). Rather than feeling special, cared for, and motivated to exercise, participants will be less confident and safe using the product. Additionally, the lack of interaction effects with moderators would imply that personalization fails to influence perceptions of product novelty and loyalty over time, failing to add additional insight towards the relationship between personalization and churn.

**Hypotheses**

Overall, the research will contribute towards mHealth app development by providing an understanding of how personalization can impact if users decide to retain or churn.

*Personalization Predicts the Degree of Churn*

**Hypothesis 1:** There will be a main effect between personalization and churn, such that when compared with no personalization features, personalized workouts and educational content will cause users to feel special, cared for, and motivated to exercise. As a result, users will develop strong loyalty to the product and the degree of churn will decrease

**Hypothesis 2:** Perceived ease of use, perceived usefulness, and user engagement will moderate the relationship between personalization and churn. Increases or decreases in moderator variables after the 5 month period demonstrate that personalization features influence novelty and utility over time, explaining the relationship between personalization and churn
Personalization Fails to Predict the Degree of Churn

Hypothesis 1: There will be no main effect between personalization and churn, such that when compared to no personalization features, personalized workouts and content breach user privacy or provide inadequate recommendations. As a result, users feel less confident and safe in using the product, which suggests that a different personalization strategy or model is more effective.

Hypothesis 2: Perceived ease of use, perceived usefulness, and user engagement will not moderate the relationship between personalization and churn. Increases or decreases in moderator variables after the 5 month period fail to demonstrate that personalization features influence novelty and utility over time, and cannot explain the relationship between personalization and churn.

Methods

This study will use an experimental approach, where participants use a fitness-focused app that provides workouts varying in aspects such as duration, intensity, and equipment requirements; a variety of educational content relevant to the user groups will be provided as well. Participants will voluntarily use the app for up to 5 months and be randomly placed into 1 of 4 groups to ensure internal validity and reduce bias. The chosen usage duration was based on a prior study (Herrmann, Kim et al., 2017), where over the course of 5 months, users’ perceived ease of use, perceived usefulness, and exercise frequency decreased significantly because the app did not enhance their workouts.
Research Setting and Participants

Eligible participants will be students at a university in the US aged 18-23 years old who own an iPhone 11, or a newer version of the iPhone, to ensure that the app functions smoothly and is not influenced by the operating system. Participants also need to ensure able access to the internet. University students are chosen because they are likely to be more physically active than older people and are capable of adhering to tasks. Gender will be split as evenly as possible, seeing how men tend to be more physically active than women (Zimmermann-Sloutskis, Wanner et al., 2010). Additional requirements to ensure that early-stage churn does not occur because of lack of experience with fitness apps are: participants have no prior experience using the app, the chosen app is a version of one that already exists, and participants have used similar apps before. A pre-existing app would be preferred over building an app from scratch to reduce costs towards development and provide a more refined experience. Personal data (i.e. weight, height, activity) will be collected to create an effective personalized experience.

The sample size will depend on a variety of factors (i.e. budget, equal number of men and women), but will most likely be on the lower end, seeing how long term usability studies struggle to recruit large numbers of participants. A recent study incorporating usability testing had 20 participants because of practical issues with budget and emphasis on qualitative feedback (Monteiro-Guerra, Rivera-Romero et al., 2022), and usability tests in empirical research rarely have more than 125 participants. A higher sample size is ideal to reduce bias and error in quantitative analysis, and unlike the aforementioned study (Monteiro-Guerra, Rivera-Romero et al., 2022), this proposal will seek to evaluate effectiveness of the intervention. Participant recruitment will be performed based on the chosen university’s best methods, such as a research platform or emails, to acquire a large sample size.
Due to the 5 month research period, it makes sense to compensate participants for their time and effort. Compensation can be either in the form of a set dollar amount, gift card, or entry into a raffle to win a prize, depending on budget constraints. However, the degree of compensation should be carefully considered, as it could coerce or influence participant activity and self-assessment responses (Largent, Lynch 2018).

Variables

Personalization

The app’s personalized features will provide personalized workouts and personalized feedback. Participants will be randomly assigned to 1 of 4 groups, each of which has a different combination of personalization characteristics.

- **Group 1:** Has personalized workouts and personalized educational content (Workouts, Educational Content)
- **Group 2:** Has personalized workouts but no personalized educational content (Workouts, No Educational Content)
- **Group 3:** Has no personalized workouts but has personalized educational content (No Workouts, Educational Content)
- **Group 4:** Has neither personalized workouts nor personalized educational content (No Workouts, No Educational Content)

An onboarding process will provide initial data for the app to personalize content. Personalized content will be data-driven to reduce user input bias (Tong, Quirez et al., 2022).
The personalization system will update continuously through usage. Users who enjoy personalization features want apps to regularly update, rather than provide baseline personalization once without ever adapting (Rabbi, Aung et al., 2018). However, most fitness apps only capture user behavior and preferences at one point in time, which decreases the value of personalization over time (Tong, Wing et al., 2013). If a pre-existing app is chosen, a great example is the Nike Training Club app, which contains workout videos and educational content that adapt based on user activity and preferences.

Examples of the Nike Training Club app, which provides personalized workouts and educational content on the “For You” page

**Recommended Workouts.** Rajan et al. (2019) discovered that activity recommendations are an attribute of an effective personalization system. The recommended workouts in this study
will be found in the “For You” page, and be based on the information provided during onboarding so that workouts are tailored to the user’s preferences. Workouts will be in video form, as evidence shows that video platforms like YouTube improved exercise motivation in resistance training and high intensity interval training (McDonough, Helgeson et al., 2022). Resistance training, high intensity training, and yoga/pilates are chosen as common workouts performed on mHealth apps. Users will have the option to rate workouts after completing them, which feed into the system to continuously recommend new workouts. Additionally, workouts will be monitored based on the time and day performed, so that relevant recommendations are provided to users. For example, if a user does the same yoga workout for 2 weeks in a row on a Tuesday afternoon, it will be the first recommendation on their homepage when they open the app on subsequent Tuesday afternoons. Participants will also be able to save workouts, where saved workouts will be considered most relevant and weigh heavily into the recommendation algorithm. Participants in groups with “No Workout” as part of the condition (groups 3 and 4) will all perform the study with the same app interface that recommends random workouts irrespective of activity and feedback.

Including personalized feedback improves workout effectiveness by motivating and providing positive feelings to users (Rajan et al., 2019). Motivational messages create positive reinforcement that targets and enhances intrinsic belief in the user to motivate continued usage (Alley, Jennings et al., 2016). Therefore, instructors will motivate users throughout workouts because exercise motivation increases in the long-term when instructors are perceived as supportive (Gaesser, Maakestad et al., 2020). After completing workouts, users’ names will be utilized in messages as hearing or seeing one’s name is shown to create a positive response in the brain (Carmody & Lewis, 2007). Additionally, users will be greeted on the homepage and
congratulated with special messages after completing workouts using their names. Lastly, participants’ exercise sessions have shown to increase from 4.7 (control group) to 7.2 (intervention group) when incorporating personalized feedback (Voth, Jung et al., 2016).

**Personalized Educational Content.** Effective personalization systems tend to retain users by incorporating educational content that increases user knowledge on exercise techniques and importance (Rajan et al., 2019). Exposure to new information on a topic, regardless of what it is, has shown to increase interest in learning more about the topic, motivating further exploration (Ditta, Strickland-Hughes et al., 2020). Educational content will be accessible to users via the “For You” page, where the majority of recommended content will be articles related to topics based on the inputted preferred type of workouts. Topics related to non-selected types of workouts will be included to inspire interest in diversifying workout types. Users will have the option to rate educational content so that the system continues to recommend relevant content. Participants in groups with “No Feedback” as part of the condition (groups 2 and 4) will all perform the study with the same app interface that does not provide personalized messages or educational content based on activity and feedback.

**Surveys (Perceived Ease of Use and Perceived Usefulness)**

The research will use a total of 2 self-assessment questionnaires (or surveys), one for perceived ease of use (PEOU) and one for perceived usefulness (PU). When developing surveys, there are 4 important steps: conceptualization, development, assessing content validity, and refinement to assess construct validity and reliability (O’Brien & McCay-Peet, 2017). However, this study will not create its own survey, instead adapting existing reputable surveys to measure
user attitudes towards exercising with the app. When utilizing existing surveys, intent behind survey usage must be considered, as significant changes such as removing questions or altering question wording to change its meaning can eliminate reliability, validity, and dimensionality (O’Brien & McCay-Peet, 2017).

The TAM model developed by Davis (1989) has proven that PEOU and PU impact users’ attitude towards using technology, subsequently affecting behavioral intention to use, then actual system usage (Cho et al., 2020). To measure PEOU and PU, Davis (1989) conducted correlational, regression, and factorial analyses to ensure he designed the “Perceived Ease of Use and Usefulness User Acceptance Model” to be reliable and valid. As is, the wording of the questions is very general and not tailored to workouts and physical activity, so the wording will be altered without changing intent. The number of questions will remain the same to avoid face validity, and the 7-point ratings scales (“strongly disagree” to “strongly agree”) will remain the same as 7-point scales fall in the reliable scale range (O’Brien & McCay-Peet, 2017). There is evidence that 11-point scales would produce the most effective and normalized results, but only if the intent is to treat them as an interval scale (Wu & Leung, 2017). Thus, the scale will be treated as ordinal with uneven intervals, since interpretations of Likert scales tend to vary from person to person. For example, the difference between “strongly agree” and “moderately agree” differs from the difference between “moderately agree” and “slightly agree”. Some people may need to have an exceptional experience to “strongly agree” with a statement, while “moderately agree” and “slightly agree” may be perceived as relatively the same. Because responses are typically not normally distributed, observations for PEOU and PU will use the median of responses instead of the mean (Sullivan et al., 2013).
Technology Acceptance Model (TAM)

Another debate is in the validity and reliability of self-reported metrics, since responses do not consider individual biases, especially for post-tests (Rosenman, Tennekoon et al., 2011). Despite potential bias of memories, retrospection of user experiences can be an effective measurement as "people like happy endings and prefer experiences that improve rather than ones that get worse", assuming that participant responses are based on their most salient experiences (Kujala, Roto et al., 2011). The survey will be distributed twice throughout the study, once after the first week and once after the participant completes the study. Notably, perceptions of PEOU and PU are shown to diminish over time as a result of technology losing novelty (Drehlich et al., 2020). Thus, the difference in the median of participant responses for perceived ease of use and perceived usefulness respectively will be used for analysis. Essentially, the Likert difference scores of PEOU and PU aim to account for changes in user attitudes that may impact the relationship between personalization and churn.

Survey 1

Perceived Ease of Use (all questions on a scale “strongly disagree” - “strongly agree")
1. Learning to operate the app is easy for me
2. I find it easy to do what I want to do when using the app
3. The features and capabilities are clear and understandable
4. I find the app flexible to interact with
5. I find it easy to become familiar with the app’s functionality
6. I find the app easy to use

Survey 2

Perceived Usefulness (all questions on a scale “strongly disagree” - “strongly agree”)

1. Using the app for my workouts enables me to accomplish them quicker
2. Using the app improves my workout performance
3. Using the app for my workouts increases my productivity
4. Using the app enhances how effective my workouts are
5. Using the app makes working out easier
6. I find the app useful for working out

User Engagement (UE)

Throughout participation in this study, users’ actions will be monitored by system-captured mechanisms to measure user engagement (UE), seeing how self-reporting can increase bias and error (Tong, Quiroz et al., 2022). Engagement is commonly assessed with behavioral metrics like click through rates (CTR), app visits, and neurophysiological techniques like eye-tracking; self-assessment questionnaires like the 31-item UES (User Engagement Scale) are also commonly used (O’Brien, Cairns et al., 2018). Another approach to measure engagement is to focus on qualitative measures like focus groups, interviews, and think aloud
sessions (Short, DeSmet et al., 2018). However, these approaches either fail to give a comprehensive view of UE (CTR, app visits), rely too much on self-reporting (UES), use expensive technology (eye-tracking), or can lead to bias and unstructured data (qualitative measures). Consequently, UE will be a categorical variable measured as the average of the rate of workouts completed and the rate of educational content completed. Completion rates will be split into 3 groups: high engagement (>75%), medium engagement (>50%), and low engagement (≤50%). Ultimately, user engagement will provide additional insight towards performance of personalized workouts and educational content to better understand personalization’s impact on churn.

**Workout Completion Rate.** The following equation will be used to evaluate the workout completion rate: \[
\frac{\text{number of workouts completed} - \text{number of workouts clicked}}{\text{number of workouts completed}}.
\]
To track the number of workouts clicked, the app will track when the play button for workouts is clicked; to track the number of workouts completed, the time spent watching the video will have to match the length of the video. This method of tracking the number of workouts clicked takes into account instances where users abandon a workout or skip through the video due to lack of relevance and/or do not actually perform the workout. However, shortcomings in collected data may arise from sudden emergencies that force users to abandon workouts.

**Educational Content Completion Rate.** The following equation will be used to evaluate the educational content completion rate:
\[
\frac{\text{number of educational content fully consumed} - \text{number of educational content clicked}}{\text{number of educational content fully consumed}}.
\]
To track the number of educational content clicked, the app will measure the number of times that a user clicks on an
article; instances where users scroll to the bottom of an article will count as fully consumed. Instances where users lose interest midway through reading demonstrate that the article was either irrelevant or ineffective at sparking interest in the topic. Similar to the workout completion rate, data may be flawed from not considering scenarios unrelated to interest that force users to abandon an article. Participants may also skim contents to gauge interest, but end up not reading the article or developing interest, which would reduce the data’s validity.

**Procedure**

People interested in participating will fill out a web-based survey using software such as Qualtrics to verify that they are a proper fit for the study. Participants will then be asked demographic questions regarding their age, gender, college class year, and workout experience level (beginner, intermediate, advanced, athlete); access to technology will be assessed by asking if participants own an iPhone 11 or newer and have reliable internet access. Participants will also be asked about their prior experience and familiarity with mHealth apps that provide workouts/workout plans. Accepted participants will be asked to fill out another web-based survey with a consent form describing the purpose of the study, use of responses for publication, consent to collect personal data, and informing their right to withdraw from the study. After agreeing to the terms of the consent form, participants will receive an email with instructions for getting started.

After downloading the app, participants in groups with personalization (groups 1, 2, and 3) will open the app and be immediately prompted to complete the onboarding process by answering 5 questions.
Onboarding Questions

1. Name and age
2. Workout experience level (beginner, intermediate, advanced, athlete)
3. Preferred type of workouts (resistance training, high intensity interval training, yoga/pilates)
4. Preferred length of workouts (5-10 minutes, 10-30 minutes, 30-60 minutes)
5. Preferred number of days to workout in a week
6. Access to equipment (no equipment, dumbbells, cardio machines, gym access)

Users will not be limited to their initial inputs, and will be able to alter preferences in their profile settings. Completion of the onboarding process will then prompt the app to curate content and bring the participant to a personalized home page, where the extent of personalization depends on the group that is assigned.

At this point, the app is completely set up and personalized, and participants will use the app as desired for the remainder of the study. After 1 week of usage and at the end of the 5 month study duration, participants will complete the 2 survey questionnaires that assess PEOU and PU. Throughout the study, user engagement will be monitored whenever the workouts and educational content are interacted with. During usage, participants have the option to stop using the app whenever, for whatever reason. These participants will then be given the survey questionnaires, and be marked as “churned” \( Churn = 1 \). Participants that sustain usage for the study’s 5 month duration will be given the survey questionnaire and be marked as “retained” \( Churn = 0 \).
Data Analysis

A binary logistic regression will be used to evaluate whether personalization, perceived ease of use, perceived usefulness, and user engagement predict churn with main effects and interactions. Personalization and user engagement will be categorical variables; personalization groups will be randomly assigned, while user engagement will consist of the average of participant workout and educational content completion rates. Perceived ease of use and perceived usefulness will be ordinal variables with possible scores from 1 to 7, where each observation will utilize participant responses before and after the intervention to generate median and difference scores. Churn will be a binary variable determined by whether or not a participant reaches the end of the 5 month study duration.

The following results section will discuss appropriate statistical analyses that can provide an intricate understanding of main effects, interactions, and theoretical implications. The binary logistic regression provides valuable information, but cannot be fully understood without additional tests that evaluate goodness of fit and model performance. First, the constant effect of the predictor and moderator variables on churn need to be understood through odds-ratios and probabilities. The Likelihood Ratio Test will evaluate goodness of fit by comparing the full model with reduced models; the AIC/BIC tests will help determine the degree of model complexity that best measures churn; the AUC-ROC tests will fit the full and reduced models under a curve to evaluate model performance; the Wald test will test significance of individual coefficients. Finally, the Mann-Whitney U test will be used to investigate significant differences in the before (responses after 1 week) and after responses (responses after the intervention) in the PEOU and PU surveys.
Binary Logistic Regression

The binary logistic regression model will be the full model used to predict whether a participant will choose to retain or churn. The model will consider each predictor individually, as well as all of the interactions between predictors. Two models should be tested and compared, where the Likert difference scores are used for one, and the after scores are used for the other. The former would evaluate if prolonged usage causes responses to differ, while the latter would evaluate if responses at the end of the intervention predict retention or churn. When deciding between the models, it will be important to consider goodness of fit, coefficient significance, and discrimination between predictors based on collected data. Overall, characteristics and results of the model need to be fully understood to develop theoretical implications.

If any of the 4 personalization groups have a significant p-value \(p < 0.05\), then there is sufficient evidence that personalization has a main effect on churn. Then, significant interaction effects between personalization groups and moderator variables should be investigated to determine a more nuanced prediction of churn. Strong coefficient estimates would either be negative to signal that an increase in the coefficient leads to a decrease in churn, or slightly positive to signal that an increase in the coefficient leads to little churn. When comparing standard errors and z-values between coefficients, lower standard errors indicate that the sample mean is a more precise estimate of the population mean, and higher z-values \((z \leq \pm 1.96)\) indicate that the null hypothesis can be rejected. However, if none of the 4 personalization groups hold significance \((p > 0.05)\), then it is likely that the study design and model characteristics do not create a personalization system that reduces churn. Insignificant results would likely imply that personalized workouts and educational content create an ineffective
personalization system, and/or perceived ease of use, perceived usefulness, and user engagement do not moderate the relationship between personalization and churn.

**Odds-Ratios**

Odds ratios are calculated by exponentiating the log-odds of coefficients \( e^{\text{coefficient}} \), which more accurately represents the odds of churn from a one-unit increase of the coefficient. An odds-ratio greater than 1 implies that the coefficient leads to an increase in churn, while an odds-ratio less than 1 implies that the coefficient leads to a decrease in churn. Coefficients will likely all have an odds-ratio greater than 1, so lower values will be interpreted positively. Coefficients with odds-ratios less than 1 would convincingly demonstrate that the model provides an avenue to decrease churn. The 95% confidence intervals for the odds-ratios inform the likely range of the true odds-ratios for the population and should be considered as well. When the range of coefficient confidence intervals does not include 1, the odds-ratio is statistically significant.

**Probabilities**

The probabilities, calculated as \( \frac{\text{odds-ratio}}{1 + \text{odds-ratio}} \), are important metrics when communicating the implication of odds-ratios to stakeholders. Probabilities range from 0-1, and inform how likely the odds-ratio for a particular coefficient influences churn. The strength of probabilities depends on context, and can be determined by comparing with other probabilities and/or with stakeholder desired thresholds. For example, a common threshold is 0.5, where values above demonstrate that the odds-ratio is likely to occur for churn, and values below demonstrate that the odds-ratio is unlikely to occur for churn. Furthermore, the predictive power
of the model also influences interpretation, where well-fitted models with high accuracy (for example, high area under the ROC curve) imply that probabilities accurately represent churn outcomes. Significance can be determined by analyzing the confidence intervals, where significant coefficients have confidence intervals that do not include 1. Then, the strongest coefficient on churn would have the lowest odds-ratio and highest probability. If all coefficients have confidence intervals that include 1, then the model unsuccessfully measures personalization’s impact on churn.

**Likelihood Ratio Test**

The Likelihood Ratio Test (LRT) will compare the full model with reduced models to evaluate goodness of fit, specifically looking at which variables are contributing towards strengthening the model. By comparing the likelihood ratio statistic (LRS) with the chi-square distribution, the observed results can be compared with the expected results; a significant p-value ($p < 0.05$) would reject the null hypothesis and indicate that the full model is more effective than the reduced model. Typically, more predictors in the reduced model should have higher chi-square values as it is able to capture more complexity and be more flexible. Adding complexity without improving the model implies that certain predictors are irrelevant to churn, overfit the model, and/or causing multicollinearity.

The LRS should be performed on all combinations of reduced models, seeing how there are only 7 combinations. If the collected data set demonstrates significance when comparing the full model to all of the reduced models ($p < 0.05$), then it could be concluded that the full model effectively measures the relationship between personalization, PEOU, PU, and UE on churn. If the LRS is insignificant ($p > 0.05$) when comparing the full model to any reduced model, then the reduced model should be chosen in favor over the full model, and be treated as
the new full model. The full model should be compared with reduced models until all LRS values are significant, so the best suited model for predicting churn is utilized. Otherwise, the chosen model will lack validity, generalizability, and reliability in proving the relationship between personalization and churn in mHealth apps.

**AIC/BIC Test**

The AIC/BIC test evaluates goodness of fit in relation to the model’s complexity by comparing the respective results; for both AIC and BIC, lower values indicate a better-fitting model. AIC/BIC values can be analyzed individually or together. Individually, the BIC penalizes complex models more and favors simpler models than the AIC, and model selection depends on considerations like sample size and simplicity. Together, the difference between the two values evaluates the balance between goodness of fit and complexity between models, where larger differences indicate a stronger model and smaller differences indicate a weaker model.

Since no prior literature in this domain utilizes these tests, the most effective approach towards analysis would have to be determined after collecting data. The results from the LRT will inform if there is a reduced model that should be used for the AIC/BIC test. Then, depending on sample size and model complexity, it will become more clear if the AIC, BIC, or both should be used for analysis. Similar to the LRT, the purpose of the AIC/BIC test is to extract the most effective model for predicting personalization’s impact on churn.

**AUC-ROC Test**

The AUC-ROC (Area Under the Curve - Receiver Operating Characteristics) test evaluates how effectively the model can discriminate between predictors by providing a value
between 0 and 1. AUC measures separability while ROC is a probability curve of the AUC. An AUC of 0 indicates the worst possible predictor separation and may predict 0s as 1s and 1s as 0s; 0.5 indicates that the model does not perform better than random chance; 1 indicates that the model predicts all occurrences perfectly. ROC curves help visualize the AUC statistic by plotting the true positive and false positive rates, represented as sensitivity and specificity respectively. When looking at model curves, the strongest model is that which has the greatest area under the curve. It should be noted that this test can produce unreliable results when using poorly developed models with confounding variables, which can be improved with covariate adjustment (Hajian-Tilaki, 2013).

Depending on who is making use of the results from collected data, the threshold can be adjusted based on the desired sensitivity and specificity. Since there are 4 predictors (personalization, perceived ease of use, perceived usefulness, and user engagement), there will be 4 curves, each classifying 1 class against the other 3. The chosen model should be based on the LRT. If the model is able to discriminate between predictors based on the threshold, meaning there is little to no intersection between them, then the model accurately predicts effects on churn. Otherwise, the model may be insufficient at discriminating between predictors, and coefficients from the regression inaccurately predict personalization’s impact on churn.

**Wald Test**

The Wald Test will assess if coefficients differ significantly from 0 to determine significance. The Wald statistic is calculated by dividing the square of the coefficient estimate by its standard error, where the comparatively larger values represent a more significant impact on churn. Notably, the Wald Test will be most effective with a larger sample size and degrees of
freedom. Coefficients that involve 1 of the 4 personalization groups and provide significant p-values ($p < 0.05$) should first be identified, where the largest Wald statistic demonstrates the greatest predictability. From there, p-values and Wald statistics of interactions can be further investigated to determine moderator effects of personalization on churn. If none of the personalization groups have significant Wald statistics ($p > 0.05$), then alternative models should be tested to determine significance of coefficients.

**Mann-Whitney U Test**

The Mann-Whitney U Test will compare the median scores and user engagement rates, to see if they significantly differ between the first week (before scores) and the end of the intervention (after scores). The tested model would be determined from the combination of tests previously mentioned. A significant p-value ($p < 0.05$) provides evidence that the observations from after the first week differ from the scores after the intervention. From there, depending on the distribution of data, the mean, median, or mode of all observations for the before and after observations will be calculated to see if the after scores are greater. If both of these conditions are true, then there would be further evidence that significant coefficients from the regression impact churn with moderation from PEOU, PU, and UE. If there is no significant difference ($p < 0.05$) and/or the before scores are greater, then it may imply that PEOU, PU, and UE do not influence the impact of personalization on churn and diminish any significant interactions.

**Combining Tests to Understand Regression Results**

The mentioned statistical analyses cannot be fully understood independently, and should be interpreted together to make decisive conclusions. In terms of goodness of fit, the LRT should
be performed first to extract the most effective model, and then the AIC/BIC test will further assess its effectiveness. Once a well-fitted model is established, the AUC-ROC test will fit the predictors under curves to evaluate how distinct predictors are from one another. The Wald Test will identify the most significant coefficients that predict churn, given that the model is well-fitted and can discriminate between predictors. The Mann-Whitney U Test will determine whether or not PEOU, PU, and UE significantly differ between the before and after observations. Construction and interpretation of these 5 tests will largely depend on one another, where changes in one will likely lead to changes in the others. The chosen model should be validated by the LRT, AIC/BIC, AUC-ROC, Wald, and Mann-Whitney U tests, where inconsistent results suggest that the chosen model requires revision. Ultimately, combining these 5 tests will reveal the most effective model for performing binary logistic regression, so that the relationship between personalization and churn can be fully understood.

**Discussion**

This study seeks to build upon previous literature regarding fitness mobile apps and personalization, but with an experimental approach. Finalized participants will have active lifestyles in order to understand how apps can be developed to continuously provide value even to those that do not have detrimental physical health conditions. Primary readers would be people involved in developing fitness apps, such as researchers, marketers, developers, and product managers. Results may impact mental health experts as well if results lead to better developed apps that can be recommended to patients. In today’s world, people expect
personalization more than ever. A report states that 71% of consumers expect personalization, 76% of consumers are frustrated when there is a lack of personalization, and fast-growth companies obtain 40% more of their revenue from personalization compared to slower growing companies (Arora, Ensslen et al., 2021). Personalization is a multi-faceted concept that can be defined in numerous ways, and was chosen to be split into personalized workouts and personalized feedback to be relevant to the context of fitness apps (Rajan, 2019); 4 groups were created to identify the aspect of personalization that plays the most significant role. By performing the LRT, AIC/BIC, AUC-ROC, Wald, and Mann-Whitney U tests (and potentially other relevant unmentioned tests), the proper model should be identified.

If coefficients involving 1 of the 4 personalization groups are significant, then it implies that combinations of personalized workouts and educational content exist in predicting churn in exercise apps, as users feel special, cared for, and motivated to workout. Interaction effects involving PEOU, PU, and UE suggest that user perceptions on product novelty and loyalty influence personalization’s impact on churn in exercise apps. On the other hand, insignificant coefficients imply personalized workouts and educational content decrease user confidence and safety in using exercise apps. Insignificant interaction effects involving PEOU, PU, and UE imply that user perceptions on product novelty and loyalty do not influence the relationship between personalization and churn.

**Building Upon Previous Research**

A large hurdle is creating an app interface to create 4 distinct groups that effectively measure personalization. User adoption commonly increases in apps that consider user demographics, user characteristics, and intent in using the technology and service (Damsgaard et
al, 2007). However, one study reveals that out of 17 mobile apps tested across various studies, only 2 studies used adaptive personalization based on inputted user preferences to examine exercise behaviors (Monteiro-Guerra, Rivera-Romero et al., 2022). Furthermore, there is limited research on examining fitness activities other than walking and steps counts, obese individuals, people with healthy BMIs (body mass index), chronic illnesses, or previously active individuals (Larango, Ding et al., 2020; Hi-Park, Hwang et al., 2019). Gjestvang et al. (2022) discovered that the COVID-19 pandemic had a lesser impact on high exercise frequency individuals, which creates the opportunity to understand methods of acquiring avid exercisers as loyal users. By understanding how personalization can acquire and retain higher exercise frequency individuals, businesses can confidently acquire and retain lower exercise frequency individuals. However, there are likely characteristics of sports and exercise apps that uniquely target lower exercise frequency individuals, which should be investigated further. In terms of metrics, research has examined outcomes like user retention and engagement, but none have directly examined causes of churn. The majority of research is exploratory and does not incorporate experimental methods. Essentially, features and user attitudes that influence churn in sports and exercise mobile health apps are underexplored, and this proposal attempts to understand personalization strategies regarding workout and educational content recommendations that could impact churn.

**Unanswered Questions and Limitations**

College students in the US were chosen for practicality and to target a demographic that likely have more active lifestyles, meaning that findings may not be generalizable to other populations such as those that suffer from chronic illnesses or live in countries outside of the United States (Barkley, Lepp et al., 2019). College students also tend to have a greater
understanding of technology, which supports the study’s goals but excludes populations with less technological proficiency. Also, incorporating usability testing will likely limit the sample size to 20-125 participants, which raises considerations of implementing qualitative measures. Furthermore, data analysis does not investigate differences in participant age, gender, exercise experience, and voluntariness of use, which are key moderators in the UTAUT2 (Unified Theory of Acceptance and Use of Technology) model, and could lead to additional insights towards development. Although the UTAUT2 has seen increased use in recent years when evaluating sport and fitness apps, there is no unified approach to measuring their key constructs, especially in regards to personalization (Angosto et al. 2023). Future research could utilize the UTAUT2 model to investigate the relationship between personalization and behavioral intentions such as social facilitation, hedonic motivation, exercise identity, and price value.

The study’s design may also lead to oversights that undermine results. To start, finding a suitable app will be challenging, as it would involve either agreement from a company to test their existing app or developing a new app specifically for this study. Next, conducting the study over 5 months may be inappropriate as churn may rapidly increase after a shorter period of time (e.g. 2 months), leading to less observations at the 5 month time point. Personalization, PEOU, PU, and UE may impact churn differently at varying points in time depending on user behaviors and goals (Kaveladze, Wasil et al., 2022). Sample size may be another major constraint. Tong et al. (2022) developed their own low-fidelity app and evaluated suggestions with clinicians and researchers before providing them to users, thus a smaller sample size ($n = 26$) was necessary to undergo this time intensive process. It will be difficult to comprehensively assess personalization’s impact on churn without time series analysis and an inadequate sample size for usability testing.
Another consideration is developing an accurate recommender system that is adequate for the study’s needs. System-based data collection enables constant updates on user preferences and adjustment to activity, but requires properly trained AI models. This proposed study may overlook significant variables that weaken the personalization model’s ability to predict churn. For instance, gamification is another variable of interest that may have significant outcomes on churn by increasing user information technology (IT) identity, leading to continued usage of fitness apps (Esmailzadeh, 20201). Fitness app usage could also be dependent on exercise identity (Junaeus, S., 2015) or measures like self-efficacy, social support, and autonomous motivation (Peterson, Kemps et al., 2020). Furthermore, the personalization will require the AI to obtain enough information from users over time to provide the most accurate recommendations. Therefore, the difference scores for PEOU and PU may be skewed in favor of the scores after intervention as one week of usage provides substantially less accurate recommendations. Usage amount on a daily basis may also play a role on PEOU and PU scores, especially after the first week, so measuring the time in minutes of app usage could be considered as well. Moreover, privacy is a common concern for mobile health app users (Guo et al., 2022), and there is much debate on this topic in regards to AI. Therefore, creating an effective personalization strategy requires a deep understanding of user preferences, identities, and motivations, so that features promote continued usage.

UE is a difficult concept to measure, as there are a myriad of influential factors. The application, setting, and variables of interest largely influence how UE is defined (Taki, Lymer et al., 2017). One study utilizes the User Engagement Survey (UES) to discover that focused attention, perceived usability, aesthetic appeal, and reward are major contributors towards maintaining engagement (Holdener, Gut et al., 2020), which should be investigated further.
Another study finds that smartphone engagement (and consequently app engagement) increases when functional, hedonic, and social features support user motivation (Kim et al., 2013). Furthermore, engagement is a heterogeneous concept, meaning patterns like frequency and intensity differ between users, and decreased use over time could be an indication that users have achieved their goals in using the app and no longer need it to improve user wellbeing (Kaveladze, Wasil et al., 2022; Amagai, Pila et al., 2022). If churn will happen regardless of an app’s quality, marketing and design strategies should be utilized to reactivate previous users.

UE as a categorical variable allows a more straightforward interpretation of the metric and lowers the complexity of the model, but does not account for specific contributors, qualitative measures, differences at various time points, and differences that remain in the same UE category. To illustrate, it may be the case where a participant’s UE largely increases from 51% after the first week to 73% after 5 months, but are categorized as medium engagement for both (>50%) and disregard the jump in engagement. Another scenario is that a participant’s UE slightly increases from 49% to 51%, which would be observed as a shift from low engagement (≤ 50%) to medium engagement despite negligible improvement. Furthermore, UE changes in participant goals would shift participant engagement, which could be analyzed by attributing goal preferences with UE. Overall, maintaining engagement is extremely challenging for companies, and understanding it can greatly help marketers and developers.

**Real-world application**

Companies should prioritize incorporating personalization in their products, seeing how it often increases revenue by 10-15% (Arora, Ensslen et al., 2021). However, companies should be aware that participants are more likely to churn than to be retained, regardless of the app’s
capabilities (Amagai, Pila et al., 2022). The goal is to understand the best personalization approach to mitigate the amount of churn that occurs. Groups 2 (Workouts, No Educational Content) and 3 (No Workouts, Educational Content) are tested to inform app developers if there are specific areas of personalization that should be prioritized from the start. For example, if recommending only workouts has greater predictive power of churn than only recommending educational content, then personalized workouts should be prioritized over personalized educational content. In the same scenario, if group 1 (Workouts, Educational Content) has higher predictive power than groups 2 and 3, have lower standard errors, and/or higher z-scores, then there is further evidence towards developing personalized workouts first.

Moderator variables largely inform areas that companies should evaluate when obtaining feedback from users. Significant moderator effects in PEOU, PU, and/or UE should be emphasized throughout development of the app. PEOU, PU, and UE help inform users’ attitudes towards apps, validate or invalidate feature and quality of life updates, and ensure that development stays on track in designing an exceptional experience.

Even if personalization significantly predicts churn, future research should consider additional strategies to fully understand the relationship between personalization and churn. Although this study investigates personalized workouts and educational content, it can provide app developers a perspective towards applying personalization. Furthermore, personalization requires extensive development and is no small feat, so companies need to be confident that it is a smart allocation of resources, especially if an app is developed from scratch. At the end of the day, personalization benefits both users and businesses, and this study should provide insights towards developing mobile health apps to retain the most users possible. Personalization through
personalized workouts and educational content may be prioritized with significant findings, whereas an alternative approach to personalization should be considered with insignificant findings.

References


mediation analysis of the role of self-efficacy and barriers. *Journal of Medical Internet Research, 17*(8). [https://doi.org/10.2196/jmir.4142](https://doi.org/10.2196/jmir.4142)


Holdener, M., Gut, A., & Angerer, A. (2020). Applicability of the User Engagement Scale to Mobile Health: A Survey-Based Quantitative Study. *JMIR mHealth and uHealth, 8*(1), e13244. [https://doi.org/10.2196/13244](https://doi.org/10.2196/13244)


Alley, S., Jennings, C., Plotnikoff, R. C., & Vandelanotte, C. (2016). Web-based video-coaching to assist an automated computer-tailored physical activity intervention for inactive adults: A randomized controlled trial. *Journal of Medical Internet Research, 18*(8). [https://doi.org/10.2196/jmir.5664](https://doi.org/10.2196/jmir.5664)


Vaghefi, I., & Tulu, B. (2019). The Continued Use of Mobile Health Apps: Insights From a Longitudinal Study. *JMIR mHealth and uHealth, 7*(8), e12983. https://doi.org/10.2196/12983


engagement in eHealth and mHealth behavior change interventions: Viewpoint of methodologies. *Journal of Medical Internet Research, 20*(11). [https://doi.org/10.2196/jmir.9397](https://doi.org/10.2196/jmir.9397)


Holdener, M., Gut, A., & Angerer, A. (2020). Applicability of the User Engagement Scale to Mobile Health: A Survey-Based Quantitative Study. *JMR mHealth and uHealth, 8*(1), e13244. [https://doi.org/10.2196/13244](https://doi.org/10.2196/13244)


https://doi.org/10.2196/30766

https://doi.org/10.4258/hir.2019.25.1.12