university of groningen

faculty of science and engineering

Tailoring Motivational Features to the Preference of Users in a Habit App with Self-Chosen Activities and Goals

Sönke Steffen

university of groningen / faculty of science and engineering

University of Groningen

Tailoring Motivational Features to the Preference of Users in a Habit App with Self-Chosen Activities and Goals

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Human-Machine Communication
at University of Groningen under the supervision of
Prof. dr. F.C. Cnossen (Artificial Intelligence, University of Groningen)

Sönke Steffen (S3846504)

August 31, 2021

Contents

Abstract

In recent years, smartphone apps intended to aid habit formation have increasingly become commonplace. A frequent criticism of these habit apps has been the lack of theoretical framework explaining which motivational mechanisms make them effective. Research in fields like Gamification and Behavior Change Theory is concerned with describing different motivational features implemented in those apps and their effects on motivation and habit retention. However, it has repeatedly been criticized that research mostly focuses on the domains of health-related behaviors and education, with other domains being largely neglected. Further, individual characteristics of different users, as well as the effects of intrinsic and extrinsic motivation are not sufficiently taken into account. For this research, a habit app was developed which gives users freedom to pursue any habit of their interest. Different motivational features were implemented which are compatible with all kinds of habits. This allows the app to serve as a framework for investigating a broader amount of domains in a more controlled setting. An experiment was conducted to check whether tailoring the motivational features to user preferences based on preference-ratings and user behavior improves habit formation. While technical issues and a lack of participants meant that no significant effects of tailoring were found, the two measures were indeed related and might therefore indicate user preferences. Further, the app itself was found to effectively aid habit formation for many different types of habits. In summary, this research is considered a step towards generalized habit research and has shown that more domains must be considered. It has also indicated the need and suitability of dynamical tailoring mechanisms.

1 Introduction

Over the past years, a large amount of smartphone apps targeting habit formation and changes of behavior have been pushed onto the market, promising to guide users towards their goal by applying different motivational strategies (Villalobos-Zúñiga & Cherubini, 2020). Motivational features like achievement badges or feedback charts are supposed to give users incentives to stick to their goal. Despite those promises, it has repeatedly been found that these apps and their features are often not based on solid research, highlighting the need to develop a consistent theoretical framework (Schmidt-Kraepelin, Toussaint, Thiebes, Hamari, & Sunyaev, 2020; McKay, Wright, Shill, Stephens, & Uccellini, 2019; Stawarz, Cox, & Blandford, 2015). Multiple fields of research have investigated the motivational mechanisms behind habit apps, how, when and whether they improve habit formation, as well as which motivational features work best. First, the field of Gamification attempts to describe design elements like the aforementioned achievements, which originate in computer games, extracting their motivational aspects and their efficiency (Sailer, Hense, Mayr, & Mandl, 2017; Koivisto & Hamari, 2019). One theory often used to explain the motivational effects of these features is Self-Determination Theory (SDT) (R. M. Ryan, Rigby, & Przybylski, 2006; Villalobos-Zúñiga & Cherubini, 2020). This conveys the effects of different types of motivation on task performance, mainly focusing on different levels of extrinsic vs. intrinsic motivation (R. M. Ryan & Deci, 2000). Intrinsic motivation refers to any inherent aspects of the task motivating the user. For an intrinsically motivated user, the task itself is enjoyment enough to perform it without hesitancy. Extrinsic motivation on the other hand refers to any motivation that is not driven by the contents of the task itself, as is the case when being paid to do an unappealing task. Additionally to Gamification and SDT, Behavior Change Theory (BCT) results from the perspective of clinical interventions into undesired and desired behaviors. BCT has been used to describe and implement behavior change methods, similar to the design elements in Gamification (Lustria et al., 2013; Michie et al., 2011).
Besides the lack of an overarching theoretical framework, it has also been criticised that habit and Gamification research mainly focuses on only a few distinct domains, namely health and education (Koivisto & Hamari, 2019). There is a need to extend the domains considered, combined with investigating differences in which motivational features work better or worse for which domains (Klock, Gasparini, Pimenta, & Hamari, 2020). Additionally to differences between domains, individual differences between users play a role in the effectiveness of interventions and different motivational features (Hamari, Koivisto, & Sarsa, 2014). A lack of personalization based on these individual differences has been criticized for a majority of apps (Klock et al., 2020; Koivisto & Hamari, 2019). Tailoring certain aspects of the intervention to different user characteristics or user preferences promises to improve its effectiveness (Klock et al., 2020; Hamari et al., 2014; Lustria et al., 2013). It can further help mitigating negative effects that have been observed for some generally promising motivational features on a sizeable minority of users.
The following research question result from the issues described and was tackled in this research:

> Can dynamically tailoring different motivational features to user preferences improve the efficiency of an app supporting habit formation?

To address this question, a habit tracking app for Android was developed, called Routinary. This provided users with freedom to pursue any kind of habit with any measure and any daily goal and track their progress. Motivational features were implemented in the form of motivational cards in the app, based on different types of design elements, like achievements or statistical charts. The content of the cards is also used in motivational notifications, which are shown to the user on a daily basis. All features implemented are intended to work regardless of the chosen habit and goal. The

use of those motivational features in push notifications and on the home screen were tailored to the individual users based on two measures: The users rating all motivational features and the amount of in-app interactions of users with the different cards. A 2x2 factorial design was used, with both tailoring methods either individually applied, their ratings being combined, or the tailoring being fully random.

The paper is structured as follows. In Section 2, the theoretical background relevant to this paper is presented. The relevancy of habits in smartphone apps is presented, including how different implemented motivational features can support the formation of habits. Additionally, the research field of Gamification and its work relating motivation in apps, as well as Behavior Change Techniques (BCTs) from the domain of Behavior Change Methods are introduced. Finally, the concept of tailoring different motivational features to user characteristics is explained. Based on the presented theory, Section 3 introduces the goals of the research, as well as the research design used for the experiment including the hypotheses.

The Routinary App developed as part of this research this is presented in Section 4. The process of using Routinary to input and track habits is introduced by providing a short walkthrough and explaining the layout of the app. Motivational cards, which serve as the implementation of motivational features within the app are explained, as well as motivational push notifications which are presented to users. Section 5 presents the methods that were used to utilize the Routinary app for the experiment. This includes the concrete mechanisms used for the tailoring of motivational content to user preferences. Additionally, the procedure used in the experiment, as well as the statistical analyses used are explained.

The results of the experiment are presented in Section 6. They include the results from the experiment itself, but also further post-hoc analyses. The results are discussed in Section 7, including whether the tailoring worked, the app increased user success and how diverse the activities were users followed. Based on these discussions, five recommendations for future research are made. A conclusion is drawn in Section 8, including further recommendations.

Finally, four appendices supplement the paper. Appendix A includes multiple screenshots to depict the layout of the Routinary App more extensively. Appendix B provides a step-by-step overview of the setup process participants went through to use Routinary. Appendix C explains and shows all motivational cards that are implemented within the Routinary app. Appendix D provides supplemental results to the ones presented in the paper.

2 Theoretical Background

The use of smartphones has drastically increased over the past years, from about 3.7 billion users in 2016 to more than 6.3 billion users in the year 2020, with the number expected to rise (Statista, 2021). Subsequently, smartphone apps have been developed for all kinds of applications that support users in their daily lives. One important aspect of these apps is habit formation (Oulasvirta, Rattenbury, Ma, & Raita, 2012). A habit describes an action that is performed repeatedly and in an automated fashion as a response to certain situations, without much planning required (Lally & Gardner, 2013). Habits can be desired as well as undesired. An example for a desired habit would be the brushing of teeth right before going to bed at night. An undesired habit for some people would be grabbing a bag of chips when entering the kitchen in the evening. Habit formation describes the process of repeatedly performing a behavior in these contexts, leading to the behavior becoming automated, or habituated. This usually requires planning behaviors to be executed repeatedly until the habit is formed. In case of brushing teeth, this is mostly done during childhood by parents reminding their children to brush their teeth until the habit has formed.

Dedicated smartphone apps have been developed to aid the formation of a desired habit regarding different types of activities, with examples shown in Figure 1. Activity tracking apps like Samsung Health support users in building healthy habits like walking or doing sports regularly by allowing users to track their progress and set goals (Samsung, 2021). Further, apps not purely dedicated to habit formation often include aspects of habit formation to increase usage of the app. The app Duolingo, which is centered around helping users to learn a language includes features to support users to form a regular habit of using the app to learn the language on a daily basis (Duolingo, 2021).

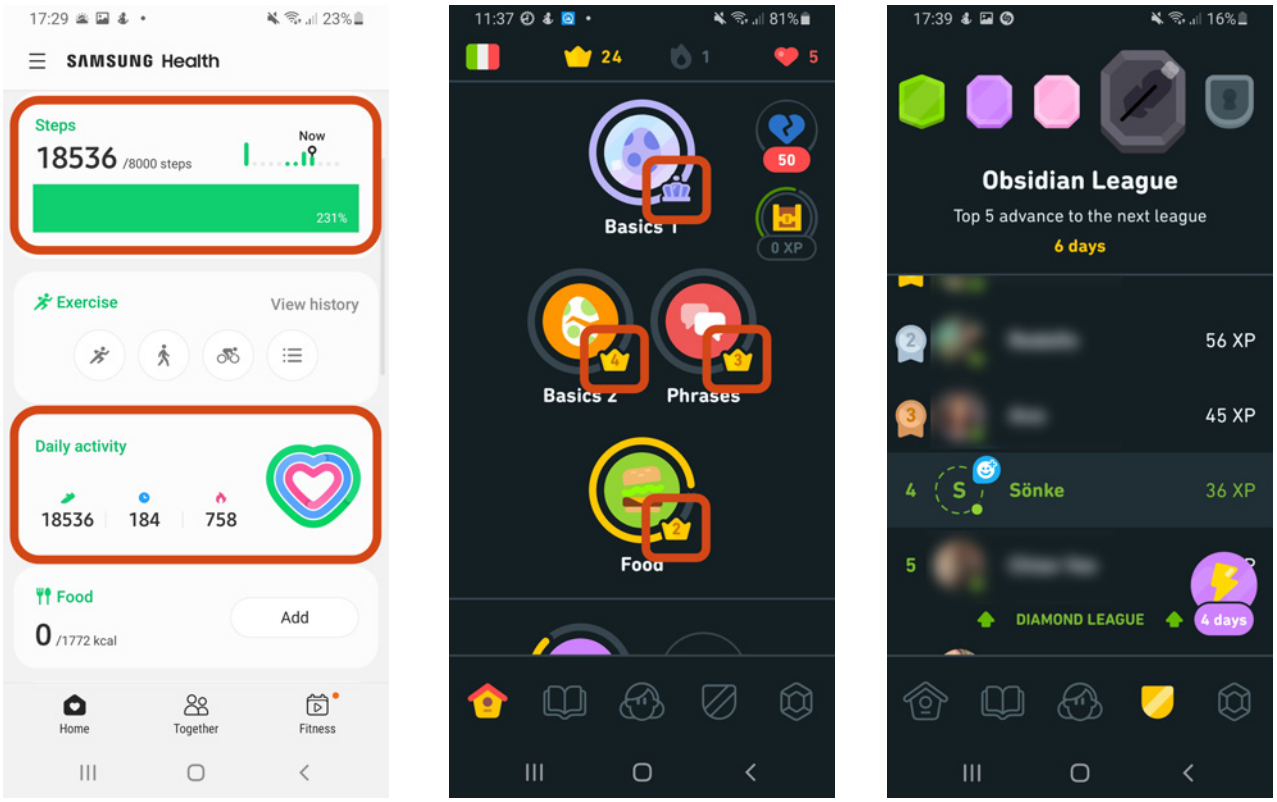
To support habit formation, features are implemented in apps with the goal of motivating users to perform their activity on a regular basis, ultimately building a habit. These features will be called motivational features in this paper, but depending on the domain of research have been termed motivational affordances, game design features or Behavior Change Techniques (Hamari et al., 2014; Sailer et al., 2017; Michie et al., 2013). Motivational features are features which are supposed to increase the motivation of users to perform an activity in some way. A large variety of different motivational features have been developed, described and investigated (Villalobos-Zúñiga & Cherubini, 2020; Michie et al., 2013). For example, Samsung Health provides activity feedback to the user by including a step tracker and allows users to relate their daily activity to a daily goal that the user can set, as highlighted in Figure 1a. Further, Duolingo provides ”crowns” highlighted in Figure 1b for achieving different levels on understanding on a language lesson, artificially rewarding users for their effort. It also includes leaderboards where experience points gained through practicing a language are compared to the ones of other users, shown in Figure 1c. By achieving a high position, users can then advance to a higher league.

Different domains of research have investigated motivational features, including how they work, how effectively they support motivation and habit formation, as well as which motivational features are most promising. This section provides a non-exhaustive look at the fields relating to habit formation in apps. First, research relating to Gamification will be outlined. Gamification describes the process of implementing features in services like apps that have a motivating or otherwise desirable effect on the user (Koivisto & Hamari, 2019). Most of the research directly or indirectly investigating habit formation in apps relates to Gamification, with motivation being the main construct investigated. The process of Gamification, as well as the current state including issues with Gamification are presented. Gamification will also be related to Self-Determination Theory (SDT). SDT describes individual differences in motivation and is often used to explain effects found in Gamification approaches (Villalobos-Zúñiga & Cherubini, 2020).

Figure 1

Examples of Motivational Features used in Habit Formation in Apps

(a) *Samsung Health* **(b)** *Duolingo* **(c)** *Duolingo Leaderboard*



Notes. Examples of motivational features implemented in established commercial apps. The highlighted motivational features are described in the main text. **(a)** Screenshot of the Samsung Health app (Samsung, 2021). The feature shown on the top is a step counter, visualizing the number of steps relative to the user goal. The bottom one gives an overview of the general daily activity, also relative to the user goal. **(b)** Screenshot of the language learning app Duolingo (Duolingo, 2021). Highlighted are so-called crowns which act as a reward indicating to the user how advanced their abilities regarding a topic is. **(c)** Duolingo also provides a leaderboard, providing social comparison. If enough experience points are gained, the user can advance to higher leagues.

Another approach used to describe and investigate aspects of habit formation apps is drawn from research relating Behavior Change Interventions (BCIs). These BCIs are mostly used to target behavior change in regards to undesirable and desirable health-related behaviors, trying to build healthier habits (Lustria et al., 2013). This research predates computer-assisted therapies or interventions, but has been applied to investigate and design motivational features in smartphone apps (Stawarz et al., 2015; McKay et al., 2019). Therefore, research into Behavioral Change Techniques (BCTs) will be introduced in this section. BCTs are techniques to assist behavioral change in therapies and can be loosely compared to the motivational features described in Gamification research (Michie et al., 2011). It will be presented how these are used in habit apps and which issues have arisen to provide another perspective additional to the one provided by Gamification.

One of the main issues found in habit app research was that not all motivational features work equally well for all users (Hakulinen, Auvinen, & Korhonen, 2015). A proposed and used solution for this is

the approach of tailoring (Klock et al., 2020). This describes adjusting which content is presented in motivational features and how it is presented, based on user preferences (Hawkins, Kreuter, Resnicow, Fishbein, & Dijkstra, 2008). Different methods of tailoring and their current state in habit research are presented at the end of this section.

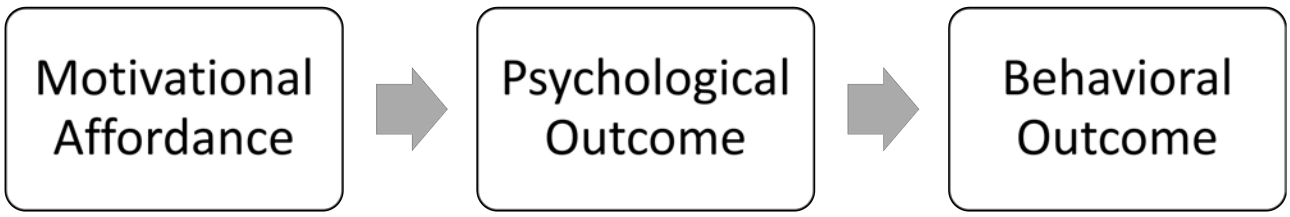
2.1 Gamification

While there is no clear consensus about the definition and conceptualization of Gamification, it is generally considered to be the design of an information system that leads to the same experiences as games do, usually with the same or similar methods to the ones used in game design (Koivisto & Hamari, 2019). This can be done intentionally, by implementing elements from game design into a service or product that is typically not considered to be a game with the goal of achieving similar psychological or behavioral outcomes, called Intentional Gamification (Hamari, 2007). It can also be done unintentionally, due to the societal or cultural impact of games leading to aspects and principles of games inadvertently affecting conventions outside game design, which has been termed Emergent Gamification. Different psychological outcomes can be targeted with Intentional Gamification, but the main one can considered to be motivation. This relates to the users motivation to continue using the service, or improve on a behavior that the service is monitoring, as has been presented in the introduction to this section. Features with this goal fall under the category of motivational features. While Gamification techniques are in practice used in a variety of applications, it has been pointed out that the majority of Gamification research focuses on the domains of learning, education and health tracking (Koivisto & Hamari, 2019). Therefore, expanding research to include a broader variety of domains has been suggested (Klock et al., 2020; Koivisto & Hamari, 2019).

2.1.1 Underlying Process

Figure 2

Gamification Process



Notes. Process underlying the effects of Gamification, as described by Hamari. Figure created based on (Hamari et al., 2014).

The process of Gamification consists of three components, visualized in Figure 2: A user is exposed to a motivational affordance which leads to a psychological outcome which subsequently leads to a measureable behavioral outcome (Hamari et al., 2014). Motivational affordances describe the characteristics of an object which determine the psychological outcome caused in the user depending on the meaningfulness in its perceived context (Zhang, 2008). In other parts of the literature these characteristics have been described as the game design elements, but they can be interpreted as conceptually identical (Sailer et al., 2017). Some examples of commonly used motivational affordances or game

design elements in Gamification include leaderboards, achievements or levels (Hamari et al., 2014). The psychological outcomes following the interaction with motivational affordances describe psychological experiences which can be observed in, and reported by, the user (Koivisto & Hamari, 2019). The concept of psychological outcomes can be interpreted and measured in a manifold of ways, with typical reported constructs being motivation, entertainment, competence and attitude (Hamari et al., 2014; R. M. Ryan et al., 2006). For example, being in third place on a leaderboard could trigger a user's motivation to perform well enough to gain second or first place. An example for a negative psychological outcome would be the lack of perceived competence when observing that one is at the bottom of a leaderboard. Finally, additionally to the psychological outcome, an accompanying behavioral outcome can and might occur, which can be measured depending on the goal and application that the Gamification has been designed for. In case of the third position on the leaderboard this might be an improved performance to achieve first place, in case of the bottom position and the lack of perceived competence an abandonment of the app might follow.

Although there is no standardized way to apply Gamification, attempts have been made to describe different types or schemes of Gamification used in practice, based on mental health apps. (Schmidt-Kraepelin et al., 2020).

2.1.2 Current State

Even though Gamification has generally been perceived as promising, reviews have emphasized a large amount of mixed results on the effectiveness of Gamification methods on motivation (Koivisto & Hamari, 2019; Hamari et al., 2014). This is apparent both when comparing the results of different studies, but also when comparing different aspects of Gamification within studies. It has even been suggested that the reported results might be skewed positively due to publication bias (Koivisto & Hamari, 2019). Multiple explanations for the mixed results exist. It has been found that the same Gamification aspects which had a positive effect on the majority of users still often had a sizeable minority for whom negative effects were observed (Hamari et al., 2014). Therefore, it was concluded that effects of different aspects in Gamification depend on the context and users, with no method fitting all users and applications equally well. This is in line with a recent meta-analysis of Gamification methods (Koivisto & Hamari, 2019). Multiple methodological and theoretical agenda points were suggested by the authors, considered to be worthy of investigation in future research. The authors found that individual differences between the user's goals and individual attributes were relevant to determine how different aspects of Gamification are perceived by the users. The analysis also found that Gamification approaches might often be too structured to account for individual preferences of users. For example, predetermined goals that do not match the user's personal goals, abilities or expectations might either overwhelm or underchallenge the user and therefore lead to counterproductive outcomes. Therefore, it was suggested to consider approaches that are less restrictive and give the user more freedom for a creative use of the Gamification-based intervention, for example by letting them determine their own goal. Finally, they emphasize the need to develop more systematic and consistent methods and instruments to investigate Gamification, allowing for more controlled experimental research.

One proposed and investigated method to account for individual differences in regards to the effects of motivational features is the method of tailoring, in which a motivational feature is adjusted to user characteristics or attributes. This has been explored in Gamification already, showing promising results (Klock et al., 2020). More details about tailoring can be found in the section on tailoring.

2.2 Self-Determination Theory

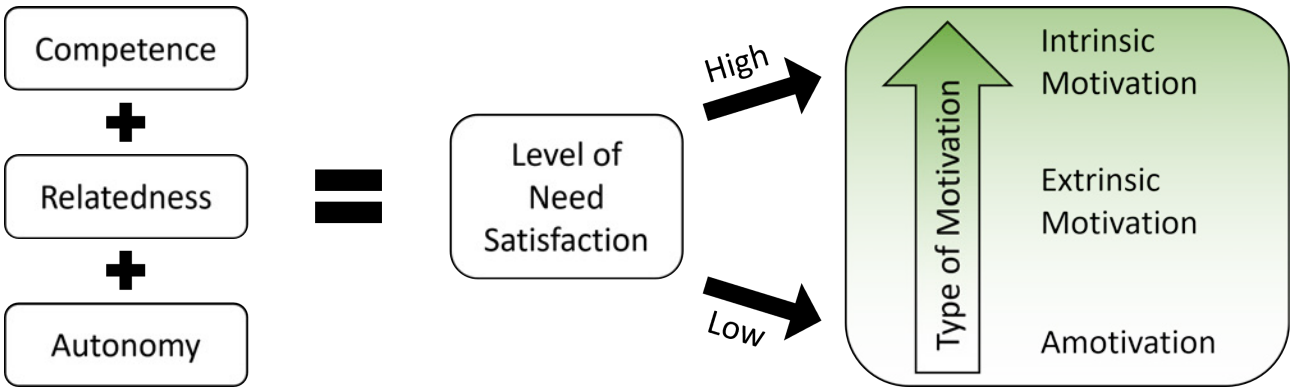
The motivational effects of motivational features in Gamification have often been described and explained using Self-Determination Theory (Villalobos-Zúñiga & Cherubini, 2020; R. M. Ryan et al., 2006). Since its inception in the 1970s, the Self-Determination Theory (SDT) has been one of the main frameworks to explain the nature of human motivation and has been continuously developed and extended since then (R. M. Ryan & Deci, 2000). A basic visualization behind the framework is depicted in Figure 3. The SDT conceptualizes the motivation to perform a task or pursue a goal by defining different motivational states, ranging from intrinsic motivation via different levels of extrinsic motivation to amotivation. While amotivation describes the total lack of personal intention to act with the behavior fully being driven by external forces, intrinsic motivation describes the inherent and unprompted pursuit of a goal at the absence of any external incentives or rewards. In between these two types of motivation extrinsic motivation is located. This describes types of motivation which to some degree are caused by external incentives, with different levels of internalization present. While attempting to determine the positive and negative effects on motivation, SDT research has identified three basic psychological needs: Competence, relatedness and autonomy. Competence relates to the feeling of being equipped with the ability to successfully achieve a desired outcome (Connell & Wellborn, 1991). Relatedness describes the need to be securely embedded and valued in the social environment. Finally, autonomy comprises the feeling of being able to personally initiate actions and pursue goals, based on one's own personal values. These psychological needs are interpreted as basic to human nature, in that personal growth can only follow when they are satisfied, also linking them to a closer state of intrinsic motivation (R. M. Ryan & Deci, 2000). In contrast, extrinsic motivation and its regulatory styles follow as a consequence of unsatisfied psychological needs in relation to a pursued goal.
Interpreting psychological needs as innate implies that any individual differences between needs are a consequence of a lack of satisfaction, instead of how strongly the need has developed in the first place (Deci & Ryan, 2008). Therefore, interventions based on SDT have mostly targeted the satisfaction of needs to correct for any negative motivational states. This assumption has recently been challenged. For example, in a meta-analysis it was found that an absence of satisfaction of needs was only weakly linked to negative motivational outcomes (Van den Broeck, Ferris, Chang, & Rosen, 2016). The authors concluded that while the satisfaction of needs accounted well for the level of positive outcomes, negative outcomes must separately be caused by some kind of need frustration, also termed thwarting. Therefore, the SDT has been extended by the implementation of need frustration in measurements like the Basic Psychological Needs and Frustration Scales (BPNFS) (B. Chen et al., 2015).

2.2.1 SDT Approaches to Gamification

The SDT has often been proposed and utilized as a theoretical framework and underpinning to explain the motivation behind Gamification and habit apps, further providing insights into avenues of improvement (Hosseinpour & Terlutter, 2019; Sailer et al., 2017; R. M. Ryan et al., 2006). As mentioned, multiple studies have found that Gamification itself does not automatically lead to increased motivation. They further concluded that the effectiveness of Gamification methods depends on the types of motivational features used, as well as the context in which they are used and the characteristics of the users (Sailer et al., 2017; R. M. Ryan et al., 2006). It was for example found that motivational features like badges or achievements support the feeling of competence, while features like teammates or avatars supported the feeling of social relatedness (Sailer et al., 2017). Therefore, the respective motivational features would be more effective with users who had a need for com-

Figure 3

Principles of Self-Determination Theory



Notes. The ideas behind Self-Determination Theory. The basic psychological needs that must be satisfied in regards to pursuing a task are competence, relatedness and autonomy. The overall level of need satisfaction then determines which type of motivation develops. At high levels of need satisfaction, this equates to intrinsic motivation, meaning the behavior is self-determined. At very low levels it leads to amotivation. In between are different levels of extrinsic motivation.

petence or a need for social relatedness. Further disentangling the relations of different aspects of Gamification, the context they are used in and different user types would also aid in explaining the mixed results found in the general effectiveness of Gamification, mentioned in the previous section.

Recently, a taxonomy of motivational app features often found in habit or Gamification apps has been suggested, relating the respective features to the three psychological needs established in the SDT (Villalobos-Zúñiga & Cherubini, 2020). It was suggested that different features support different psychological needs, for example: Goal setting and motivational messages support the need for autonomy, while activity logs, self-monitoring features or rewards support competence. Finally, features like peers comparisons or performance sharing were found to support the need for relatedness. The authors found that an insufficient number of apps had features which supported all three basic needs, suggesting that this should be attended to in the motivational design of habit apps. This matches the recommendations resulting from similar investigations (Molina & Sundar, 2018; Asimakopoulos, Asimakopoulos, & Spillers, 2017). Further, the features themselves should be tailored according to the SDT. For example, users with a higher, or unsatisfied need for competence could be aided by achievements being shown, as those indicate the fulfillment of a requirement for something achievement-worthy. Additional recommendations in the papers were trying to optimize the challenges to the level of performance of the user, further satisfying the psychological needs, as well as providing personalized challenges. Similarly, in a study on motivation and user engagement in fitness tracking, it has been proposed to foster the motivation of users by giving choice to them, thereby abandoning any unnecessary restrictions (Asimakopoulos et al., 2017). For example, instead of providing a predetermined goal for all users, they would be allowed to set their own desired goals themselves. This in itself would support the need for autonomy. Being able to further adjust other features of the app gives the user the opportunity to maximize gains for all three psychological needs.

Finally, it has been well-established that the use of external rewards to foster extrinsic motivation can also undermine the more desirable intrinsic motivation (R. M. Ryan & Deci, 2000). The effects of this are depicted in a study on the motivational costs of wearing fitness trackers (Attig & Franke, 2019).

It showed that users with a tendency towards more extrinsic motivation exhibited a larger decrease in motivation when the tracker was not available anymore than users who scored higher on intrinsic motivation and therefore were less dependent on the tracking mechanism. Therefore, multiple authors have suggested that apps should not only promote extrinsic motivation, but also support the transitioning from extrinsic to intrinsic motivation. This aids in securing the long-term establishment of the habit or behavior change intrinsically, as would be preferable according to the SDT, without having to rely on the app to provide extrinsic motivation (Villalobos-Zúñiga & Cherubini, 2020; Attig & Franke, 2019).

2.3 Behavior Change Techniques

Behavior Change Interventions (BCIs) originate in the attempt to structurally investigate behavior change therapies for mostly health related behaviors like smoking cessation, physical activity or changes in nutrition/diet (Lustria et al., 2013). Within the Behavior Change Interventions, Behavior Change Techniques (BCTs) are the components of the intervention which have been designed to aid in changing the behavior, focusing on what content is delivered rather than how it is delivered (Michie et al., 2011). Behavior Change Techniques can loosely be compared to the motivational features described in Gamification research, although those traditionally tend to purely focus on motivational aspects.
Due to the fractured development of interventions and their respective techniques, multiple attempts have been made to describe and taxonomize the Behavior Change Techniques (BCT) interdisciplinarily (Lustria et al., 2013; Abraham & Michie, 2008). Arguably, the most established taxonomy is the Behavior Change Technique Taxonomy (BCTTv1) (Michie et al., 2013). The goal in the creation of the BCTTv1 was to find, describe and taxonomize BCTs which could be clearly defined, are unambiguous and are distinct without any overlap between different BCTs. After extracting a preliminary list of BCTs based on existing BCT classification systems, these were synthesized and ultimately taxonomized based on expert ratings. Despite their different origin, the extracted BCTs have a high overlap with the motivational features typically found in Gamification research. For example, the techniques of goal setting, providing rewards, or providing activity feedback can be found in both approaches (Villalobos-Zúñiga & Cherubini, 2020; Michie et al., 2013).

2.3.1 BCTs in Behavior Change and Habit Apps

The BCTTv1 and similar taxonomies have successfully been used to describe features for the design in smartphone apps relating to behavioral change and habit building in multiple applications (McKay et al., 2019; Hosseinpour & Terlutter, 2019; Stawarz et al., 2015; Conroy, Yang, & Maher, 2014). It has been commonly criticized that most commercial apps did not implement a sufficiently large number of BCTs and were additionally lacking a formal framework followed in the design of the app (Hosseinpour & Terlutter, 2019; Stawarz et al., 2015; Conroy et al., 2014). This might be responsible for the limited success of these apps, as app developers do not cover all relevant aspects for designing their intervention. Within research, researchers have additionally pointed out difficulties in comparing different interventions with regards to their effectiveness and contents due to the lack of a clear methodological framework and the focus on qualitative studies (Lustria et al., 2013). Most of these methodological issues have also been described within the domain of Gamification research.
Analyzing the effectiveness of BCTs in behavior change and habit apps has generally led to mixed results. In a meta-analysis on the effectiveness of BCTs in motivational apps to increase physical activity, all investigated BCTs were linked to supportive and contradictory findings (Hosseinpour &

Terlutter, 2019). Still, some BCTs were found to be more effective than others, with feedback, goal setting, competition and social sharing with familiar users being considered effective in a majority of the studies examined. Other features, however, resulted in contradictory and inconclusive findings. For example, the Reward-BCT lead to mixed results, with different studies showing both positive and negative effects on physical activity in both quantitative and qualitative studies. The authors suggest that the exact reasons for why and when rewards work need to be explored. This is in line with research on the effect of different achievement badges on the learning progress of students, with achievement badges being an implementation of the Reward-BCT (Hakulinen et al., 2015). While some achievement badges were indeed found to have a positive effect on the learning performance, others either had no effect, or might have even had a negative effect and impeded on learning progress. This indicates that different implementations of the same BCT can differ in their effect on the intervention outcome. The authors therefore suggest further research into individual differences regarding the effects of achievement badges, as well as the development of dynamical adaptation algorithms. This recommendation matches the conclusions drawn by Hosseinpour and Terlutter, who suggest that the generally mixed results about the effectiveness of different behavior change apps might not only be a consequence of the usage of different BCTs, but also their respective forms of implementation (Hosseinpour & Terlutter, 2019).
In summary, research on BCTs in habit and behavior change apps has shown that BCT taxonomies are promising in both describing and prescribing principles used in behavior change apps. However, further research improving the theoretical framework, as well as research into differences between BCTs and their respective forms of implementation is necessary. Finally, the influences of individual differences should be explored, allowing for further dynamical adjustments to user characteristics, called tailoring, which will be described in the next section.

2.4 Tailoring

Tailoring describes the process of adjusting communication to characteristics of the individual user with the goal of increasing and maximizing the impact of the message by increasing the relevancy of the content (Hawkins et al., 2008). This can be interpreted as the highest level on a continuum of customizing the message to users, with the lowest being the design of the message towards the whole audience and degrees of customizing towards certain groups as strategies within. However, there is no consensus about the exact definition of tailoring and the categories represented on the continuum. For alternate definitions or conceptualizations of tailoring, see (K. Ryan, Dockray, & Linehan, 2019; Noar, Grant Harrington, Van Stee, & Shemanski Aldrich, 2011).

2.4.1 Methods of Tailoring

Different methods and strategies of tailoring exist, which can broadly be conceptualized as personalization, feedback and content matching (Hawkins et al., 2008). In a personalized message, the content is adjusted in a way which tries to give the impression that the communication is specifically tailored to the user. Personalization is interpreted as the most superficial form of tailoring, not conveying any deep relationships between user characteristics and tailoring. One example for this is the mentioning of the user's name in a message, which can increase effortful processing due to its perceived personal relevance and can under certain circumstances increase motivation (Dijkstra, 2014). Next, Feedback is a commonly used strategy in which the user is provided with information relevant to themselves. This can be as simple as descriptive feedback, where the user is provided with information in regards to him- or herself, for example the current performance on a behavior, or it can be more complex,

like an evaluative feedback, putting the performance into context by comparing it to the user's goal. Finally, content matching describes the tailoring of the message by evaluating different factors that have been found to influence the behavior of interest, addressing those in the content of the message. This type of tailoring has recently become more commonplace and complex due to advancements in automatized data-driven decision making. Algorithmic computer-based tailorings towards the characteristics of each individual has been termed computer-tailoring (Krebs, Prochaska, & Rossi, 2010). It is important to mention that all previously mentioned tailoring methods can be used in combination with each other and no clear-cut distinction between the concepts exists.

2.4.2 Dynamic vs. Static Tailoring

There are an abundance of measures based on different assessment methods available to use for tailoring. Examples are sociodemographic variables being assessed via theory-driven questionnaires, or measures of behavioral performance as indicated by sensors or by manually entering the data into an app (van Velsen, Broekhuis, Jansen-Kosterink, & op den Akker, 2019; K. Ryan et al., 2019; Elbert, Dijkstra, & Oenema, 2016). An important distinction in the procedure of tailoring can be made between dynamic and static tailoring (Krebs et al., 2010). In static tailoring, the metric is measured only once before the intervention. The result is then used as a basis for all tailorings or feedbacks. However, users might change over the course of the intervention, including their attributes relating to the optimal tailoring. This means that the tailoring would become outdated and not lead to optimal or improved outcomes and might even diminish the outcome. An alternative is therefore dynamic tailoring. This describes the continuous adjustment of the measure, or re-tailoring, to changes in the user characteristics, optimally before each feedback. Dynamic tailoring has repeatedly been found to outperform static tailoring (Lustria et al., 2013; Krebs et al., 2010).

3 Research Goals and Design

3.1 Goals of the Present Research

Researchers in both the fields of Gamification and Behavior Change Interventions have pointed out the need for a more structured, controlled experimental investigation of the motivational features used (Koivisto & Hamari, 2019; Lustria et al., 2013). They also interdisciplinarily emphasized the need to investigate more than the limited amount of domains and goals which research regarding habit apps has focused on. Currently, it is common to compare highly specialized interventions or apps with completely different designs. This leads to numerous uncontrollable design-based covariates which confound any meaningful results that can be drawn from comparing different apps. To allow for the structured investigation of different activities/habits in one app, a goal of this research is to design and develop an app which allows for full freedom in regards to the habit or activity that is implemented, including its respective goals, measurements and features. This means that the app should allow for the implementation of any quantitatively measureable goal for any kind of habit. The user should in principle also be able to customize all settings, including the habit, goal, goal direction, measurement unit and period in which the goal is measured. This design would then allow for the comparison between different features in regards to a variety of underexamined habits, goals and users in a more controlled setting than what can considered to be the standard at the moment. Multiple motivational features should motivate the user to pursue the goal and ultimately form a habit based on the chosen activity.
Further, the architecture of the app should include mechanisms for tailoring and personalization based on user characteristics, allowing for the intervention and motivational features themselves to be tailored. This supports the widespread demand for user autonomy and the ability to account for individual differences, as pointed out in the theoretical background. (Villalobos-Zúñiga & Cherubini, 2020; Hosseinpour & Terlutter, 2019; Koivisto & Hamari, 2019). The established context- and user dependency of Gamification and similar approaches means that distinct motivational features might work for different users. As has been pointed out, features in Gamification approaches that have shown to positively affect most users nearly always had some users for whom it had a negative effect (Hamari et al., 2014). It has been proposed that context-dependent differences might be explainable due to the user demographics, individual differences of psychological needs according to the SDT, or differences in the characteristics of the habit and goal itself, as well as many other factors involved (Hamari, Hassan, & Dias, 2018; Hakulinen et al., 2015; Hosseinpour & Terlutter, 2019).
Regardless of the correct theory behind the context- and user-dependent effects, the goal in this study is to measure user preferences regarding different motivational features and develop two tailoring mechanisms, one using static tailoring and one using dynamic tailoring to tailor the preferred motivational features to the users.

3.1.1 Research Question

The research question tackled in this study is: Does tailoring motivational features to the preferences of users improve habit formation in an app with customizable habit selection and goal setting?

3.2 Design

To investigate the research question stated above, multiple motivational features from different classes of BCTs or game elements are implemented in the app. The users should be fully free to enter any habit or activity they would like to develop, and select any respective goal based on any quantitative

Table 1

Experimental Design

Condition	Tailor to Questionnaire	Tailor to User Behavior
1	X	X
2	X	-
3	-	X
4	-	-

Notes. Table showing all four conditions used in the research. These consist of all possible combinations regarding the two tailoring mechanisms, with either having both implemented, only one of them implemented or no tailoring at all.

unit of measure. Then, a measurement of the user's interest in the different motivational features should extract the preferred features and subsequently allow for the more prevalent use of their content as well as the features themselves. This can for example be done by using the feature content in notifications shown to the user, or showing the features themselves on the home screen of the app. Using the content of preferred motivational features is ultimately supposed to increase motivation in users and therefore increase target performance on the selected habit. Two different approaches to measure the interest in the respective features are supposed to be investigated, based on the principles of static and dynamic tailoring (Krebs et al., 2010). For the static tailoring, a questionnaire will be administered before the usage of the app. In this, the user is required to indicate the interest in each feature. This should capture user preferences regarding the different features, with the preferences supposed to reflect underlying characteristics that allow for the most efficient features to be extracted per user. The second approach follows the principle of dynamic tailoring, collecting user data during the usage of the app. Here, the interest is reflected in the amount of interaction of the user with each feature. This measure can be continuously updated.

These two classes of tailoring, tailoring to the questionnaire and tailoring to the user behavior, are then investigated in the experiment. The tailoring mechanisms are embedded in a 2x2 experimental design, as shown in Table 1. Condition 1 includes both measures, which are combined to determine the user's interest in the respective features. Conditions 2 and 3 solely use tailoring to the questionnaire or tailoring to the user behavior respectively. Finally, condition 4 is the control condition, which does not use any tailoring, but instead randomly selects the motivational features for the respective content shown to the user.

Four outcome measures are used to indicate the effectiveness of the motivational features on habit formation. These are supposed to convey information on how successful participants pursued their habit and how well users stuck to actually using the app. The ratio of successful daily entries refers to the amount of entries where users successfully achieved their goal. The number of entries refers to the total amount of entries made. The number of logs conveys how many interactions the users in total had with the app features. Finally, the last day of login represents the retention or abandonment rate of the app itself.

3.2.1 Hypotheses

Four hypotheses follow from the research question and experimental design: 1. Both tailorings (conditions 1-3) should individually and collectively lead to a better performance in relation to the performance goal than non-tailoring (condition 4). 2. Both tailorings should have lower rates of abandonment than non-tailoring (same conditions as above). This will be indicated by the length of usage of the app, as it can be assumed that a significant amount of users abandons using the app altogether. 3. Combining both tailoring methods (condition 1) should lead to better results on both outcome measures compared to the individual tailoring methods (condition 2 & 3). 4. The data-driven, dynamical tailoring should improve over time, as more data comes in (condition 3), while the static, questionnaire-based tailoring (condition 2) is expected to outperform in the beginning, but then decay as the measure becomes outdated.

4 Routinary App

The goal of this research was to investigate whether tailoring different motivational features of an app to the preference of users improves habit formation. For this purpose, a smartphone app for Android OS-based phones was developed, called ”Routinary”. This app is supposed to aid users in developing a new habit. Users can track the progress of any habit they want in relation to a self-set daily goal, supporting the development of a personal routine. Motivational features inside the app are then supposed to motivate users to pursue their habit more successfully and avoid abandoning both the goal and usage of the app. Two distinct implementations of the motivational features exist, Motivational cards and notifications. Motivational cards are cards within the app with different content that represent motivational features, for example graphs to show the user's recent progress. Notifications are push notifications that are shown to the user once per day, conveying motivational content based on the aforementioned motivational cards.

This section describes the Routinary app and its functionality in regards to habit formation. First, a short walk-through provides an overview of how the app is used and how the user sets up the habit and goal and finally pursues it. Then, the structure of the main window layout of Routinary is presented, explaining all relevant aspects. Finally, the two implementations of motivational features, namely motivational cards and notifications, are presented. In this section, the Routinary app itself is presented, details on the experiment including how the tailoring mechanisms ultimately work, and the procedure of the experiment follow in the Methods, Section 5.

4.1 Walk-Through of App

The walk-through will provide the most essential information on the usage of Routinary to setup and track the habit, not all aspects of the app in general and relevant to the experiment will be mentioned. Details on the procedure of the experiment are found in Section 5.7, while step-by-step screenshots of all aspects of the app can be seen in Appendix B.

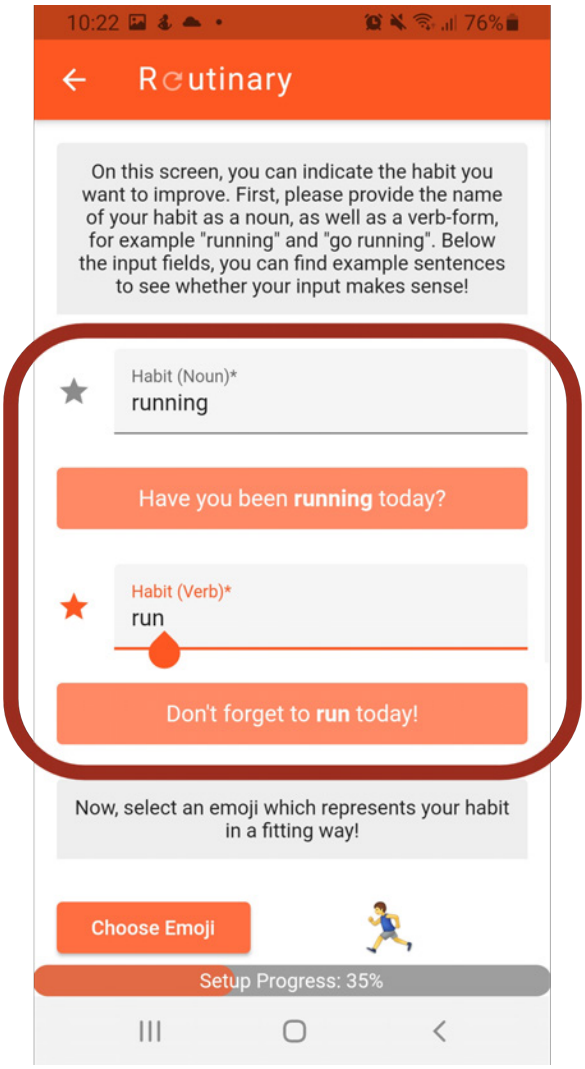
One premise of the research setup was the idea of giving users freedom to decide which activity they want to build a habit on. Therefore, Routinary allows to customize the habit activity, measure and daily goal at the setup of the app. A user first has to enter the activity that the habit is based on. This includes the noun and verb of the activity, for example running and run, as visible in Figure 4a. Both verb and noun are used within the app to adjust messages to the activity of the user. Additionally, users have to enter the measurement unit that their activity is based on (not shown). In case of pursuing running, the progress might for example be measured in meters (m), kilometers (km) or minutes (min.), but users can customize this as they like as long as the measurement is quantitative. Finally, users have to indicate the daily target level that they want to achieve as a habit, as indicated in Figure 4b. Users can only input and track one habit and not follow multiple habits at the same time.

After finishing the setup process, users can start tracking their habit. The main window at startup of the app is shown in Figure 5a, with the layout being described in the next subsection. The motivational cards in this window, as well as other windows are accessible at any time by users and are intended to help them in tracking their progress and increasing motivation to pursue their goal. This is aided by two types of daily notifications. The aforementioned motivational notifications are supposed to increase motivation in users, while daily reminder notifications remind users to enter their daily progress. Finally, to enter the progress for a day, users can press the ”Plus”-button on the main window, allowing them to select a date for which to add the amount of activity done. This is entered in a separate window, shown in Figure 5b.

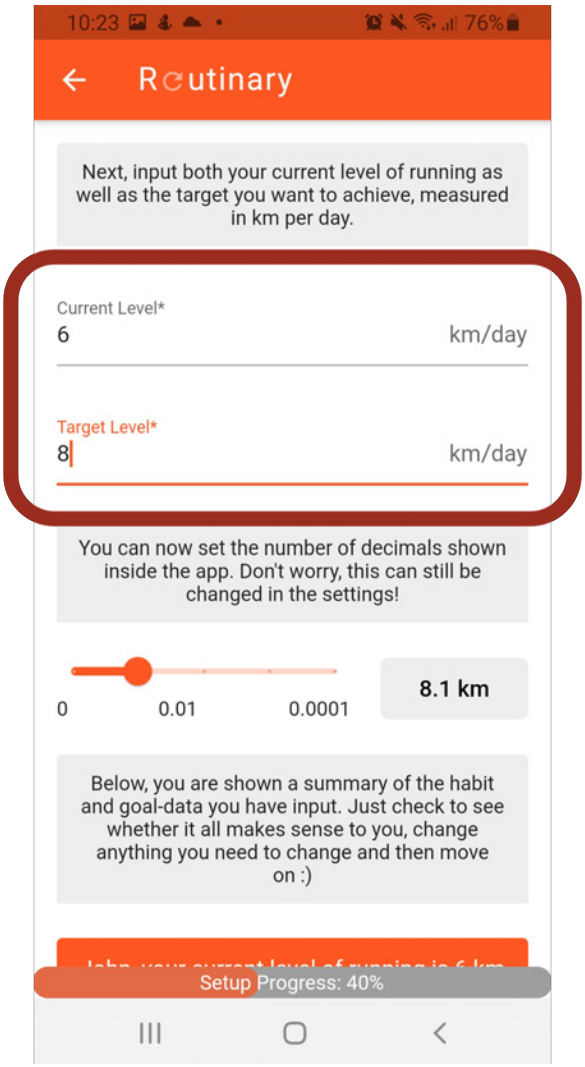
Figure 4

Setup of the Habit Activity and Goal

(a) *Entering the Habit Activity Noun and Verb*



(b) *Entering the Target Level*



4.2 Layout of Main Window

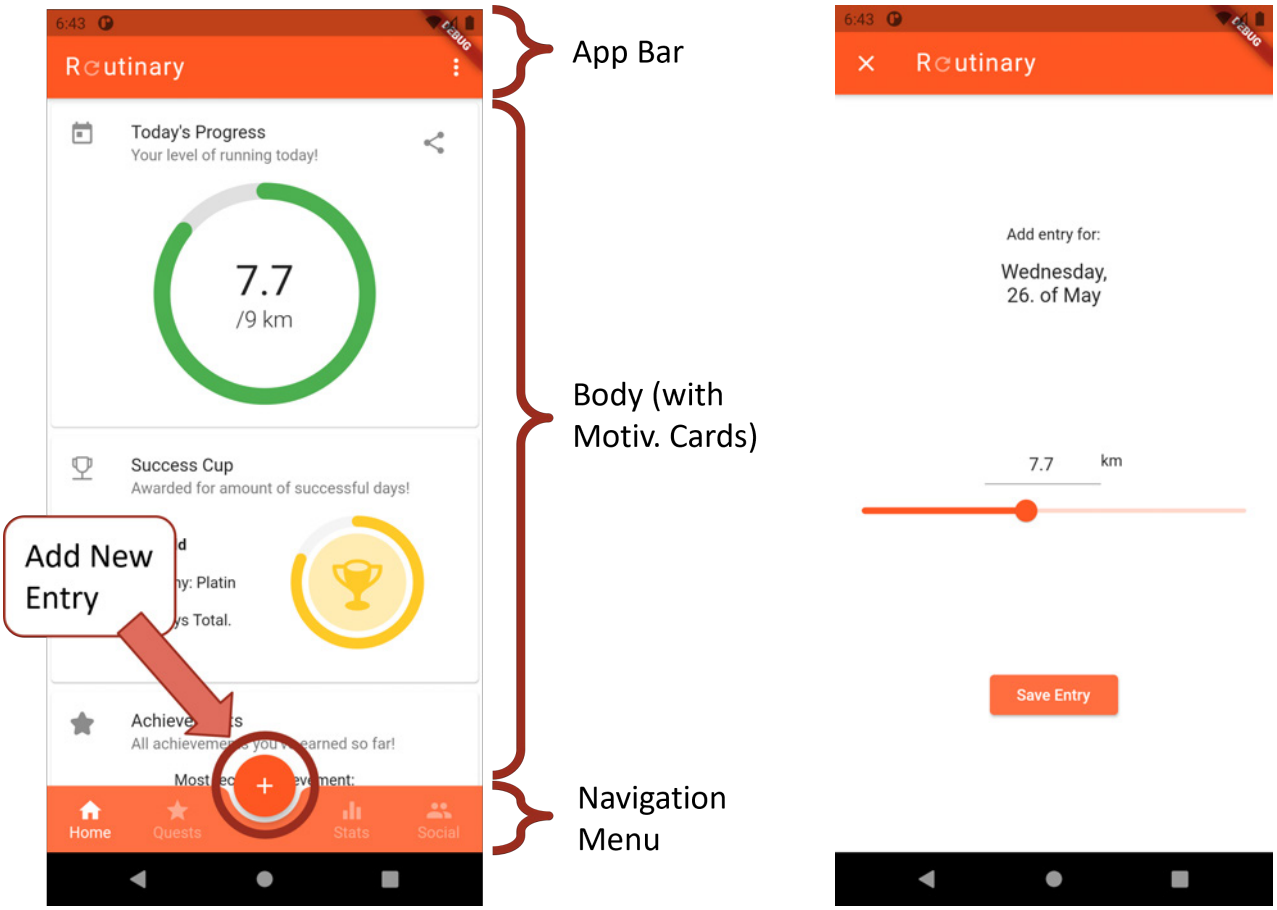
The main window is roughly divided into three parts, the App Bar on the top, the Body in the middle and the Navigation Menu on the bottom. The App Bar shows the Routinary logo on the left, as well as a so-called navigation drawer on the right, which on click provides an expanded menu with secondary options. These include the settings of the app, a contact option, an about section as well as the option to logout. The Body is at the center of the screen. This contains the motivational cards, which are the main aspect of the app and allows users to interact with them. In total, nine different of these motivational cards are available. More details on the motivational cards, including their structure and theoretical background can be found in the next subsection. A detailed listing of all motivational cards including screenshots is presented in Appendix C.

On the bottom of the screen, the navigation menu is shown. This allows to switch between different windows, or screens, each containing different motivational cards in the body. Splitting the motiva-

Figure 5

App Structure and Data Entry

(a) *Main Window* **(b)** *Data Entry Window*



Notes. (a) The basic layout of the main window, including the App Bar, Body and Navigation Menu. (b) The user has the choice of either entering daily progress using the slider, or the entry field.

tional cards up onto multiple windows is done to avoid having to scroll down too far to reach some of the motivational cards. It is used to indicate interest in the respective cards of the different windows, whenever the user opens it. The windows Quests, Stats and Social use fixed cards, so always have the same cards in the same order. The Home screen on the other hand is flexible. It has three slots for motivational cards. The first slot is always filled with the card ”Today’s Progress”, which contains a circular progress bar showing the current day’s progress relative to the daily goal. The second and third slot are tailored to the user’s preferences, which means that the cards here can also be found in one of the other windows by default and then additionally are shown here, providing easier access to their content. A further explanation of this can be found in the methods, also explaining the experimental conditions regarding the tailoring mechanisms which determine the cards being shown in these positions.

Finally, embedded inside the navigation menu is an action button with a plus symbol, which is surrounded by a notch. This button allows the user to add a new daily entry. When pressed, it opens a popup menu on the bottom to select whether to add an entry for today, yesterday or another day. In

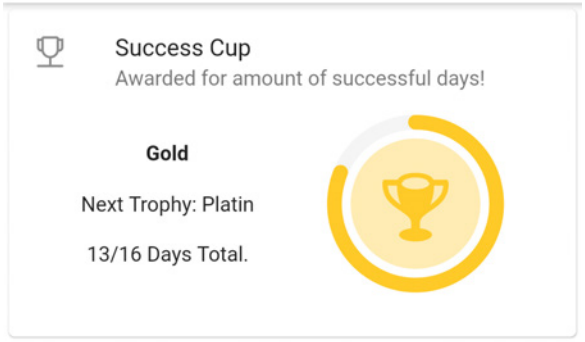
case of selecting the other day, a date selector in form of a calendar is shown to specify the date. After selecting the date, the user can enter the performance for the selected date in a separate window, as shown in Figure 5b and then save the new entry, or overwrite the existing one.

4.3 Motivational Cards

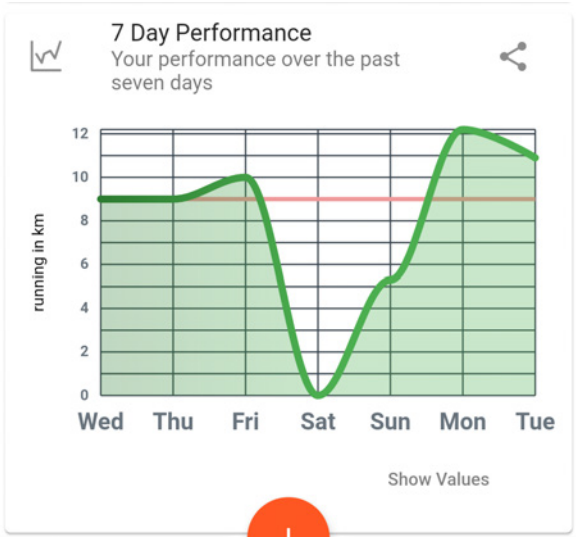
Figure 6

Motivational Cards/Features Content.

(a) *Success Cup-Card*



(b) *7 Day Performance-Card*



Notes. (a) In-app screenshot of the Success Cup, which rewards the user with a colored and ranked badge for the total number of successful days. (b) The 7 Day Performance-card, which serves the user with a graph representing the progress relative to the target over the last 7 days.

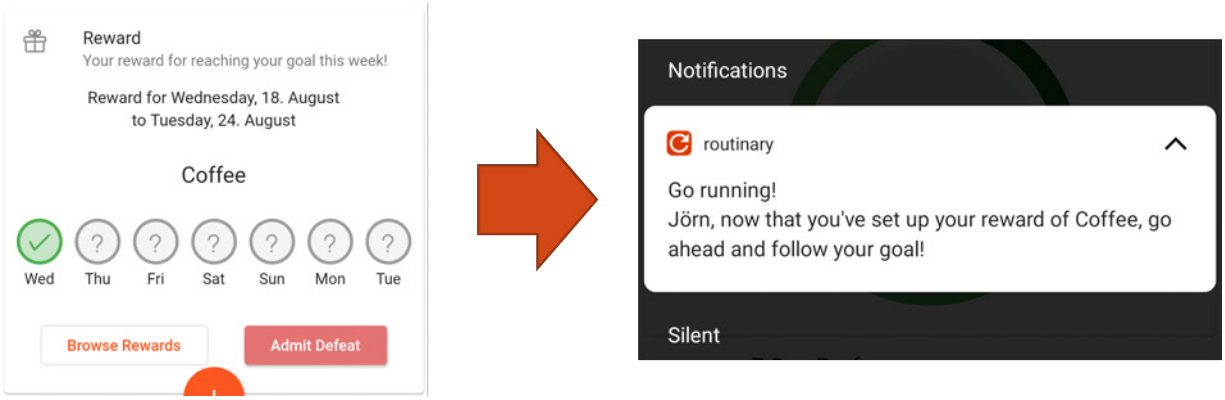
Essential to the experience of using the app are the motivational cards which can be found on the main screen. The goal of these is to motivate the user to achieve their personal target. Each card is based on multiple and differing established motivational features, as described in the theoretical background. One example is the motivational card ”Success Cup”, shown in Figure 6a. Following the taxonomy presented in (Villalobos-Zúñiga & Cherubini, 2020), this would be coded as a badge, therefore the motivational feature it is classified as is a ”reward”. In comparison, the ”7 Day Performance”, shown in Figure 6b shows a progress graph and would be classified as history. However, both cards also have content that can be shared, which additionally classifies them as ”performance sharing”, so both cards use multiple motivational features that overlap to some degree. For a detailed listing of all motivational cards, including an explanation of their functionality, see Appendix C.

It is important to emphasize that it was not the goal to compare the different clusters found in any of the named taxonomies, or represent all of them in the motivational cards. Instead, the different motivational cards are supposed to appeal to different users, so that preferences between those cards can be detected and used for the tailoring mechanisms, which are explained in Section 5.

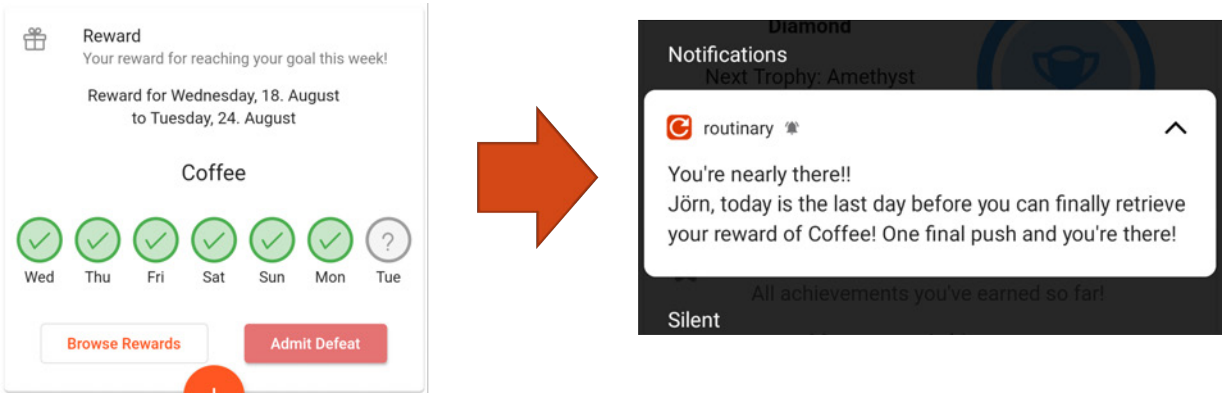
Figure 7

Notifications based on Motivational Cards

(a) *Card and Notification based on a new Reward*



(b) *Card and Notification right before achieving a Reward*



Notes. Examples of how the content of motivational cards is used as a basis for the content of notifications. The motivational card "Reward" allows users to set up a reward for sticking to the self-set goal for 7 days in a row, in this case coffee. **(a)** When the user has just setup the reward, a notification based on the "Reward"-card would refer to the newly setup reward to motivate the user. **(b)** On the last day before retrieving the reward, the motivational notification would have a different content, trying to motivate the user by referring to the soon reached reward.

4.4 Notifications

Push notifications are messages sent to users at specific times during the day. These are shown even when the app is not opened and appear in the notification menu on screen, although depending on the specific phone and operating system, the exact delivery mechanism can differ. Technically, they have been implemented using the Android Alarm Manager, which is prone to compatibility issues with phones by some manufacturers like Huawei.

Two kinds of notifications are used, a reminder and a motivational statement. Both are shown on a daily basis at a set time, unless they are turned off in the settings menu. The time of show can also be changed for each. The reminder is a simple notification, reminding the user to enter their daily habit progress. This is not tailored depending on the condition, so all participants receive the

same type of reminder. The default time to which the reminder is set is 20:00 local time, which can be customized both during the setup of the app, as well as afterwards in the settings menu. The motivational statement is, as the name suggests, supposed to motivate the user to follow the habit goal. The content of motivational notifications is based on the motivational cards. This means that each motivational card has different messages depending on the state it is in that can be used in motivational notifications. An illustration of this is shown in Figure 7, where the notification content is based on different states of the motivational card ”Personal Reward”. Multiple of these messages are implemented for each of the motivational cards. This makes it possible to tailor which of the motivational cards is used as basis for the notification in the experiment. The exact mechanism of this is explained in the next section.

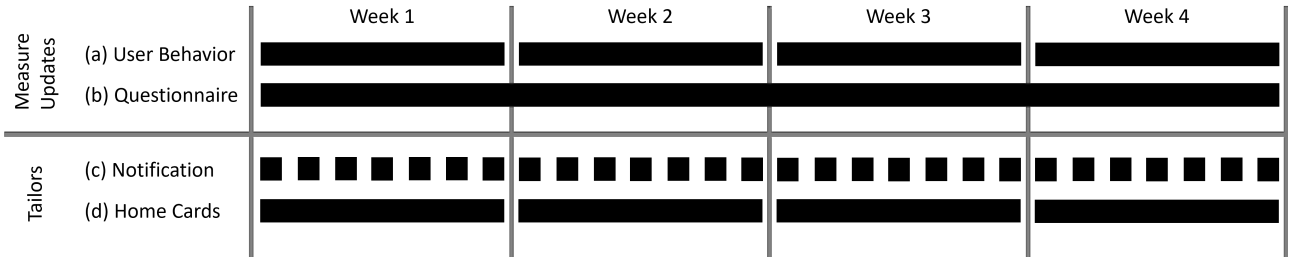
5 Methods

This section presents the methods relating the experiment. For the experiment, the Routinary App presented in the previous section was utilized to investigate the goal of tailoring motivational features to user preferences in an attempt to improve habit formation, as indicated in the theoretical background. Two types of content in the app are tailored to the preferences of users: Which of the motivational cards in the Routinary app are shown on the home screen and what content from the motivational cards is used in the motivational notifications. This selection is based on the user preferences regarding the motivational cards, meaning which of the motivational cards the user has shown most interested in.
Two measures are used to assess the user preferences of the different motivational cards, one being used for static and one for dynamic tailoring. The static tailor is based on a questionnaire shown to the user at the setup of the app, with all motivational cards presented to the user, so the level of interest in each card can be indicated. The dynamic tailoring is the "User Behavior" measure, which is based on the recorded user behavior. For this, user interactions with the different motivational cards are logged. Higher levels of interaction with a card means a higher level of interest, therefore indicating that the card is more preferred by the user.
The section is structured as follows. First, a general overview of the tailoring setup of the experiment is given. Then, the tailored content is more specifically explained, including which content is tailored and how it is tailored based on the user preferences. Then, the two measures of user preferences are presented. Based on this tailoring setup, the specifics of the experimental procedure are explained. Finally, the statistical analyses are described.

5.1 Tailoring Setup of the Experiment

Figure 8

Tailoring Timeline



Notes. Timeline of the updates of measures and their effects on the deployment of tailored content throughout the experiment. Note that the measures used for each participant depend on the experimental condition. The beginning of each bar indicates an update to the measure in (a) and (b) or a new deployment of content in (c) and (d). **(a)** While the user behavior data is constantly logged, the measure used for tailoring is only updated once per week, at the beginning of the week. **(b)** The questionnaire is only administered once at the beginning of the experiment with the ratings being used throughout the whole experiment. **(c)** Motivational notifications are shown each day. As (a) and (b) imply, notifications throughout each week are based on the same measures, as the UB measure does not update during the week. **(d)** Home cards are updated at the beginning of each week, coinciding with the update of the UB measure and being based on this updated measure.

As presented in the theoretical background, different users are expected to be interested in different motivational cards and their respective contents. Therefore, one can assume that presenting irrelevant

or uninteresting content has a lower positive or even a negative effect on the motivation of the user regarding both the goal and the usage of the app itself. Consequentially, by presenting content specific to the preferences of the user, motivation should increase and performance improve.
The tailored content of the app includes the second and third positioned cards shown on the home screen, as well as the motivational notifications. As explained in Section 4.3, the home screen is shown at startup of the app. While all motivational cards are always found in the same position at their respective screens, the second and third position on the home screen are tailored according to what motivational cards a user has shown the highest level of interest in. The position of the tailored cards on the home screen is visible in Figure 9. The motivational notifications are push notifications shown once per day. As presented extensively in Section 4.4, the content of these notifications is based on the motivational cards with all cards having their own notification content conveying information based on the card content. Similar to the motivational cards on the home screen, the cards used as basis for motivational notifications are tailored based on which cards users are most interested in.
To measure the interest users have in the cards two types of preference measures are used, a card questionnaire and a user behavior measure (UB measure). The card questionnaire is used for static tailoring, while the UB measure is used for dynamic tailoring. This allows for the effects of static and dynamical tailoring to be compared. The card questionnaire presents all cards to the user to be rated for user interest on a scale from 1 to 5, with a higher rating representing higher levels of interest in the feature. It is applied only once for the purpose of tailoring, before the start of the experiment. A post-experiment questionnaire is applied, but not used for tailoring, but only to compare for changes in preferences between the beginning and end of the experiment. The measure used for dynamical tailoring is based on recordings of the user behavior (UB measure). Various interactions with the different motivational cards are recorded and determine how interested the user is in the motivational cards, with more interactions indicating a higher level of interest in the respective card. While user behavior is constantly logged, the card ratings based on user behavior are only updated once per week at the beginning of the week. This update is based on all logs available up until that point in time.
The timeline of the tailoring within the four weeks of the experiment is summarized and explained in Figure 8.

5.2 Determining the Tailored Home Cards

Let C be the set of all n motivational cards, or motivational features.

$$C = \{c_1, c_2, ..., c_n\} \tag{1}$$

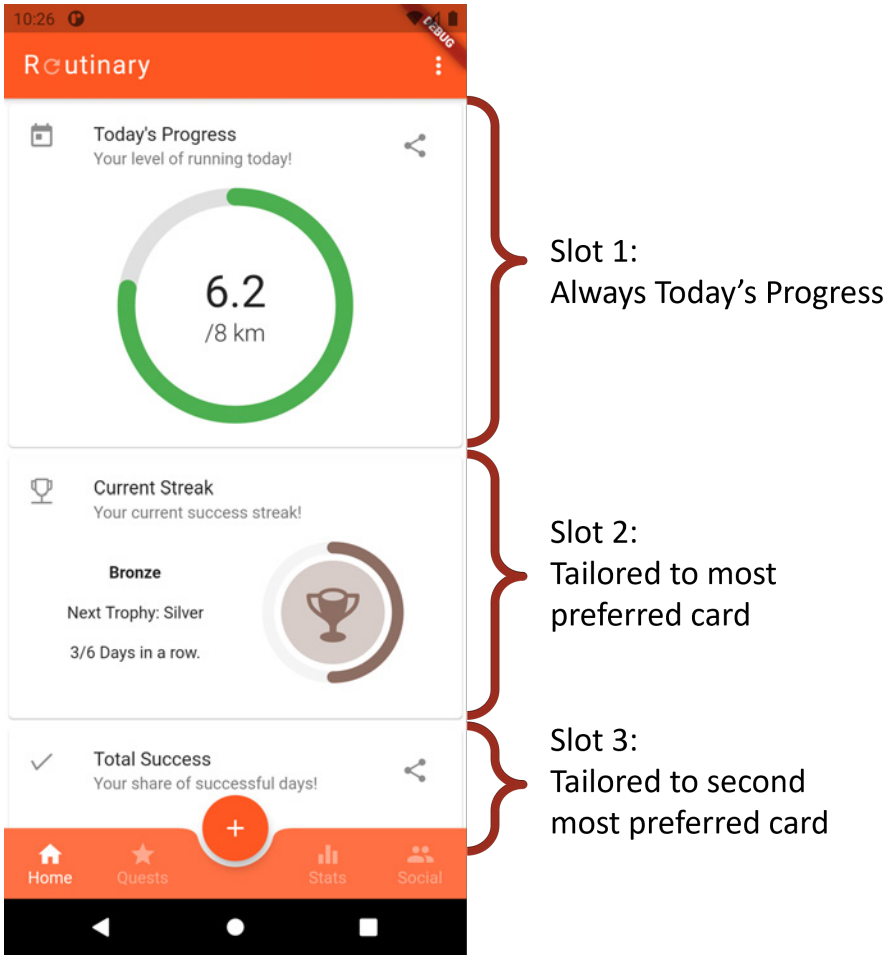
The target is to find the preferred cards per user of all cards in C. These are determined based on the different levels of interest I users have in all cards. These interest levels are contained in the set

$$I = \{I_c \text{ for all } c \in C\} \tag{2}$$

with the interest in each card c being noted as I_c. The variable I_c with the highest value would therefore refer to the most preferred motivational card. The measures used to determine interest values are presented later.
The second and third card on the home page, α and β, are then determined based on the highest values of interest shown in the cards. This means that the motivational card c used for the second slot α is determined by which of the cards the user has shown the highest interest I_c in. The formula to retrieve the most preferred card is therefore:

Figure 9

Positions of Tailored Second and Third Home Card



Notes. Home screen with the position of the tailored cards. The first slot is always fixed as showing the motivational card "Today's Progress". The second slot shows the most preferred motivational card according to the tailoring, which in this case is the "Current Streak"-card. Finally, the third slot shows the second most preferred card, which here is "Total Success", although it is not fully visible in the screenshot.

$$\alpha = \{c \text{ for } I_c \in I : \max I\} \tag{3}$$

The third card slot on the home screen, β, is determined by the second-most-favoured card, or the most-liked card of the remaining cards, being:

$$\beta = \{c \text{ for } I_c \in I \setminus \{I_\alpha\} : \max I\} \tag{4}$$

5.3 Determining the Content of Motivational Notifications

As presented in Section 4.4, the content of the motivational notifications is based on the different motivational cards. To determine which motivational card notifications is based on, the same set C of Cards from Equation 1, as well as the same interest values I from Equation 2 are used. Contrary to the determination of the home cards, the interest in each card only determines its probability to be

used for the notification content. This is done to avoid too much daily repetition of the same content being presented over and over again, instead varying the content shown regularly.
The interest values are interpreted as meaningful, which means that for example a value of $I_c = 0.6$ would mathematically indicate double the interest in a feature than $I_c = 0.3$. The probability $P(c)$ of each card c to be used in the notification content therefore is computed based on the fraction of the interest I_c shown in the card c, compared to the combined sum of the interest values I in all cards C. This is calculated as:

$$P(c) = \frac{I_c}{\sum_{i=1}^{c} I_i} \tag{5}$$

and would mean that in the previously mentioned example, the card with $I_c = 0.6$ has twice the chance of being selected than the card with $I_c = 0.3$.

5.4 Tailoring to Questionnaire

Before the experiment started, all participants were presented with each motivational card, a description of its workings and the question "How much are you interested in this widget?", see Figure 12. They then have to respond by rating each card on a 5-point Likert-scale from "Not at all" to "Extremely". This questionnaire is used as an indicator of interest in the respective cards, with the scoring being used for the tailoring.
For tailoring to the questionnaire, the likeliness of the card being shown on the home screen or being used for the notification content is therefore based on the ratings in the questionnaire.
As we have seen in Equation 2, the interest values in the different cards were contained in the set I. The way these are determined depends on the experimental condition. The partial interest values of the cards C based on the questionnaire will be classified in the set I^q, consisting of:

$$I^q = \{I_c^q \text{ for all } c \in C\} \tag{6}$$

The likert-scale in the questionnaire is coded from 0, representing "Not at all" to 4, representing "Extremely". Ratings for all cards in C are contained in the set R.

$$R = \{R_c \text{ for all } c \in C\} \tag{7}$$

Each response R_i in R is mapped linearly to retrieve the inferred questionnaire-based interest value I_i^q according to the function:

$$f : 0.2 * R + 0.2 \mapsto I^q \tag{8}$$

At a full rating of "Extremely", the interest value I^q will therefore be set to the maximum of 1, while at the lowest rating of "Not at all", it will still be at 0.2. This is to still have a small chance of the card being chosen, even if the reported interest is low.
To control for possible changes of preferences over the course of the experiment, the same questionnaire is presented to the participant again after the experiment is finished. For the tailoring itself, only the pre-questionnaire is used which receives no update over the course of the experiment.

5.5 Tailoring to User Behavior

Within the app, user interactions with the menus and motivational cards are logged. The recorded interactions are used as an indicator for the preferences regarding different motivational cards. Multiple

Table 2

Tailoring Factors per Motivational Card

Card	Window	Others	Card	Item	Shared	Refresh
Success Cup	0.6	0.2	-	X	X	-
Achievements	0.6	0.2	X	X	X	-
Personal Reward	0.6	0.2	X	X	X	-
Five Day Chart	0.7	0.3	-	-	X	-
Total Success	0.7	0.3	-	-	X	-
7 Day Performance	0.7	0.3	-	-	X	-
Personal Motivation	0.6	0.2	-	-	X	X
Accountability	0.6	0.2	X	-	X	-

Notes. The table both shows which factors are implemented per motivational card, with all implemented ones being marked with an "X". It also shows the factor determining the impact that opening the window has, as well as the combined impact of all other logged user behaviors. Detailed explanations of the user interactions can be found in Section 5.5.

interactions with the cards are factored in, which will be presented in this section and are visualized in Figure 10. These include the opening of the window in which the motivational card is present, interacting with or opening the card itself, interacting with or opening items like achievements or individual rewards that are present within some cards, sharing a screenshot of the content of the card and refreshing the content of the card. Not all cards have the same interactions available, for a listing of the respective behaviors available for each card as well as their factors in computing the probability of the card being used, see Table 2.
Similar to the set containing the partial interest values from the questionnaire in Equation 6, a set containing the partial interest values regarding the user behavior is defined:

$$I^u = \{I^u_c \text{ for all } c \in C\} \tag{9}$$

The interest values in I^u are determined based on a combination of all user behaviors mentioned, which will be presented in this subsection, with the combining equation being provided at the end. The exact formulas and proportions used to determine the interest-values in the cards were first based on assumptions about the way the respective features work. Based on insights from a pilot run, these formulas were then tweaked in an attempt to balance the interactions accordingly.

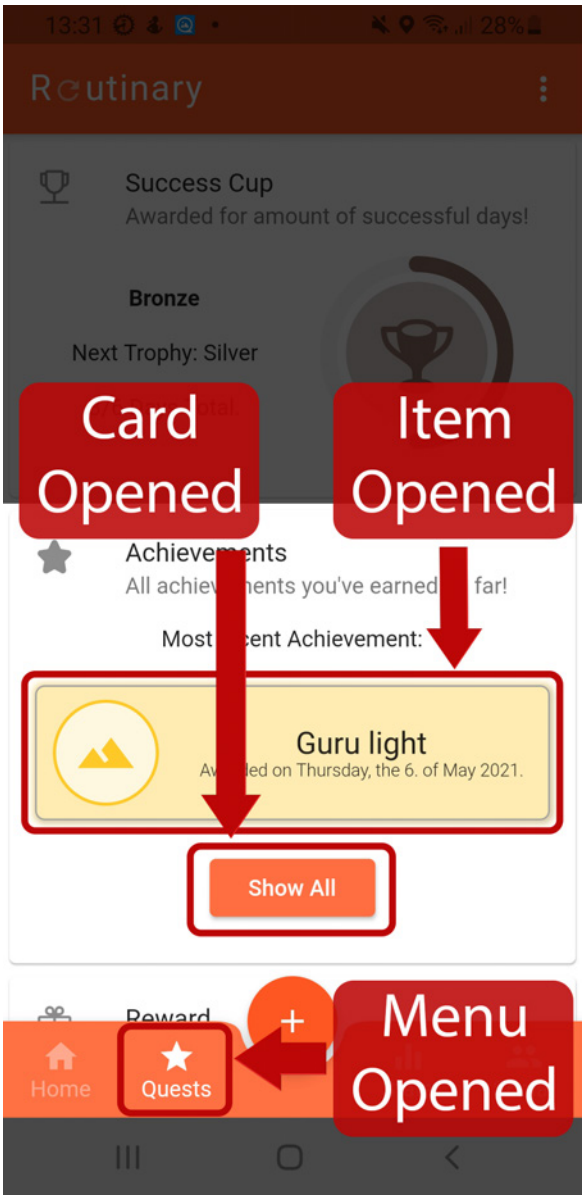
5.5.1 Window Opened

The action of opening the window relates to pressing the option in the navigation bar which shows the window containing the motivational card. This also means that opening this window leads to increased preference-values for all 2 or 3 motivational cards that are part of this window, as it can not be reliably distinguished for which of the cards the user has opened this window. As can be seen in Table 2, this is the most influential single factor in the probability of the card being shown. The reason for this choice is that opening the window containing the card is both necessary and sufficient to view the contents of the card, with all other interactions being optional.

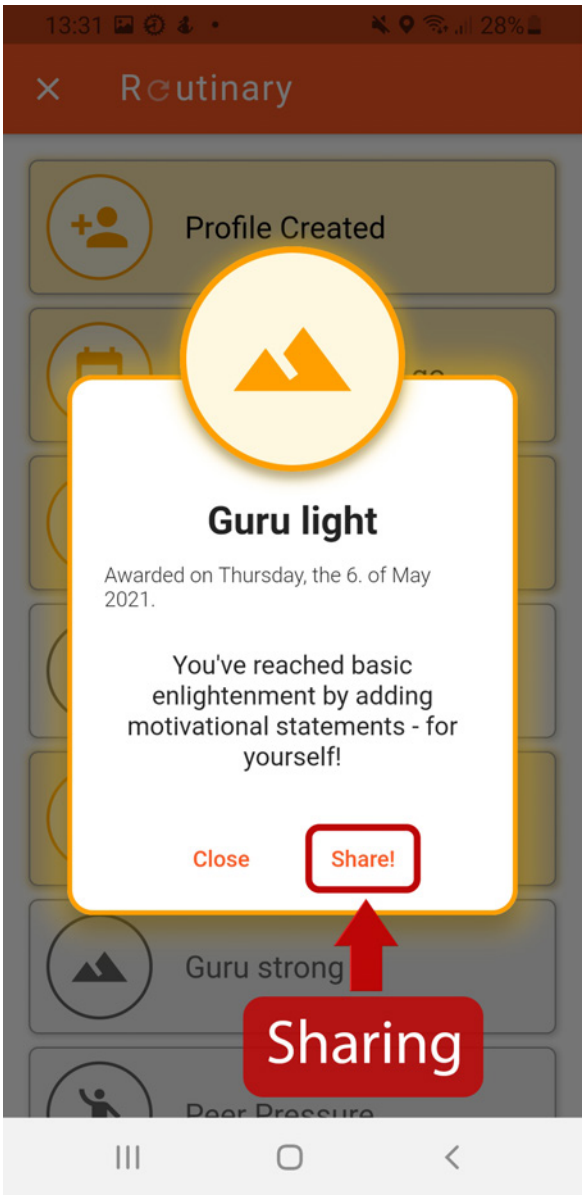
Figure 10

In-App User Interactions

(a) *Achievement Card with User Interactions*



(b) *Opened Achievement with Sharing-Option*



Notes. Example of the user behaviors determining the liking of the different motivational cards. This example shows the logged interactions with the achievement-card. Be aware that different cards have different available interactions, as can be seen in Table 2.

The different windows include multiple motivational cards each. This is why the action of opening a window is related not to the set of cards, but to a set of windows W, representing the three available windows:

$$W = \{w_1, w_2, w_3\} \tag{10}$$

The values representing the preferences of the respective windows in W can be found in the set A^m

$$A^m = \{A^m_w \text{ for all } w \in W\} \tag{11}$$

The preferences regarding different windows A^m_w and their content is inferred based on how often they have been opened relative to the other windows. This is conveyed in the simple formula:

$$A^m_w = \frac{m_w}{M} \tag{12}$$

with m_w being the number of times the respective window was opened and M being the total amount of openings of all available windows.

5.5.2 Card Opened

The next interaction used is opening or interacting with the motivational card itself. The way this interaction is implemented differs between cards. It also has to be differentiated from the next interaction of "Opening an Item", as this requires multiple items to be interactable with inside each card. "Card Opened" however only requires one single interaction per card. As an example, opening the window listing all achievements qualifies as "card opened", while opening an achievement inside this window, or on the base card where the most recent achievement is shown qualifies as "item opened", as multiple achievements are available.
Preferences regarding different cards based on how often they have been opened are contained in the set A^o. This only entails cards from C which have the feature of card opening implemented, so:

$$A^o = \{A^o_c \text{ for all } \{c \in C : c \text{ is openable}\}\} \tag{13}$$

The preference is based on two aspects. The first one is based on whether the respective card c has ever been opened, regardless of how often, see:

$$o^r_c = \begin{cases} 1, & \text{if } o^t_c \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where o^t_c is the total number of times that the card c was opened or interacted with. The formula leads to o^r_c either equating 1, when the card has ever interacted with, or 0 when it hasn't. The result then embedded in the function to compute A^o_c:

$$A^o_c = 0.5 * o^r_c + 0.5 * \frac{o^t_c}{O} \tag{15}$$

where half of the preference rating is determined by whether the card c has ever been opened, so o^r_c, while the other half depends on the fraction of the total number of openings of card c, o^t_c, by the overall number of card openings of all cards in C, O. The second part of Equation 15 mirrors the approach used in Equation 12.

5.5.3 Item Opened

As explained above, the action of opening an item relates to the interaction with one of a limited set of items that are part of a motivational card. For example, in the achievement card, all achievements, which qualify as items, can be listed, and those that have been achieved are openable. Clicking one of these items counts as an item opened, with the full achievement consequentially being shown to the user, as seen in Figure 10b.

The preference values for the opened items and their owning cards are contained in the set A^i. The set follows the same logic as used in Equation 13, with the cards from C represented which have the item opened-functionality implemented, see Table 2.
Similar to the Card Opening equation, the equation for the opening of the item consists of two summed expressions. The first one follows the same logic as Equation 14, so:

$$i_c^r = \begin{cases} 1, & \text{if } i_c^t \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

with i_c^t being the total number of items in c opened, which as well leads to i_c^r equating 1 if any item has been opened or 0 if no items have been opened. i_c^r is then integrated in the following equation:

$$A_c^i = 0.5 * i_c^r + 0.5 * \frac{i_c^t}{i_c^a} \tag{17}$$

which is similar to Equation 15, with i_c^t being the total number of opened items. However, instead of using the total number of interactions of all cards as a denominator, here, i_c^a is used, representing the total number of available items that are openable in the card, irrespective of interactions in other cards. Therefore, the fraction $\frac{i_c^t}{i_c^a}$ represents the fraction of items that were opened of all available, openable items.

5.5.4 Shared

Sharing refers to the action of sharing the content of a card or, optionally, an item that can be opened in the card. An example of this option is visible for the achievement card in Figure 10b. When pressing the share-button or share-icon, the OS-own share-menu will be opened. This allows to share a screenshot of the content of the card, for example via social media, email or in any kind of messenger.
The set of card preferences A^s again follows the pattern shown in Equation 13, although the share-functionality is implemented in all cards in C. The equation to compute the interest based on the sharing also follows the same pattern as Equation 15 for the card opening, so:

$$A_c^s = 0.5 * s_c^r + 0.5 * \frac{s_c^t}{S} \tag{18}$$

with s_c^r being 1 in case any sharing of a specific card or card-content has taken place and 0 in case it has not, the same logic used in Equation 14. Also, the fraction follows the same logic as in Equation 15, with s_c^t being the amount of times the card has been shared and S being the total amounts of sharing taking place in all cards.

5.5.5 Computing the Interest-Values per Card

The interactions laid out above now determine the user behavior interest values I^u. First, the set of preferences based on the different user interactions tracked can be collected in A, consisting of:

$$A = \{A^m, A^o, A^i, A^s, A^r\} \tag{19}$$

However, as is transparent from Table 2, not all interactions are available for all persuasive features. Logically this means that the set of preferences for each card only includes those interactions, which

are implemented. As an example, the set for the card Success Cup, A_{suc}, would only include the interaction values for opening the menu, interacting with the item and sharing the content:

$$A_{suc} = \{A^m_{suc}, A^i_{suc}, A^s_{suc}\} \tag{20}$$

Further, for the determination of I^u the interaction values in A are split into two parts, with each part having a different influence on the outcome depending on their own factors. For this, the interaction value determined on the opening of the respective menus, A^m, is treated separately from all other interactions, which are therefore part of the subset A^e:

$$A^e = \{A \setminus A^m\} \tag{21}$$

The reason for this separation is that opening the menu/window, which is included in all features, is a necessary and sufficient condition to access the included cards. It therefore has the largest impact on I^u, as is visible from the second column in Table 2, with the factor determining the impact of A^m ranging from 0.6 to 0.7, with slight differences depending on insights from pilot testing. The interaction values included in A^e are supposed to mostly separate the respective cards within the window and have a much smaller combined impact, with the factor ranging from 0.2 to 0.3. Note that both factors are not supposed to add up to 1, again correcting for differences in the user behavior regarding each card. Ultimately, the interest in each card based on user behavior, I^u_c can be determined using the formula:

$$I^u_c = a_c A^m_c + b_c \frac{\sum_{i=1}^{n} A^e_{ci}}{n(A^e_c)} \tag{22}$$

This consists of two parts: The first part represents the impact of opening the menu, A_c, which is scaled by the respective factor a_c that can be read out from the Table 2. The impact of all other interactions A^e_c is added, with all interaction values being averaged and scaled by the factor b_c. As pointed out, number n of members of the set A^e_c differs per card c, which is the reasoning for the averaging-process. To provide an example of Equation 22 in practice, the formula to determine the user behavior-based interest in the card Success Cup, I^u_{suc} is given, with the elements being available for lookup in Table 2:

$$I^u_{suc} = 0.6 * A^m_{suc} + 0.2 * \frac{A^i_{suc} + A^s_{suc}}{2} \tag{23}$$

5.6 Computing Interest-Values per Condition

As mentioned in Section 3.2, four conditions are included in the experiment, using the 2x2 factorial design shown in Table 1. The combination of interest values based on the questionnaire, I^q and the user behavior, I^u is quite straightforward. Condition 1, which means tailoring to both methods, just uses the average of both to compute the interests I in all cards, being:

$$I = \frac{I^q + I^u}{2} \tag{24}$$

Condition 2 and 3 only use one of the two tailoring mechanisms, therefore the overall interest is solely based on the respective interest values. For condition 2, tailoring to the questionnaire this means:

$$I = I^q \tag{25}$$

While for condition 3, tailoring to the user behavior the equation is:

$$I = I^u \tag{26}$$

Finally, condition 4 is random and therefore does not use any interest ratings, instead the cards are assigned in a fully random order.

5.7 Procedure

In this section, the experimental procedure will be described. The procedure was approved by the Research Ethics Review Committee (CETO) of the University of Groningen. A concise overview of all phases and the steps in each phase is depicted in Figure 11. More details, including screenshots of the respective screens shown to the participant can be found in Appendix B.

5.7.1 Recruitment of Participants

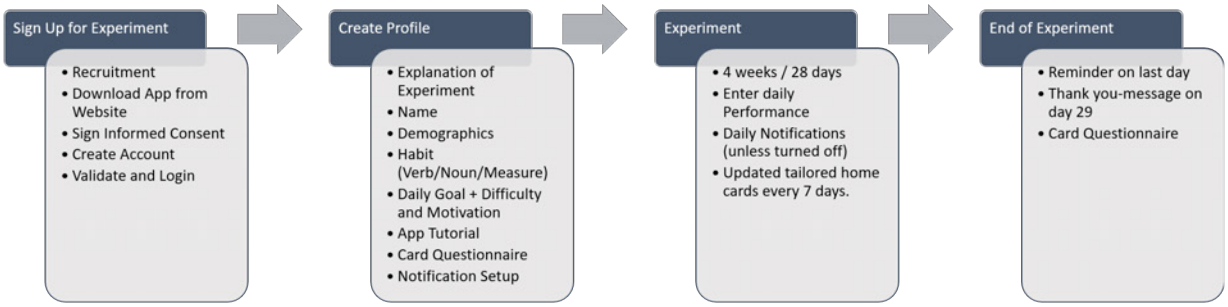
Recruitment of participants was done via multiple platforms on the internet. These included LinkedIn, Twitter, Facebook and multiple Email-Lists of university students, for example of the students of Artificial Intelligence at the University of Groningen. While there were no limits in regards to the type of participant signing up for the study, the nature of the experiment as being conducted through a smartphone app, as well as the recruitment channels used will likely have lead to a recruitment bias towards university students with an interest in technology or psychology.
The recruitment message included a link towards a website hosting more information regarding the app and experiment itself, as well as a download link and QR-code including instructions to download an .apk-file which allowed for the installation of the Routinary-app on Android phones. The website can be found at www.oenkophon.de/routinary. It also included a contact form for questions and comments regarding the experiment.
No financial or other rewards were given for participating in the experiment, although participants were able to continue using the app after the the experiment was concluded.

5.7.2 Signing Up and Creating Profile/Habit

After the participant has downloaded, installed and opened the app for the first time, an informed consent is shown. This includes an explanation of the context and purpose of the experiment, methods of data collection, usage and storage as visible in Appendix B, as well as information on how to contact the researchers for questions and comments, withdraw from the study and the contact information of the Data Protection Officer of the University of Groningen. Progressing to the next step is only possible if all points of the consent form have been agreed to.
Next, the participant has to create an account for Routinary using a valid email address and a password, with the e-mail address having to be confirmed by clicking a validation link in an email. This allows to login to the account and create the in-app profile. When creating the profile, the participant is first presented with a short tutorial explaining how the habit and goal as well as the experiment are structured. The participant is then asked to provide a name, which is purely used inside the app and not transmitted to the experimenter. In the next step, demographic data can be provided, although this step is optional. This includes age, nationality, education, gender and proficiency in the English language.
The participant is then supposed to enter the habit pursued in form of a verb and noun, including feedback sentences which provide information on whether the form of the term entered is correct. Additionally, an emoji representing the habit and the measurement unit are selected and entered. On

Figure 11

Experimental Procedure



the next screen, the participant enters the estimated current daily performance of the habit, as well as the target level which serves as the goal. Finally, two questions are posed, ”How difficult do you expect your goal to be?” and ”How motivated are you to achieve your goal?”, both having to be entered on a 5-point Likert scale from ”Not at all” to ”Extremely”. This serves two purposes. First, this measure can be used to check whether difficulty and motivation have an effect on the outcome. Related to this, both difficulty of the goal and the user’s motivation might mediate the effect of tailoring on the outcome and can therefore be included in the final model investigating the effect of tailoring on outcome.

Finally, a short tutorial of the app layout is shown to the participant, where all relevant aspects are pointed out. This includes the cards themselves, their header, content, the share button, the bottom navigation bar of the app and the action button to add new entries. Afterwards, the participant goes on to the card questionnaire.

5.7.3 Card Questionnaire

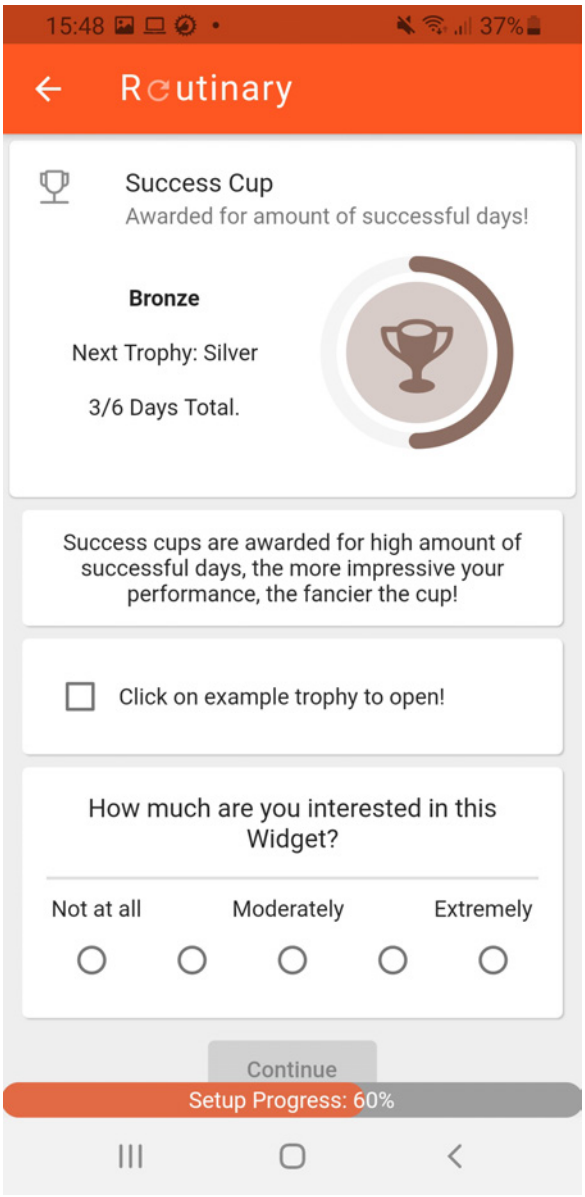
The card questionnaire is used to determine the condition of tailoring to the questionnaire. It sequentially shows all cards that are available in the app and requires the participant to rate them based on interest, responding to the question of ”How much are you interested in this widget?” using a 5-point likert scale from ”Not at all” to ”Extremely”. An example of a screen is shown in the questionnaire in Figure 12a. It shows an example of the card itself, a short description, if applicable a checkbox to indicate that it can be clicked/opened with the example of this being visible in Figure 12b, and finally the question itself. This procedure is repeated for all nine cards.

Finally, two screens relating the daily notifications are shown. On the first screen, the participant has to send a test notification and indicate whether this has worked as intended. On the second screen, the participant can turn two types of daily push notifications on or off and can determine the time of day when they shall be shown. The first notification is a simple daily reminder, reminding users to enter their data for the day. The second notification is the motivational notification which entails a tailored message depending on the condition. This is targeted towards increasing the users activity, as described before. By default, both notifications are turned on and set to 12 am and 8 pm, respectively. After potentially adjusting these settings, the participant has finished the setup of the app and can proceed with using the app. While the duration of the whole setup process can be assumed to vary between participants, it is expected to take between 15 and 20 minutes. Afterwards, the participant can use the app freely and the experiment itself starts.

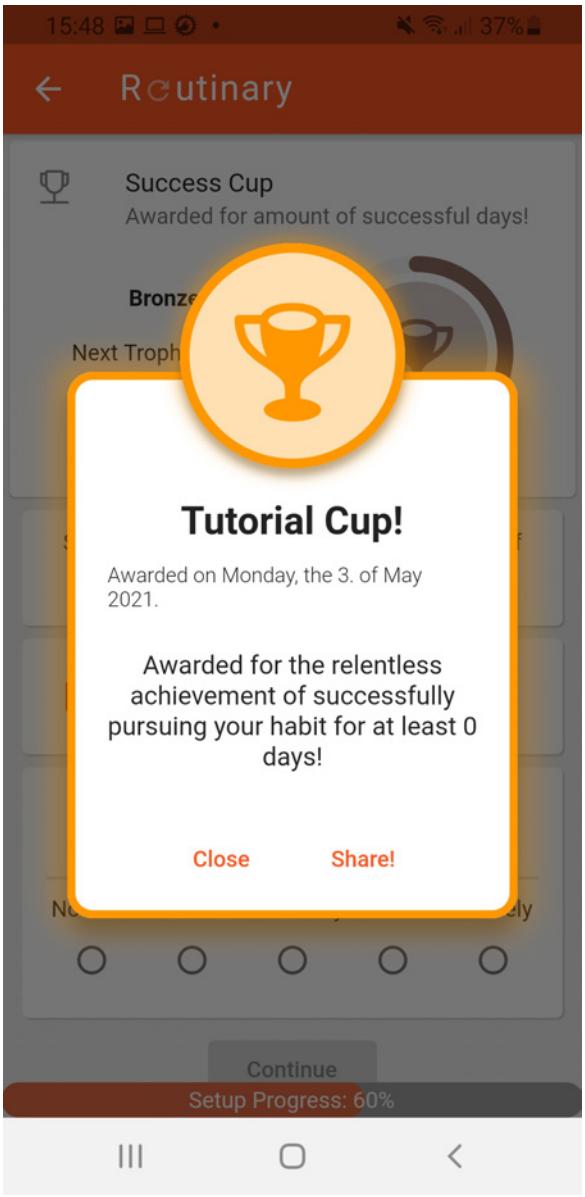
Figure 12

Card Questionnaire

(a) *Card Questionnaire Screen*



(b) *Opened Example of Success Cup*



Notes. Example of the card questionnaire, with the elements of the card itself, a short description, an optional check to indicate interaction with the card and the 5-point likert scale question to indicate interest in the card.

5.7.4 Experiment

The experiment ran for 28 days, or 4 weeks, with the day of the setup being day 1. The participant received daily motivational and reminder notifications on their phone, unless those were turned off or were not working due to restrictions by the operating system. Participants were supposed to enter their performance on their habit goal for each day of the experiment, although missing some entries or abandoning the app as a whole was expected for a sizeable amount of participants and was factored

into the hypothesis. Participants were able to use the app as they like, including any and all features they were interested in. The usage was tracked and, depending on the experimental condition, used as basis for tailoring. Every seven days beginning with day 1, the second and third card shown on the home screen were updated based on the recalculated tailoring, as detailed previously.

5.7.5 End of Experiment

On the first opening of the app on day 28, so the last day of the experiment, the participant was greeted with a popup message indicating the upcoming end of the experiment, asking to add all missing entries of the habit performance. On day 29, so after the experiment has finished, the participant was presented with a thank you message, indicating the end of the experiment and asked to respond to a final questionnaire. This questionnaire was identical to the one at the beginning of the experiment, sequentially running through all cards again, indicating the preference per card. After this, the experiment had ended. While the participant was able to continue using the app, any further use or entered data did not affect the experiment anymore.

5.8 Statistical Analysis

All analyses of the data were done in R, version 4.1.0. Preprocessing of the participant's data logs into formats that were more accessible for the use in R was done in Python, version 3.8.5. Both the R Markdown files and the Python-script are available as supplemental material.
Multiple descriptives were reported. For all descriptives, the mean, standard deviation and number of observations is reported. Descriptives include the outcome measures and age. It also includes the pre- and post-experiment reported measures of motivation to reach the daily goal, self-reported difficulty to reach the goal and the post-experiment indicated perceived helpfulness of the app in reaching the goal. Further, the level of education as well as the nation of origin was reported, but only limited to the number of times a certain level of education or nation was mentioned.
It was checked whether the two measures of interest in the motivational cards, the Card Questionnaire and the User Behavior Measure, are correlated. To check this, both ratings were standardized and Pearson's R was used to determine the level of correlation between the participant's measures. Then, to disentangle the effects of tailoring conditions on the outcome variables, a complex model (LME/GAMM) was intended to be fitted to the data. A forward selection procedure was used, which besides the condition-based tailoring mechanisms $tailorUB$ and $tailorQuest$ and its interaction term included prospective confounders in the data. However, as no model improved over the base model ($outcome \sim tailorUB * tailorQuest$), the reported model was a basic linear model, including interaction.
Additionally, the effects of self-reported motivation to achieve the daily goal, as well as perceived difficulty of the goal on the outcome measures was tested using simple linear models. Both measures were reported by participants pre- and post-experience on a likert-scale. The same was done for the effect of age on outcome.
For further insights on the suitability of using the measures to tailor the features towards the individual users, post-hoc tests were applied. This investigated how well for each user the three most favored motivational features according to the questionnaire predicted the three most favored features according to user behavior. A two-proportions z-tests was applied two times. First, it was applied comparing the ratio of matching based on the individual user ratings to the random matching rate. Second, it was used to compare the ratio of matching between the top 3 favored features over all participants in the questionnaire and the top 3 per user according to the UB measure.

Further tests were applied to investigate how effective the app supported users in their habit formation. One way to assess how well users were motivated to continue using the app is the retention rate. This is captured by the fraction of users continuing usage of the app at different points in time after installing the app. The basis for these calculations are the 33 accounts which were created inside the app. Retention rates were compared for the days 1, 7 and 28 to retention rates of other Android apps, although the last had to be compared to the retention rate after day 30 according to the two sources, as no data for day 28 was available.
Habits and their measures were both checked on their variety and outcome performance differences. Informally, the variety of habits and measures are reported. Informal groups of different types of habits were created post-hoc by the researcher, as no formalized way of grouping was possible beforehand. Notifications were investigated informally on how well they worked technically and how well they conveyed the habit names in the messages. After it turned out that not all notifications were presented to users, a linear model was used to check whether the ratio of shown notifications has an influence on the outcome variables. Finally, linear models were fitted to check for systemic influences of variables like age, motivation or difficulty on the success of participants. Finally, a linear model checked the effect that the amount of user interactions with the motivational features had on the ratio of success.

6 Results

In total, 33 accounts in Routinary were created. However, only 26 users (78.8%) finished the full setup process, including entering the habit and goal, as well as finishing the tutorial. 7 users (21.2%) did not finish the setup, either because they never actually logged into their account after creating it, or because of quitting during the setup process. An additional 5 users (15.2%) finished the setup, but had no recorded activity beyond this and never opened the app beyond the setup procedure, which left 21 participants (63.6%) with some recorded activity data for the further analysis.

6.1 Demographics

Out of the 21 participants, 19 at least partially responded to the optional demographic survey. Ten of the participants (52.6%) identified as female, 8 (42.1%) as male and 1 (5.3%) as other. The median age was 24, while the mean age was 25.2 years. Age spanned a range from 17 to 62 years. A wide variety of nationalities were represented. Six participant indicated that their nationality was German, five Dutch, two Belgian, two Spanish and three from other European countries. One participant indicated Vietnam. 18 participants reported their educational level. Eight participants were graduates, six undergraduates, three high school graduates and one was a middle school graduate.

6.2 Outcome Variables

Four outcome measures are used to indicate the success in regards to app usage. Last Day refers to the last recorded day of app usage, reflecting how long users were motivated to continue using the app. The Ratio of Success is the amount of activity entries where the self-set goal was reached, relative to the total amount of entries. Next, the Number of Entries represents how many entries the participant has made over the course of the experiment, irrespective of whether those were successful. Finally, the Number of Logs conveys the activity regarding the use of motivational features. Every user interaction captured for the User Behavior-measure contributes to this, without any weightings applied. This measure is both used as an outcome measure, as well as a predictor, therefore being distinct in concept from the other outcome variables.
A matrix depiction of the daily entries per participant over the course of the 28 days-long experiment is depicted in Figure 13. This includes an indication of whether the self-set target by the participant was achieved or not. It is visible that most participants did not enter their data for the majority of days and some users seemingly only entered the data whenever the target was achieved.

6.2.1 Distribution of Outcome Variables

The distribution of all outcome variables amongst participants are shown in Figure 14. The number of daily entries ranged from 0 to 27 out of the full 28 days (M=7.29, SD=8.79, n=21). Visually inspecting the distribution shows that it does not seem to be normal, but either positively skewed or an exponential distribution, although the small sample size limits any proper conclusions. The user interaction with the app, as indicated by the number of logs ranged between 0 and 140 interactions during the experiment (M=30.52, SD=36.39, n=21). The ratio of successful entries again ranged from 0, meaning no successful entries, to 1, meaning only successful entries (M=0.56, SD=0.41, n=18). As visible in Figure 14 (c), 6 participants only entered their data whenever they fulfilled their goal, while five participants never achieved their goal at all. Further, three participants never entered any data, leading to no ratio being represented for them. The remainder of the participants ranged somewhere

Figure 13

Daily Entries by Participants

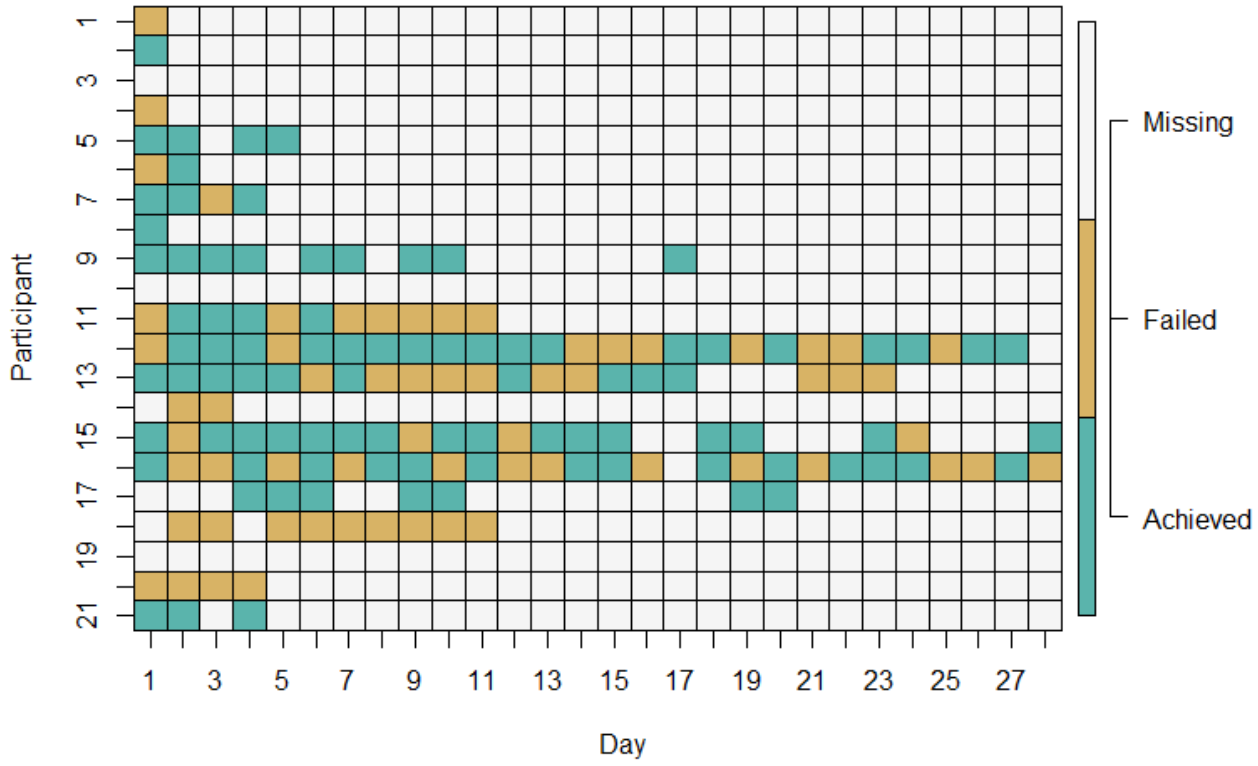
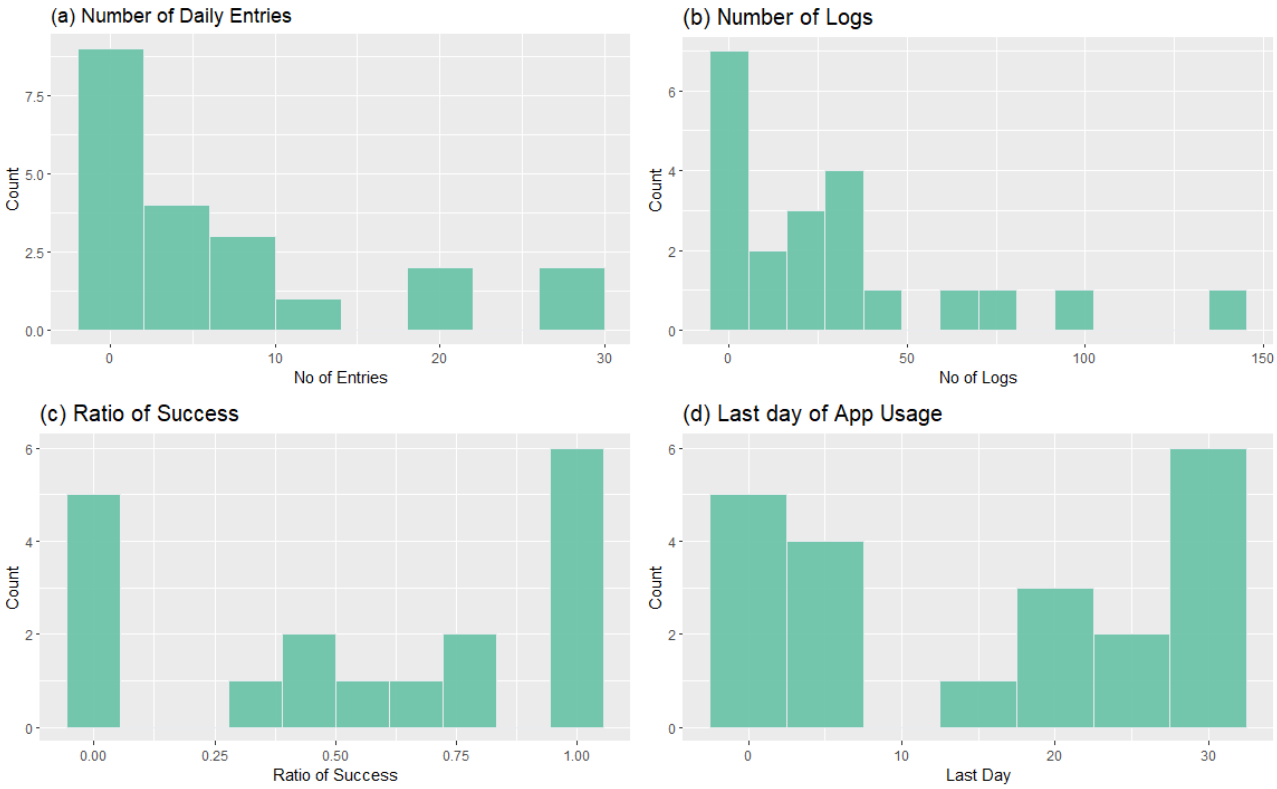


Figure 14

Distributions of Outcome Variables



Note. Histograms of the distribution of outcome variables amongst participants, which are: (a) The number of daily entries in the app, (b) The number of user interactions logged inside the app, (c) The ratio of success, which is the amount of successful entries relative to the amount of unsuccessful entries. Finally, (d) shows the last day of app usage, before the participant did not open the app anymore.

Table 3

Correlation Table of Dependent Values

Dependent Variable	*n*	*M*	*SD*	1	2	3	4
1 - Last Day	21	15.24	11.62	-			
2 - No of Entries	21	7.29	8.79	.00***	-		
3 - Ratio of Success	18	0.56	0.41	.46	.75	-	
4 - No of Logs	21	30.52	36.39	.00**	.00**	.29	-

Notes. $*p < .05$. $**p < .01$. $***p < .001$.

between .36 and .8 in their ratio, indicating a mix of successful and unsuccessful entries. This could be an indication of a multimodal distribution, although again the sample size does not allow for any further conclusions. The final measure is the last day of app usage, day 28, representing using the app until the end of the experiment. This ranged from 1 to the full 28 days, visible in Figure 14 (d) (M=15.24, SD=11.62, n=21). It can be summarized that a high number of early dropouts in the first couple of days is followed by a relatively stable phase of the remaining participants, with a quite steady dropout rate over the later duration of the experiment. In total, 5 participants finished the experiment, meaning they logged in on the last day. The distribution here does again not indicate normality and might be multimodal, but no clear trend emerges.

6.2.2 Correlations of Outcome Variables

The correlation table, showing the correlations of all outcome variables with each other is presented in Table 3, together with the descriptives of each variable. As expected, the use of overlapping concepts lead to high correlations between the outcome variables, with the exception of the ratio of success, which does not correlate significantly with any of the other measures.

6.3 User Statistics

Of the 21 participants that started the experiment and recorded some activity, only five finished the experiment. For the purpose of this experiment, the condition for finishing was to at least once login after the 28 days are over to receive the message about the experiment being over, as well as filling out the short post-experiment questionnaire. One additional participant opened the app after the experiment was finished, but did not fill in the post-experiment questionnaire. All other 15 participants stopped opening the app at some point during the experiment.

6.3.1 User Retention Rates

The user retention rates during the experiment are summarized in Table 4, compared to the values reported for Android apps in general, as well as health and productivity apps specifically (A. Chen, n.d.; Statista, 2020). The retention rate for more than one day of usage was 63.6%, with attrition due to not finishing the setup, not logging in at all, or not logging in anymore after the first day. Although no statistical test can be applied and therefore no significant difference can be inferred, this

Table 4

User Retention Comparisons

Retention Rates	Android Apps	Health Apps	Productivity Apps	Routinary
1 Day	29.2%	20.2%	17.2%	63.6%
7 Days	17.3%	8.5%	7.2%	39.4%
28/30 Days	9.6%	4%	4.1%	18.2%

Note. User retention rates during the experiment compared with results from an investigation of user retention in Android apps (A. Chen, n.d.), as well as specifically for health and productivity apps (Statista, 2020). Retention rates after 1 and 7 days are compared, as well as the results after the full 28 days, which are compared to 30 days from the other sources, as no data from 28 days is available.

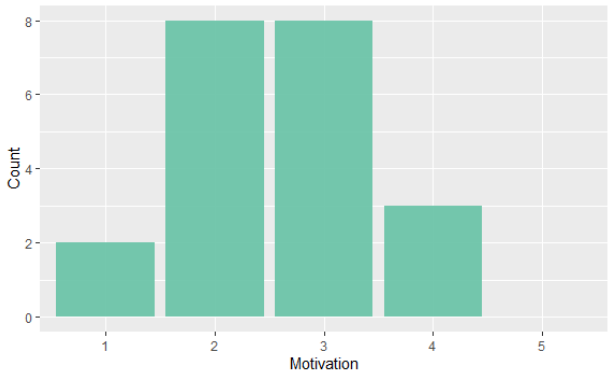
is a much higher retention rate than what has been reported for android apps in general, and for health and productivity apps especially. This pattern is also visible for the 7 day retention rate, which at 39.4% is notably higher than the reported comparative values. Of those users who were retained after day 1, 61.9% remained after day 7, also outperforming the retention rates from the other sources. Finally, while the two sources report a 30 day retention, they can only be compared to the full 28 days retention of the experiment, which at 18.2% is about twice as high as the 30 day retention rate for android apps in general and even more than four times the retention rate reported for health and productivity apps after 30 days. Relative to the users remaining after day 7, 46.2% of users were still retained after day 28.

6.3.2 Motivation

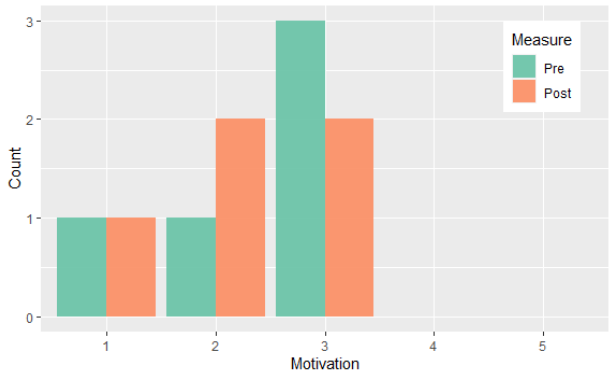
Figure 15

Motivation Ratings

(a) *Pre-Experiment Motivation*



(b) *Pre- vs. Post-Experiment Motivation*



Notes. Motivational ratings on a Likert-scale from 1 to 5. **(a)** Measured before the experiment (n=21). **(b)** Measured before and after the experiment, limited to participants which finished the experiment and therefore provided both measures (n=5).

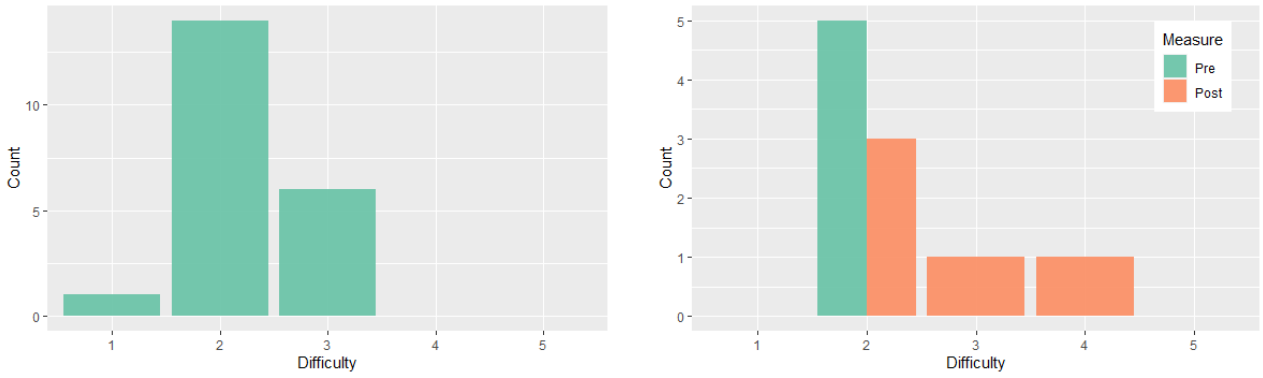
Motivation ratings at the start of the experiment are depicted in Figure 15a, while a comparison of the motivation post-experiment to the pre-experiment ratings limited to those who provided both measures are shown in Figure 15b. Pre-experiment mean motivation for all participants was 2.57 (*SD*=.87, *n*=21) on a scale from 1 to 5. Post-experiment motivation for the participants who finished the experiment was 2.2 (*SD*=.84, *n*=5), which was slightly lower than the pre-experiment ratings for the same group of participants at 2.4 (*SD*=.89, *n*=5). Due to the low number of participants finishing the experiment, no statistical conclusions can be drawn from the difference between both ratings.

6.3.3 Difficulty

Figure 16

Difficulty Ratings

(a) *Pre-Experiment Difficulty* **(b)** *Pre- vs. Post-Experiment Difficulty*



Notes. Difficulty ratings on a Likert-scale from 1 to 5. **(a)** Measured before the experiment (*n*=21). **(b)** Measured before and after the experiment, limited to participants which finished the experiment and therefore provided both measures (*n*=5).

Similar to the descriptive results for motivation, the perceived difficulty ratings of achieving the self-set goal before the start of the experiment are visible in Figure 16a, while the post-experiment ratings are compared to the same group of finishing participants in Figure 16b. Overall pre-experiment difficulty mean was at 2.24 on a scale of 1 to 5 (*SD*=.54, *n*=21). The mean for the post-experiment difficulty was 2.6 (*SD*=.89, *n*=5), which was higher than the indicated pre-experiment difficulty by the same group of participants at 2 (*SD*=0, *n*=5).

6.3.4 Helpfulness

The perceived helpfulness of the app was only reported after the end of the experiment, again on a scale from 1 to 5. Ratings are visualized in Figure 17. The mean helpfulness was 2.6 (*SD*=1.14, *n*=5).

6.4 Habits and Measures

The full list of habits and their respective measures used by the participants in the app is shown in Table 5. The 21 participants specified 14 different activities as their desired habits. In total, six different types of measures were used by the participants. While there is no straightforward way to group the

Figure 17

Post-Experiment Ratings of Perceived App Helpfulness.

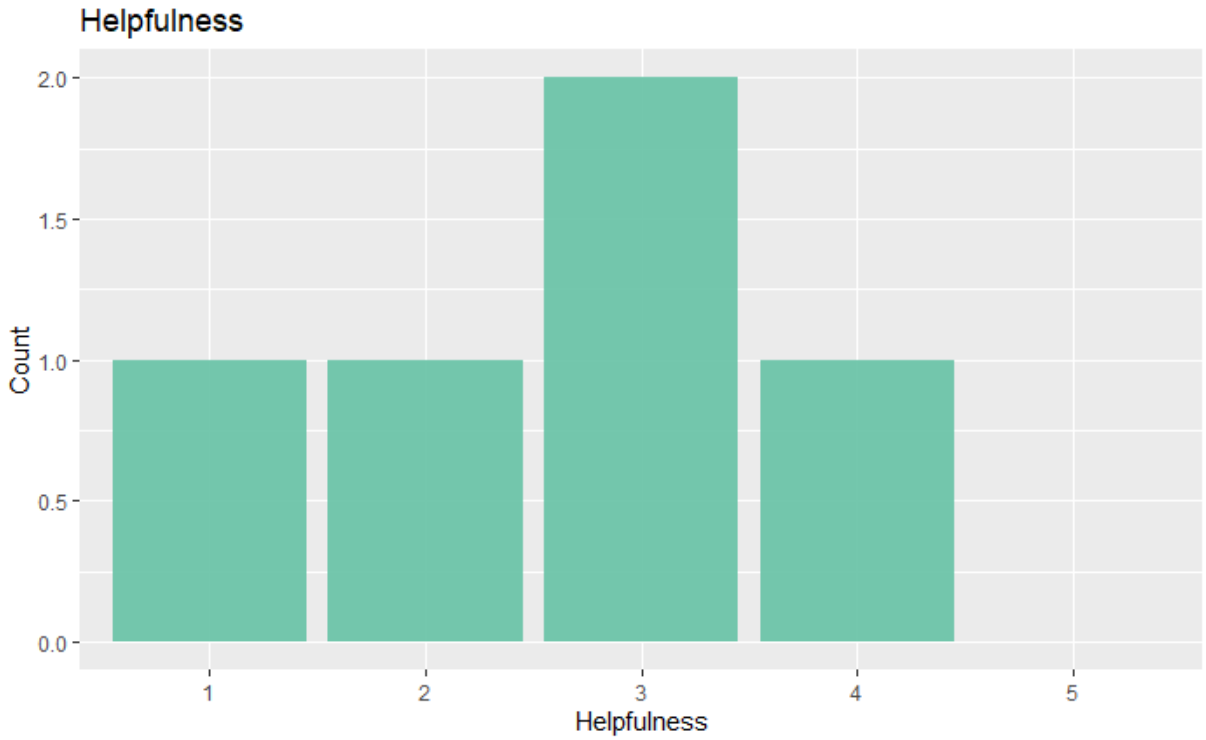


Table 5

Habits and Measures Used by Participants

Habit Group	Habit	Measure	Count
Arts	Guitar	min.	2
	Writing	min.	1
Sports	Exercising	min.	5
	Being Outside	min.	1
	Biking	km	1
	Running	km	1
	Walking	min.	1
Other	Awake bef. 10	min.	1
	Flossing	x	1
	Japanese	min.	1
	Meditating	min.	1
	Reading	pages	2
	Speaking out	x	1
	Drinking water	l/glasses	2

activities, a clear majority relates to sports or moving, in total 9 participants (42.9%) entered this as their goal. Another three participants pursued an artistic habit activity (14.3%). Nine more participants (42.9%) entered habits drawn from miscellaneous categories, for example learning, bodily and mental health, influencing the daily structure or even a habit relating character development. Due to the small sample size, no statistical tests comparing the habit formation success between habit groups were carried out.

6.5 Notifications

6.5.1 Notification Test

In the setup process, two users reported that the notification test was unsuccessful, indicating that the test notification was not visible to them. However, for the first user, multiple logs of scheduled and shown notifications were recorded, including one which was tapped and opened, an interaction which requires the notification to be shown. The second user as well had multiple logs for notifications which were scheduled and shown. It is therefore likely that notifications were visible for those two participants as well and that the negative indications were a consequence of a delay in the delivery of the test notification, or a misclick on the part of the participant. However, this can not be finally established.

6.5.2 Notification Settings

As users were able to turn the notifications off, it was investigated how many participants chose this option and what plausible causes for their decision might have been. Three users turned the motivational notifications off at some point during the experiment. While one of them was never shown a notification, two had 18 and 21 notifications shown to them respectively. In both cases, turning off the notification had been preceded by a period of days in which they did not enter any data. This might imply that a general lack of interest in the app caused the deactivation of the notifications. More participants might have chosen to eliminate notifications by fully removing the app from their phone, but it cannot be investigated whether the stopping of usage is related to deinstalling the app, or just not opening the app while still receiving further notifications.

6.5.3 Inconsistencies in the Presentation of Notifications

Investigating the notification logs indicated that notifications were not always shown. More precisely only 59.3% of days, notifications were shown, ranging from 0% to 100% between users, (M=59.3, SD=34, n=17). This statistic includes those users who turned notifications off, but it has also been reported by participants that notifications suddenly stopped working for multiple days, only to start showing again later. This indicates that there were either issues relating to the app functionality, preventing the notifications to be shown, or issues with the device or Android Alarm Manager, responsible for scheduling and showing the notifications. This can not be clearly determined, neither what influence this had on user behavior. To check whether the amount of notifications shown had an effect on the outcomes, linear models were fitted to test the effect of the ratio of notifications on the outcomes. Users who received fewer motivational notifications had a lower ratio of successful entries than those who received more notifications, $F(1,13) = 3.11$, $p = .10$, $R^2 = .19$, $R^2_{adjusted} = .13$. Additionally, they recorded less daily entries over the course of the experiment, $F(1,15) = 2.16$, $p = .16$, $R^2 = .13$, $R^2_{adjusted} = .07$. Both trends, however, were not significant.

6.5.4 Notification Content

An informal investigation into how well the technical implementation of the notification content tailoring worked was done, as the content was based on the user-chosen habit nouns, verbs, as well as units and some other features, like the self-set rewards. Some issues with this arose, as the entered terms did not always make sense in all notifications shown to participants. All examples that are shown here have the user-names replaced with the term "Name", and the tailored content is highlighted in italic.
It was obvious that sentences were not always smooth or grammatically sound. Two examples for the differing smoothness of motivational messages based on self-set rewards are:

> "Name, today is the last day before you can finally retrieve your reward of *You Can Drink A Special Beer*! One final push and you're there!"

> "Name, only 2 days to go until you can retrieve your reward of *The Last Bag Of Popcorn*! Go, get it!"

Other sentences turned out to be fully grammatically incorrect:

> "Name, on average you've *write* 18min. each day over the past week! This is only 45% of your goal, but you still have time to improve!"

However, even in case of unsmooth or grammatically incorrect sentences, the content of notifications was still understandable.

6.6 Motivational Features

The comparison of standardized ratings of all motivational cards according to the measures of user behavior and the card questionnaire is shown in Figure 18. The questionnaire ratings reflect the pre-experiment ratings, while the user behavior measure reflects all interactions that were made over the course of the experiment. To investigate whether the two preference measures might be suitable for approaches of tailoring, two more post-hoc analyses are presented. First, it is checked whether the two measures are correlated. Second, it is tested whether the top three preferred cards according to the questionnaire are able to predict the top three preferred cards according to user behavior.

6.6.1 Correlation between Questionnaire and User Behavior(UB)-Feature Ratings

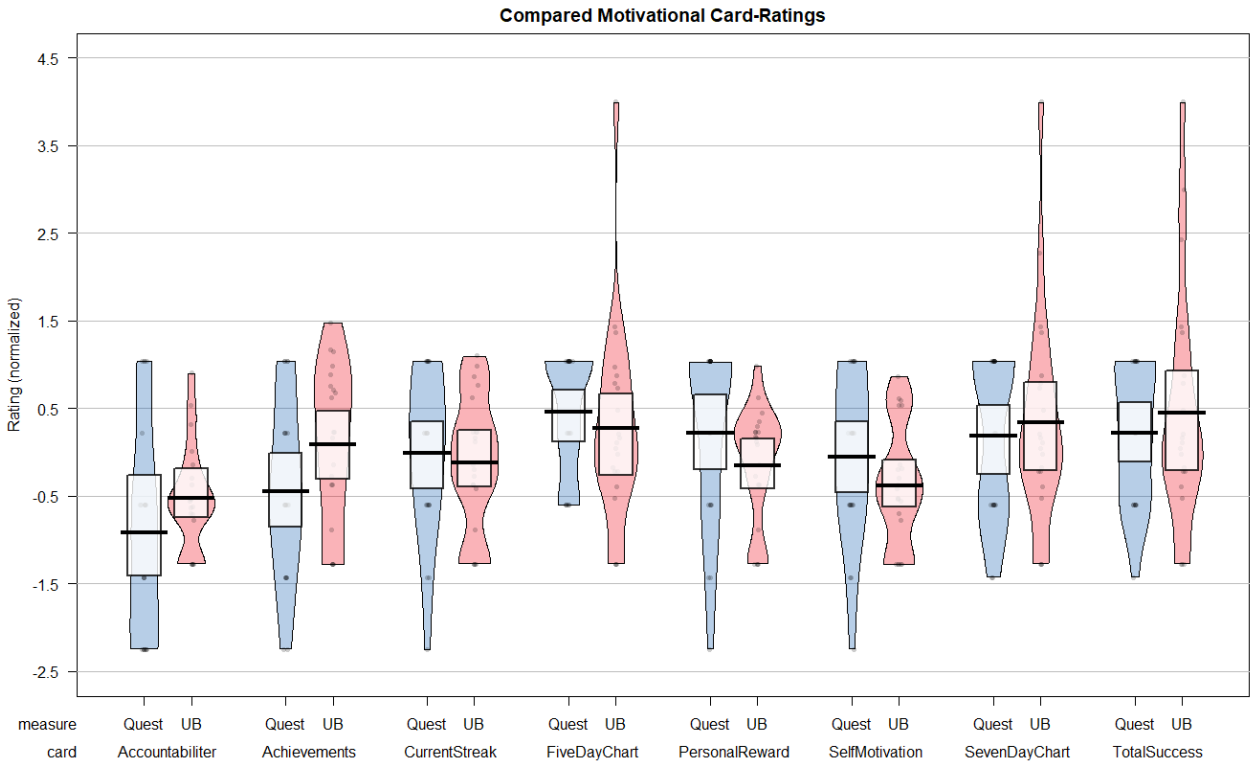
The relationship between the standardized ratings in the questionnaire and the ratings based on the User Behavior are visible in Figure 19. A Pearson correlation showed that this relationship indeed is significant, $r(152) = .3, p < .001$. This indicates that the questionnaire ratings are a significant predictor of the actual user behavior.

6.6.2 Predicting Top 3 Features per User Based on Questionnaire

The intended use of the preference measures is to extract the favorite motivational features of the user. Therefore, the match of the top 3 preferred motivational features according to both measures was additionally investigated. This is done per user. First, the three highest rated motivational cards of each user according to the questionnaire are extracted. Then, the three preferred, or most used motivational cards according to the user behavior measure are extracted as well. The proportion of

Figure 18

Standardized Ratings of each Motivational Card



Notes. Ratings both based on the Questionnaire (Quest) and the User Behavior-Tracking (UB). Ratings include computations from all participants, not just from those who were assigned to the respective tailoring conditions. Ratings of each measure are normalized (z-scores) for better comparison. Two participants are not shown as they gave all cards the same rating, $n = 19$.

Figure 19

Relationship between Ratings

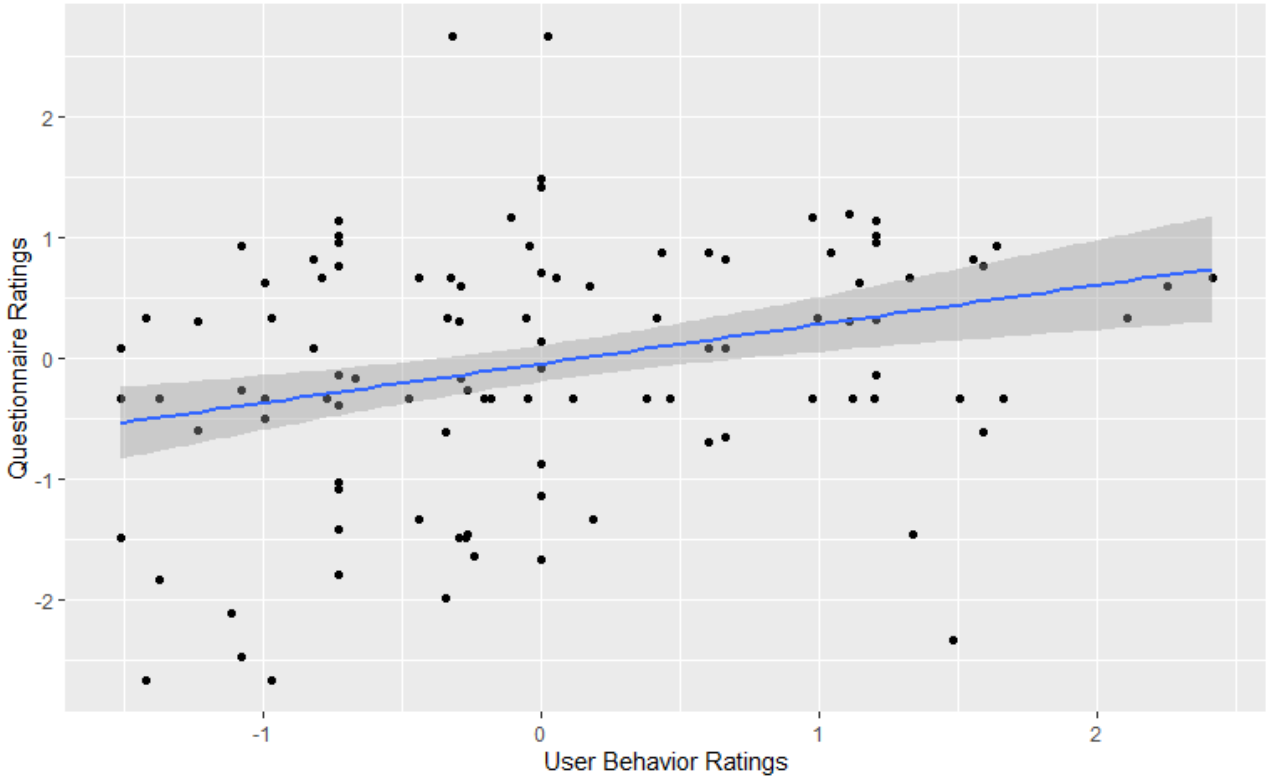


Table 6

Sample Sizes per Condition

Tailoring	UB	Not UB
Quest.	3	4
Not Quest.	9	5

Notes. UB = Tailoring to the User Behavior, Not UB = Not Tailoring to UB. Quest. = Tailoring to the Pre-Experiment Questionnaire, Not Quest. = Not Tailoring to the Questionnaire.

matching cards refers to how many of the three highest rated cards in the questionnaire are also in the three most used according to the user behavior. For example, if two of three cards for one participant match, this would translate to a 66.6% matching rate. For comparison, the proportion of matching cards if predicted randomly would be 37.5%. A Two-Proportions Z-Test was applied to compare the actual matching rate of all participants to the random prediction of motivational cards. The proportion of matches between the tailoring conditions turned out to be significantly higher at 55.6% (35/63), $\chi^2 = 3.46$, $df = 1$, $p = .031$.
However, as was visible in Figure 18, some motivational cards are overall rated higher than others. Therefore, there are two possible explanations for the higher prediction ability of the questionnaire: One would be that it is capable of capturing the individual differences regarding motivational card preferences. However, it is also possible that some motivational cards are generally more likely to appear in both measures as they are on average rated higher, regardless of individual differences. Therefore, the proportion of individual ratings was additionally compared to a second proportion, covering the overall top 3 used ones. This was again lower than the individual tailoring matches at 42.9% (27/63), but the difference turned out not to be significant, $\chi^2 = 1.56$, $df = 1$, $p = .106$. .

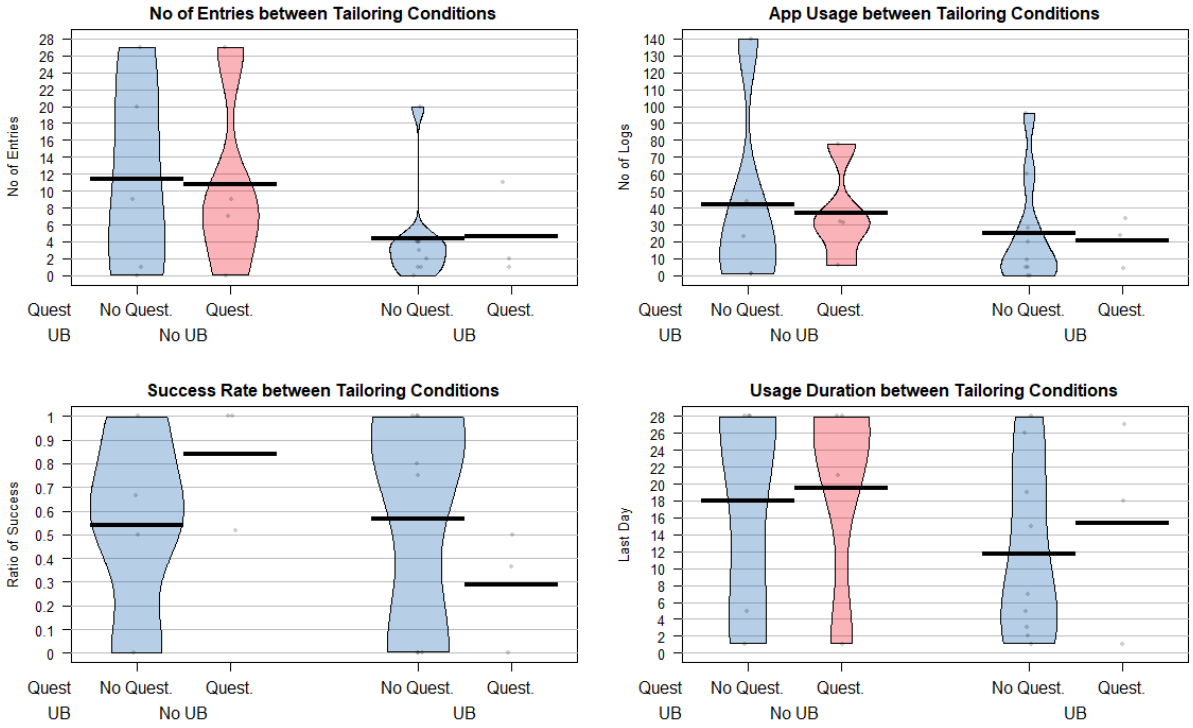
6.7 Hypotheses

6.7.1 Effects of Tailoring on User Performance

Due to the large amount of participant dropout and the random allocation within the app to the four different conditions, the sample size per condition turned out to be very uneven, as indicated in Table 6. The largest difference here was the difference between the UB-Questionnaire condition, with only one third of the participants allocated to the UB-Not Questionnaire condition. Additionally, any investigations into differences between the conditions were hampered by the overall small sample size of 21.
As explained before, two mechanisms were supposed to be affected by the tailoring conditions: The selection of the second and third card shown on the main screen, as well as the content of the motivational notifications shown to participants. However, in the post-analysis it turned out that due to technical issues one of the mechanism did not work as intended: The tailoring of the two cards on the home page was based on the card ratings in the questionnaire for all participants, instead of the respective tailoring mechanism intended. The tailoring for the notification content worked as intended, however, as shown above, the notifications were not always visible to participants and some of them turned them off intentionally.
Comparisons of the 2x2 design results comparing the tailoring conditions for all four outcome condi-

Figure 20

Differences between Tailoring Conditions on Outcomes



tions are visible in Figure 20. It is immediately apparent that not enough participants took part to draw any conclusions. In an attempt to fit a suitable model to the data, no model outperformed the base model of $outcome \sim 1$. The results of the linear model testing for differences between the experimental conditions and therefore an effect of tailoring including the interaction between both mechanisms, $outcome \sim tailorUB * tailorQuest$, is shown in Table 7. None of the effects were significant.

6.7.2 Effect of Feature Use on the Ratio of Successful Entries

A simple linear regression model tested the effect of User interactions, as indicated by the number of logs per user, on the ratio of success. To restrict outlying effects of users whose success ratio might have been skewed by only a small amount of entries, only users with at least 4 daily entries were included in the analysis, $n = 11$. However, the reported trends also held up when all participants were included in the analysis. The model itself did not return significant, $F(1,9) = 1.7$, $p = .22$, $R^2 = .16$, $R^2_{adjusted} = .07$. However, a trend was visible in the regression coefficient, $B = .0038$, also visible in the visualized resulting linear model, see Figure 21. This indicates that the more users interact with the motivational features, the more successful they are in pursuing their goal.

6.7.3 Further Investigations on Outcomes

Further effects, like the effect of motivation, difficulty or age on the outcome measures were investigated, with results summarized here. The full results, including models and visualizations of those can be found in Appendix D. For most of these, the sample sizes were too low to be able to draw any statistically significant conclusions from any of the models, as well as being able to properly

Table 7

Interaction Model of Tailoring Conditions on Outcome

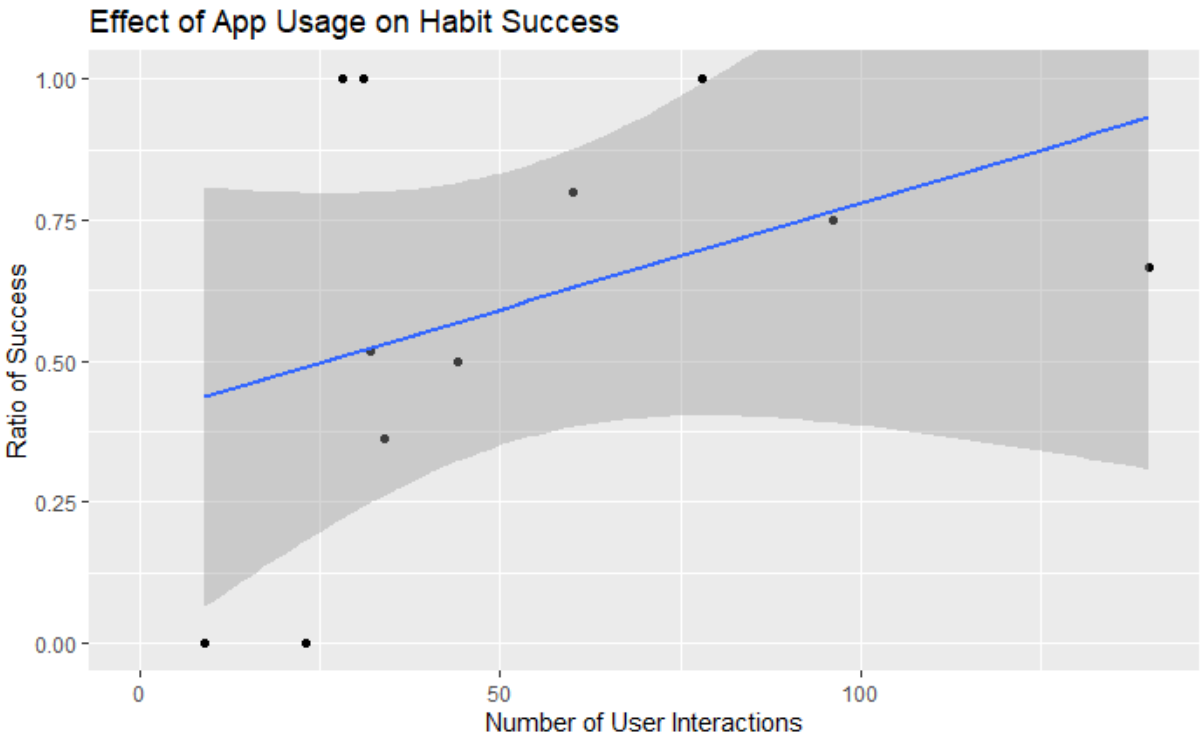
	Dependent variable:			
	Last Day	No of Entries	No of Logs	Ratio of Success
	(1)	(2)	(3)	(4)
TailorQuest	1.5	−0.7	−5.0	0.3
	(8.1)	(5.9)	(25.8)	(0.3)
TailorUB	−6.2	−7.1	−17.0	0.03
	(6.7)	(4.9)	(21.4)	(0.3)
TailorQuest:TailorUB	2.1	1.0	0.9	−0.6
	(11.4)	(8.3)	(36.3)	(0.4)
Constant	18.0***	11.4***	41.8**	0.5**
	(5.4)	(3.9)	(17.2)	(0.2)
Observations	21	21	21	18
R^2	0.1	0.2	0.1	0.2
Adjusted R^2	−0.1	0.000	−0.1	−0.02
Residual Std. Error	12.1 (df = 17)	8.8 (df = 17)	38.4 (df = 17)	0.4 (df = 14)
F Statistic	0.5 (df = 3; 17)	1.0 (df = 3; 17)	0.3 (df = 3; 17)	0.9 (df = 3; 14)

Notes: *p<0.1; **p<0.05; ***p<0.01

The model fitted was $Outcome = TailorQuest * TailorUB$, investigating the effect of the tailoring conditions including their interaction on the outcome variables. None of the models have any significant coefficients and the inclusion of further variables, or the extensions to more complex models were able to improve the fit.

Figure 21

Effect of App Usage on Habit Success

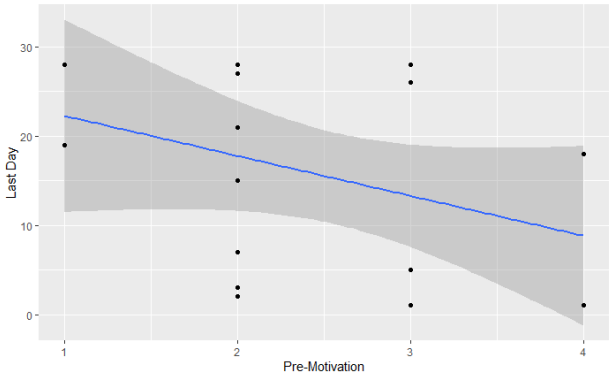


investigate whether the assumptions for the appropriateness of fitting linear models were satisfied. Especially all post-experiment measures had a very low number of subjects (n = 5), therefore not providing reliable insights. However, some trends were visible which are considered worth noting.
One interesting, albeit insignificant trend was visible regarding the effect of pre-experiment motivation on all outcome measures (n = 21). As the example in Figure 22a shows, a higher motivation to pursue the self-set goal as reported before the experiment was related to lower usage duration. This trend also held for all other outcome measures. The post-experiment measures (n = 5) reveal the opposite trend: Higher motivation was related to a higher success in the app usage, as well as higher intensity of usage.
The effect of the perceived difficulty of the habit goal showed a trend for all four outcome measures, both for the pre- and post-experimental measures. Higher difficulty was related to lower success, lower duration of usage and lower intensity of usage. No statistical significance was reached, though.
Finally, the linear model of the effect of age on all outcome variables showed a positive trend for all four measures. However, only the effect of age on the ratio of success, shown in Figure 22 turned out to be significant, $F(1,14) = 19.56$, $p = < .01$, $R^2 = .61$, $R^2_{adjusted} = .58$. The regression coefficient, $B = .10$, revealed that an increased age was linked to a higher level of success.

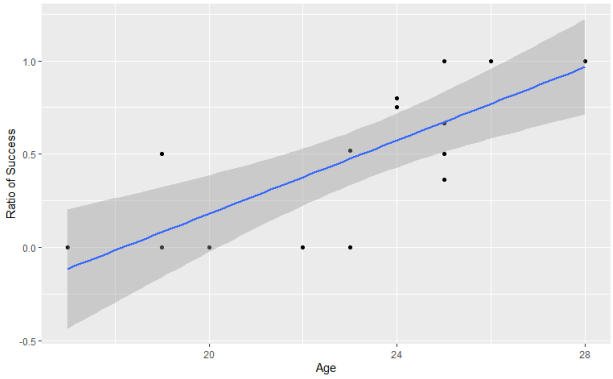
Figure 22

Card Questionnaire

(a) *Effect of Pre-Motivation on Last Day*



(b) *Effect of Age on the Ratio of Success*



Notes. Example of the card questionnaire, with the elements of the card itself, a short description, an optional check to indicate interaction with the card and the 5-point likert scale question to indicate interest in the card.

7 Discussion

The goals of this research have been to: 1. Establish a framework in form of an app which allows to generally investigate influences of app features on outcome and success in habit and behavior change apps. As part of this, it was explored how well the app supported different types of habits and measures, both in regards to the general app functionality and the utilization of motivational features supporting users. 2. Develop, implement and inspect concrete tailoring mechanisms that adjust the content of motivational features to user preferences, as indicated by user ratings and measures of user behavior.
Multiple limitations can be assumed to affect all results and possible conclusions that could be drawn from them. First of all, the low number of participants was further worsened by a comparably low average number of daily entries. Also, users differed in their patterns of entering data and abandonment rates were quite high. Some of the participants only inputted their progress whenever they achieved their goal, while others seemed more open to also input data on days were they failed achieving their target. The already low number of participants was further worsened by the imbalanced assignment to the four experimental conditions, leading to one condition having only three participants assigned to it. Further, technical issues lead to the card tailoring mechanism not working. Notifications, which were elementary for delivering the content tailored based on experimental conditions, also did not work reliably for every participant. Ultimately, these limitations make it impossible to assure that the assumptions for the statistical models fitted on the data are satisfied and that the trends which were observed can be generalized to the whole population.

7.1 Did the Tailoring Mechanisms Work?

The assessment of tailoring conditions was hampered by the limitations mentioned above. It is therefore not a surprise that no differences in any outcome measures between the four tailoring conditions were found and no indicative trends were observed.
Post-hoc investigations into the effects of motivational features and the two measures used for tailoring towards them however, led to some relevant insights into both how well the measures worked, as well as the suitability of the approach for further research. It was found that questionnaire-ratings of the motivational cards and the user behavior-based ratings correlated significantly and highly. This shows that the results of the questionnaire predict user behavior. The results also show that differences in the preferences of motivational features exist, something already explored in previous research with other measures (Hamari et al., 2018).
Due to the limited amount of available data, the idea of segmenting the user behavior-based data into four different weeks was abandoned. Instead, the analysis was based on all data of the experiment, disregarding any changes in user preferences that might occur over the course of the experiment. If such change in preferences exists, the correlation found might actually underestimate the overlap in measured construct between the two preference-measures. This is because some of the relation will be masked by the changed preferences in the user behavior measure towards the end of the experiment. However, this cannot be fully established within the limitations of this experiment.
Finally, the goal of the tailoring is not to maximize correlation between the measures, but to be able to find the favourite features of each user. Therefore, one post-hoc test checked whether the user's favorite 3 features according to the questionnaire are able to predict the favorite 3 according to User Behavior. This turned out to be significantly better than selecting features at random. While the prediction using the individual user's top 3 features missed significancy when compared to the overall top 3 rated features, it still scored higher. It cannot be said with all certainty whether tailoring to the

user's individual preferences is definitely the best option, or whether some cards are simply overall preferred, but based on the observed trends one can suggest that both effects exist. This can be expected, as both individual differences, but also overall differences in which motivational features work better on the overall population have been observed repeatedly (Villalobos-Zúñiga & Cherubini, 2020; Hamari et al., 2018; Morrison, Yardley, Powell, & Michie, 2012).

7.2 Did the App Increase User Success?

There were multiple indications for a positive effect of app usage on habit success overall. Productivity and health apps similar to Routinary have an even lower user retention rate than average smartphone app, which is already considerably low (Statista, 2020; A. Chen, n.d.). Promisingly, the retention rates of users over the course of the experiment far outperformed what has been reported in the literature, even up to about four times as many users were retained as is usual. This can be an indication of a successful motivational effect of the app on users, drawing more users into long-term usage than is the case with common, average apps in this category. However, it has to be mentioned that the setup of the experiment likely confounded these numbers. The fact that the app was part of an experiment meant that the download of the app itself already was more of a hurdle than the usual one-click download in a pre-installed app store on the phone. This might have lead to some users with low motivation perceiving the download itself as too energy- and time-consuming, therefore not being covered by the statistic as dropouts. In case of a normal app store, they might have just downloaded the app easily and then dropped out soon later, therefore being covered by the user retention as an early dropout. Additionally, the types of users in the experiment might not reflect those that constitute the population of overall smartphone users. Due to the complex download and the pathways used to advertise the experiment, the user base was highly educated and therefore likely of higher socioeconomic stand than the general public. It has been shown that health interventions are less successful for people with lower socioeconomic stand, especially when participation in those interventions is voluntary (White, Adams, & Heywood, 2009). These factors limit the comparison with retention rates from app stores. However, one thing that likely lowered the retention rate in this app was the convoluted, time-consuming setup process, partially as a consequence of it being part of an experiment. Additionally, some aspects of the app were malfunctioning, potentially further lowering retention rate, as indicated by the insight that a higher rate of notifications shown was linked to higher success, even if insignificantly. While the impact of all of these confounding factors on the user retention is difficult to assess, the high retention rate still can be seen as highly promising. Fixing some off-putting elements, as well as streamlining the setup process could certainly further increase the user retention rate.
It was also observed that users who interacted more with the motivational features tended to have a higher rate of successful daily entries. Despite the aforementioned limitations regarding both measures, this can be a further indication that the motivational features might have lead users to be sufficiently motivated to reach their goal, putting in the extra effort to get an achievement, a streak cup or any other rewarding aspect of the features relevant to the user.

7.3 Did the App Work for different Types of Habits?

Even though only few participants took part in the research, the app itself was used for a high variety of habits. These reached far beyond the classical paradigms of gamification apps, which typically relate to the domains of health and learning (Koivisto & Hamari, 2019). While a common theme indeed was relating sports and movement, there were many niche activities like flossing or speaking

out one's mind. For many of them, it would be unlikely to find specialized apps, as well as impractical to develop these with motivational features tailored to them. This supports the original intention of allowing users to freely set up their own activities and goals, as well as the need to widen research into what motivational features work for which kinds of habits. These conclusions support similar insights from previous research regarding the need for more generalized habit app research (Klock et al., 2020; Asimakopoulos et al., 2017). Routinary might be a valuable tool for both applications. The app architecture allows for the comparison of different paradigms on different habits in a quite controlled way, making it suitable for more thorough experimental research. Additionally, no outcome differences between habit groups were observed. While only a low number of participants were represented and therefore each habit was only followed by very few participants, this does show promise regarding a generalized applicability of the app. Unfortunately, the low number of participants per habit also means that it cannot be reliably investigated whether there are differences in regard to which features worked well for which type of habit.

7.4 Possible Effect of Intrinsic Motivation on Habit Formation

An inverse, or at least non-positive effect of prior level of motivation on habit formation and app usage duration was found. This is contrary to the intuitive idea and assumption in behavior change research that higher motivation leads to higher performance on the goal. However, as motivation was assessed through a simple likert-scale self-report this result has to be interpreted with caution regarding its limited reliability of assessing a multifaceted concept like motivation (Fulmer & Frijters, 2009).

One explanation for the non-positive or negative effect of motivation on the outcome measures is that the motivation measure might be a reflection of the intrinsic motivation by the users, related to the fully self-determined choice of habit in the research. The extrinsic motivation provided by the motivational features of the app might then actually undermine the intrinsic motivation of those users with higher self-reported motivation, a phenomenon well-established in self-determination theory (R. M. Ryan & Deci, 2000). Therefore, the features of the app would not do any better in motivating users with higher self-reported motivation, as any intrinsic motivation is overshadowed by the induced extrinsic motivation. This would imply that the user preferences-based tailoring mechanisms in this research did not sufficiently take into account the user's levels of extrinsic and intrinsic motivation.

A second explanation is that users with higher intrinsic motivation simply do not need the support of an extrinsically motivating app in achieving their goal. Therefore, participants scoring high on intrinsic motivation would abandon the usage of the app, as they perform the habit regardlessly and do not perceive the features to be helpful. This explanation would be coherent with previous research showing that already formed habits are tied to less impactful effects of motivational features, as users are not depending on them for performing the habit. (Yang & Koenigstorfer, 2021).

Finally, it has to be emphasized that the trend observed is not significant and can be a result of the aforementioned limitations of the experiment itself. To disentangle the explanations put forward for this phenomenon, further research is required. This includes investigating the levels of intrinsic and extrinsic motivation in participants both pre- and post-experiment with established tools. In relation to this, measuring the performance on the habit also after abandoning the app would help establishing whether user retention is related to habit performance.

7.5 Recommendations

7.5.1 Development of Habit Apps for Multiple Domains

The use of a large variety of habits by the participants has further emphasized the need to develop habit apps that allow more freedom in their application of habits of different kinds. This implies the need to develop appropriate motivational features as well, which technically and conceptually work for a larger variety of habits. Comparing the suitability of different motivational features for different habits still remains a necessity, further expanding on the promising, but statistically limited results of this experiment.

7.5.2 Improvements to the Routinary App

The generalized way in which Routinary was designed for the implementation of different habits showed a lot of promise. However, besides the obvious technical issues with notifications not being shown reliably and one of the tailoring mechanisms malfunctioning, some conceptual issues arose. Especially tailoring the text content seemed not to work out for all habits. While the context and logic of the content itself seemed to work out well for all varieties of messages, as they were kept quite generic, grammatical errors appeared. Methods of Natural Language Processing (NLP), like Grammatical Error Correction, might be suitable to extract the meaning of the terms and adjust the grammatical form when used in different sentences, avoiding grammatical errors and making the sentence structure more fluent. It would also forfeit the need for users to enter both a noun and a verb for their habit, further simplifying the setup process.
Helpful in this context would also be the creation of a taxonomy of habits, with groupings allowing to further tailor motivational features to different groups of habits and their characteristics. One starting point to this would be the taxonomy of leisure-activities in relation to different psychological needs (Tinsley & Eldredge, 1995), although the existing taxonomy would have to be extended beyond specialized leisure-activities.

7.5.3 Repeat Approach in Established App

Many of the main limitations that impeded the research were due to the own development and deployment of an app. This lead to a small, non-representative user base which went through an uncommon setup process, compared to a regular, production-level app. Therefore, the main recommendation relating to this research would be to repeat the approach taken here, utilizing an established app with a large user base, a common approach in Gamification and habit research (Hamari et al., 2018). This eliminates issues of small sample sizes and therefore makes it possible to experimentally test multiple conditions, as well as drawing proper conclusions from the retention rates, for once because two versions of the app can be compared, but also because the retention rate of the app itself is more suitable to be compared to other apps available in the same app store. It would also ensure that the sample of users reflects the population of app users, contrary to what was observed in the research. The biggest issue will be finding a generalized habit app that actually allows for the approach to be implemented, as most apps are specialized on a limited amount of activities from few domains. This was, after all, the original cause for the motivation to develop Routinary.

7.5.4 Use of Big Data-Based Dynamical Tailoring

In this research, the user interaction-values per card were computed using fixed formulas as there was not enough data to check which of the individual interactions, like opening the card or interacting

with the item, ultimately predict success or are linked to the questionnaire. A larger user base would allow for more elaborate investigations of the algorithms presented in this research. For example, factor analyses could extract the significant individual user interactions more easily, allowing to use the results for tweaking the parameters instead of the rather intuition-based approach taken in the initial design. Further, the inclusion of more user characteristics, for example based on demographics, could further optimize the tailoring and even be used as an indicator for which user interactions are predictive of success for which user group. The impact of each variable in the formulas could then be tailored itself to the users. All aforementioned proposals regarding dynamical tailoring are coherent with recommendations from a recent paper concerning tailoring in gamification (Klock et al., 2020).

7.5.5 Further Exploration of Intrinsic and Extrinsic Motivation

Based on the findings that pre-experiment motivation might have had a negative effect on habit formation using the app it is advisable to further investigate the effects different levels of intrinsic and extrinsic motivations have. This would allow to tailor the features in a way that helps transitioning from extrinsic to intrinsic motivation, depending on the motivational state of the user and the effects of each feature on it, a recommendation matched by previous investigations (Villalobos-Zúñiga & Cherubini, 2020). As some users seemed to drop out or turn off the notifications whenever they received feedback that directly or indirectly made them aware of a lack of current progress, it is also recommended that the effect of need frustration is further evaluated using the appropriate tools (B. Chen et al., 2015).

8 Conclusion and Further Recommendations

This research investigated the suitability of a dynamical tailoring method for tailoring motivational features, based on measures of user behavior and ratings of the features by users. It has been found that ratings by the users predicted their actual usage of different features and were suitable to extract their favored motivational features. However, whether the selection of these improves the outcome of the intervention could not be established. Still, the introduced methods have shown promise in the use of tailoring for further research.
The Routinary app itself showed promise in supporting habit building and it is suggested that improvements are made to the structure and some features. Utilizing the app in the context of normal app stores and streamlining the setup process would allow for a more thorough investigation of its suitability and effects in habit building and behavioral change. Beyond this specific app, this research has shown that implementing approaches like these systematically in already existent and established apps with large user bases is highly recommended for more extensive and reliable data.
Finally, as has been established previously, Gamification research needs to expand the investigated domains beyond the current scope of mostly focusing on learning and health-related behaviors. The variety of activities used for habit building by users of Routinary further strengthens this point. A more structured approach to habit and motivation research would help in generalizing the findings and applying them beyond the narrow applications which are currently the standard. This includes systemizing habits or activities, as well as patterns in how motivational features relate to them and which user characteristics are capable of explaining those. Supporting the transition from extrinsic to intrinsic motivation has to be part of this research as well, making sure that the behavioral change persists beyond the use of the app itself.

References

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health psychology*, *27*(3), 379.

Asimakopoulos, S., Asimakopoulos, G., & Spillers, F. (2017). Motivation and user engagement in fitness tracking: Heuristics for mobile healthcare wearables. In *Informatics* (Vol. 4, p. 5).

Attig, C., & Franke, T. (2019). I track, therefore i walk–exploring the motivational costs of wearing activity trackers in actual users. *International Journal of Human-Computer Studies*, *127*, 211–224.

Chen, A. (n.d.). *New data shows losing 80% of mobile users is normal, and why the best apps do better.* Retrieved from `https://andrewchen.com/new-data-shows-why-losing-80-of-your-mobile-users-is-normal-and-that-the-best-apps-do-much-better/`

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., ... others (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and emotion*, *39*(2), 216–236.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes.

Conroy, D. E., Yang, C.-H., & Maher, J. P. (2014). Behavior change techniques in top-ranked mobile apps for physical activity. *American journal of preventive medicine*, *46*(6), 649–652.

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian psychology/Psychologie canadienne*, *49*(3), 182.

Dijkstra, A. (2014). The persuasive effects of personalization through: name mentioning in a smoking cessation message. *User Modeling and User-Adapted Interaction*, *24*(5), 393–411.

Duolingo. (2021). *Duolingo (version 5.24.3) [mobile app].* Google Play Store. https://play.google.com/store/apps/details?id=com.duolingo&hl=de&gl=US.

Elbert, S. P., Dijkstra, A., & Oenema, A. (2016). A mobile phone app intervention targeting fruit and vegetable consumption: the efficacy of textual and auditory tailored health information tested in a randomized controlled trial. *Journal of medical Internet research*, *18*(6), e147.

Fulmer, S. M., & Frijters, J. C. (2009). A review of self-report and alternative approaches in the measurement of student motivation. *Educational Psychology Review*, *21*(3), 219–246.

Hakulinen, L., Auvinen, T., & Korhonen, A. (2015). The effect of achievement badges on students' behavior: An empirical study in a university-level computer science course. *International Journal of Emerging Technologies in Learning*, *10*(1).

Hamari, J. (2007). Gamification. *The Blackwell Encyclopedia of Sociology*, 1–3.

Hamari, J., Hassan, L., & Dias, A. (2018). Gamification, quantified-self or social networking? matching users' goals with motivational technology. *User Modeling and User-Adapted Interaction*, *28*(1), 35–74.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th hawaii international conference on system sciences* (pp. 3025–3034).

Hawkins, R. P., Kreuter, M., Resnicow, K., Fishbein, M., & Dijkstra, A. (2008). Understanding tailoring in communicating about health. *Health education research*, *23*(3), 454–466.

Hosseinpour, M., & Terlutter, R. (2019). Your personal motivator is with you: a systematic review of mobile phone applications aiming at increasing physical activity. *Sports Medicine*, *49*(9), 1425–1447.

Klock, A. C. T., Gasparini, I., Pimenta, M. S., & Hamari, J. (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies*, 102495.

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, *45*, 191–210.

Krebs, P., Prochaska, J. O., & Rossi, J. S. (2010). A meta-analysis of computer-tailored interventions for health behavior change. *Preventive medicine*, *51*(3-4), 214–221.

Lally, P., & Gardner, B. (2013). Promoting habit formation. *Health psychology review*, *7*(sup1), S137–S158.

Lustria, M. L. A., Noar, S. M., Cortese, J., Van Stee, S. K., Glueckauf, R. L., & Lee, J. (2013). A meta-analysis of web-delivered tailored health behavior change interventions. *Journal of health communication*, *18*(9), 1039–1069.

McKay, F. H., Wright, A., Shill, J., Stephens, H., & Uccellini, M. (2019). Using health and well-being apps for behavior change: a systematic search and rating of apps. *JMIR mHealth and uHealth*, *7*(7), e11926.

Michie, S., Abraham, C., Eccles, M. P., Francis, J. J., Hardeman, W., & Johnston, M. (2011). Strengthening evaluation and implementation by specifying components of behaviour change interventions: a study protocol. *Implementation Science*, *6*(1), 1–8.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, *46*(1), 81–95.

Molina, M. D., & Sundar, S. S. (2018). Can mobile apps motivate fitness tracking? a study of technological affordances and workout behaviors. *Health communication*.

Morrison, L. G., Yardley, L., Powell, J., & Michie, S. (2012). What design features are used in effective e-health interventions? a review using techniques from critical interpretive synthesis. *Telemedicine and e-Health*, *18*(2), 137–144.

Noar, S. M., Grant Harrington, N., Van Stee, S. K., & Shemanski Aldrich, R. (2011). Tailored health communication to change lifestyle behaviors. *American Journal of Lifestyle Medicine*, *5*(2), 112–122.

Oulasvirta, A., Rattenbury, T., Ma, L., & Raita, E. (2012). Habits make smartphone use more pervasive. *Personal and Ubiquitous computing*, *16*(1), 105–114.

Ryan, K., Dockray, S., & Linehan, C. (2019). A systematic review of tailored ehealth interventions for weight loss. *Digital health*, *5*, 2055207619826685.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, *30*(4), 344–360.

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380.

Samsung. (2021). *Samsung health (version 6.18.7.005) [mobile app].* Google Play Store. https://play.google.com/store/apps/details?id=com.sec.android.app.shealth&hl=de&gl=US.

Schmidt-Kraepelin, M., Toussaint, P. A., Thiebes, S., Hamari, J., & Sunyaev, A. (2020). Archetypes of gamification: Analysis of mhealth apps. *JMIR mHealth and uHealth*, *8*(10), e19280.

Statista. (2020). *Mobile app user retention rate worldwide 2020.* Retrieved from https://www.statista.com/statistics/259329/ios-and-android-app-user-retention-rate/

Statista. (2021). *Number of smartphone users from 2016 to 2021.* Retrieved from https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwi

de/

Stawarz, K., Cox, A. L., & Blandford, A. (2015). Beyond self-tracking and reminders: Designing smartphone apps that support habit formation. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (p. 2653–2662). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2702123.2702230 doi: 10.1145/2702123.2702230

Tinsley, H. E., & Eldredge, B. D. (1995). Psychological benefits of leisure participation: A taxonomy of leisure activities based on their need-gratifying properties. *Journal of Counseling Psychology*, *42*(2), 123.

Van den Broeck, A., Ferris, D. L., Chang, C.-H., & Rosen, C. C. (2016). A review of self-determination theory's basic psychological needs at work. *Journal of Management*, *42*(5), 1195–1229.

van Velsen, L., Broekhuis, M., Jansen-Kosterink, S., & op den Akker, H. (2019, Sep 06). Tailoring persuasive electronic health strategies for older adults on the basis of personal motivation: Web-based survey study. *J Med Internet Res*, *21*(9), e11759. Retrieved from https://www.jmir.org/2019/9/e11759 doi: 10.2196/11759

Villalobos-Zúñiga, G., & Cherubini, M. (2020). Apps that motivate: a taxonomy of app features based on self-determination theory. *International Journal of Human-Computer Studies*, *140*, 102449. Retrieved from https://www.sciencedirect.com/science/article/pii/S1071581920300513 doi: https://doi.org/10.1016/j.ijhcs.2020.102449

White, M., Adams, J., & Heywood, P. (2009). How and why do interventions that increase health overall widen inequalities within populations. *Social inequality and public health*, *65*, 82.

Yang, Y., & Koenigstorfer, J. (2021, July). Determinants of Fitness App Usage and Moderating Impacts of Education-, Motivation-, and Gamification-Related App Features on Physical Activity Intentions: Cross-sectional Survey Study. *Journal of Medical Internet Research*, *23*(7), e26063. Retrieved 2021-07-30, from https://www.jmir.org/2021/7/e26063 doi: 10.2196/26063

Zhang, P. (2008). Motivational affordances: reasons for ict design and use. *Communications of the ACM*, *51*(11), 145–147.

Appendices

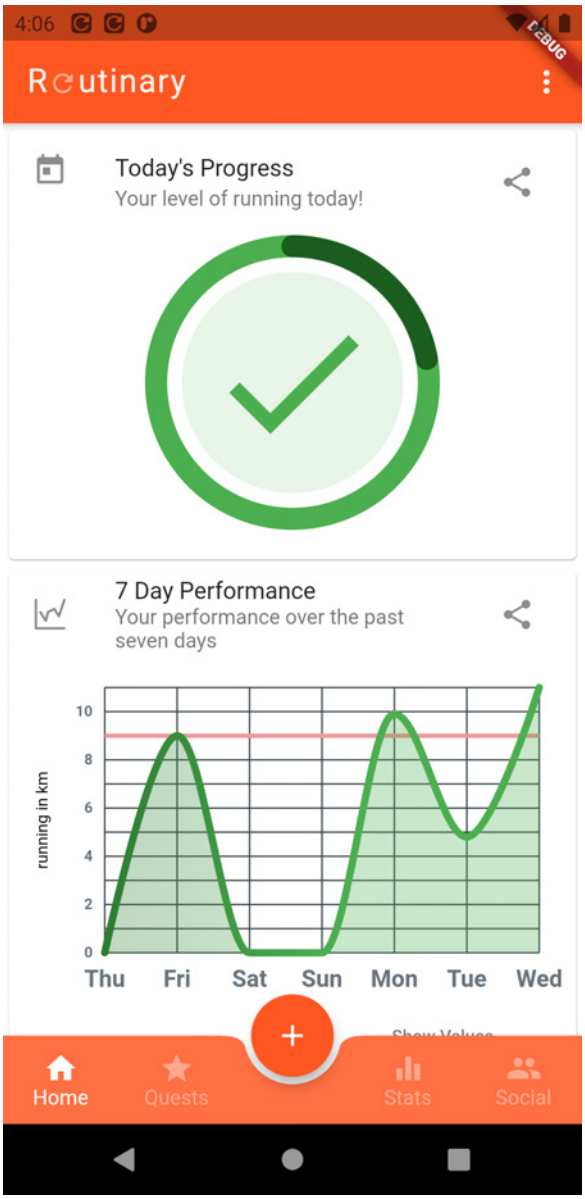
The four appendices are available as separate PDF-files, due to the large file-size caused by the amount of images. Page numbering continues and they are considered part of the report.
Appendix A presents an additional overview with explanations on the layout and features of the app itself, which should be viewed supplemental to Section 4 of the main paper. The setup process, including all relevant aspects of the app, is detailed in Appendix B. In depth explanations of all Motivational Cards that are implemented in Routinary are presented in Appendix C. Finally, supplemental results to the main paper are shown in Appendix D.

A Routinary App Layout

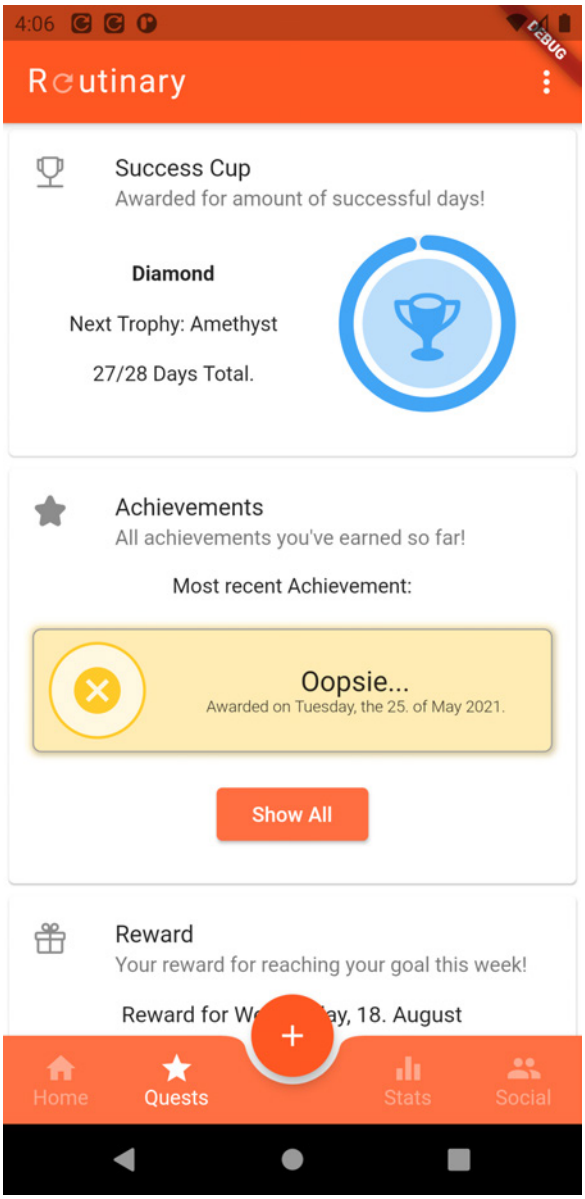
Figure A.1

Routinary Screens - Home & Quests

(a) *Home Screen*



(b) *Quests Screen*

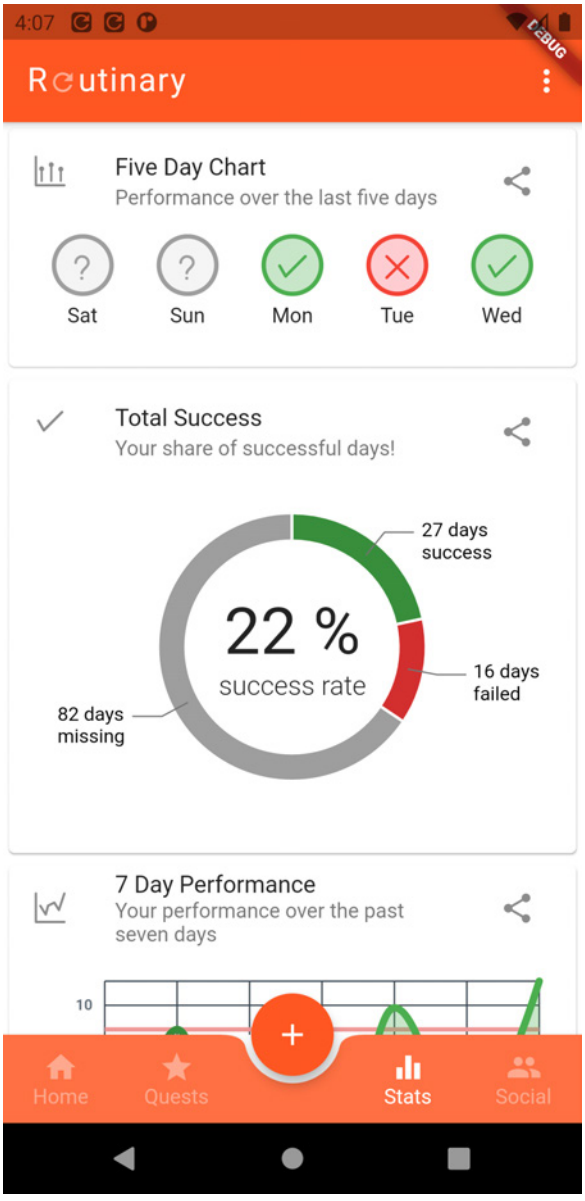


Notes. (a) The Home screen includes the Today's Progress Card, as well as the two tailored cards, of which one is visible here, but the second one only gets visible once the user scrolls down. (b) The Quests screen, including the fixed cards of Success Cup, Achievements and Personal Reward.

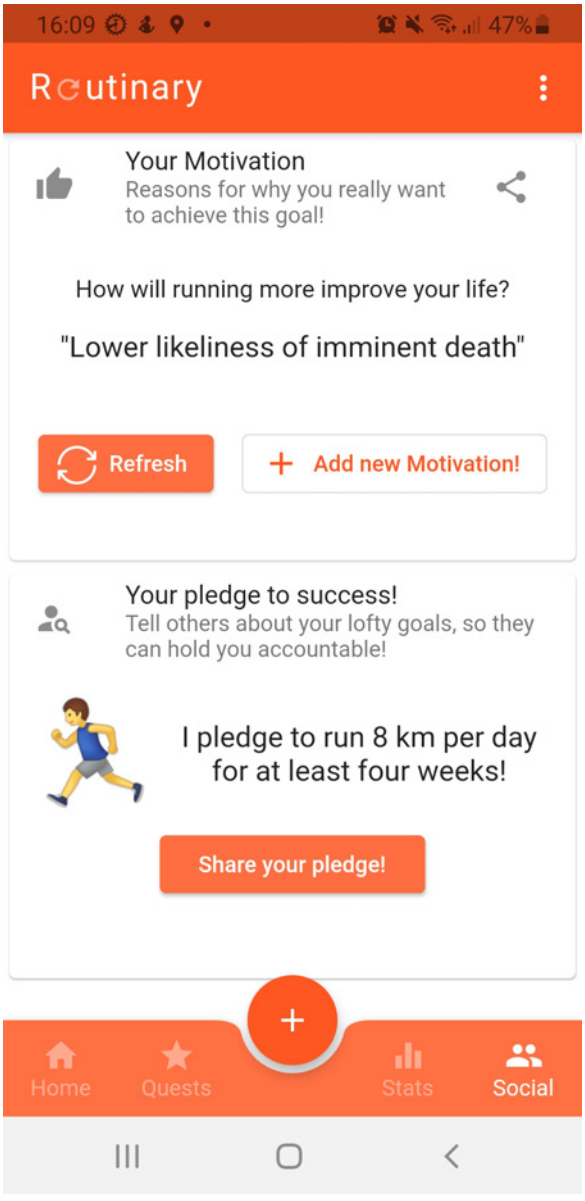
Figure A.2

Routinary Screens - Stats & Social

(a) *Stats Screen*



(b) *Social Screen*

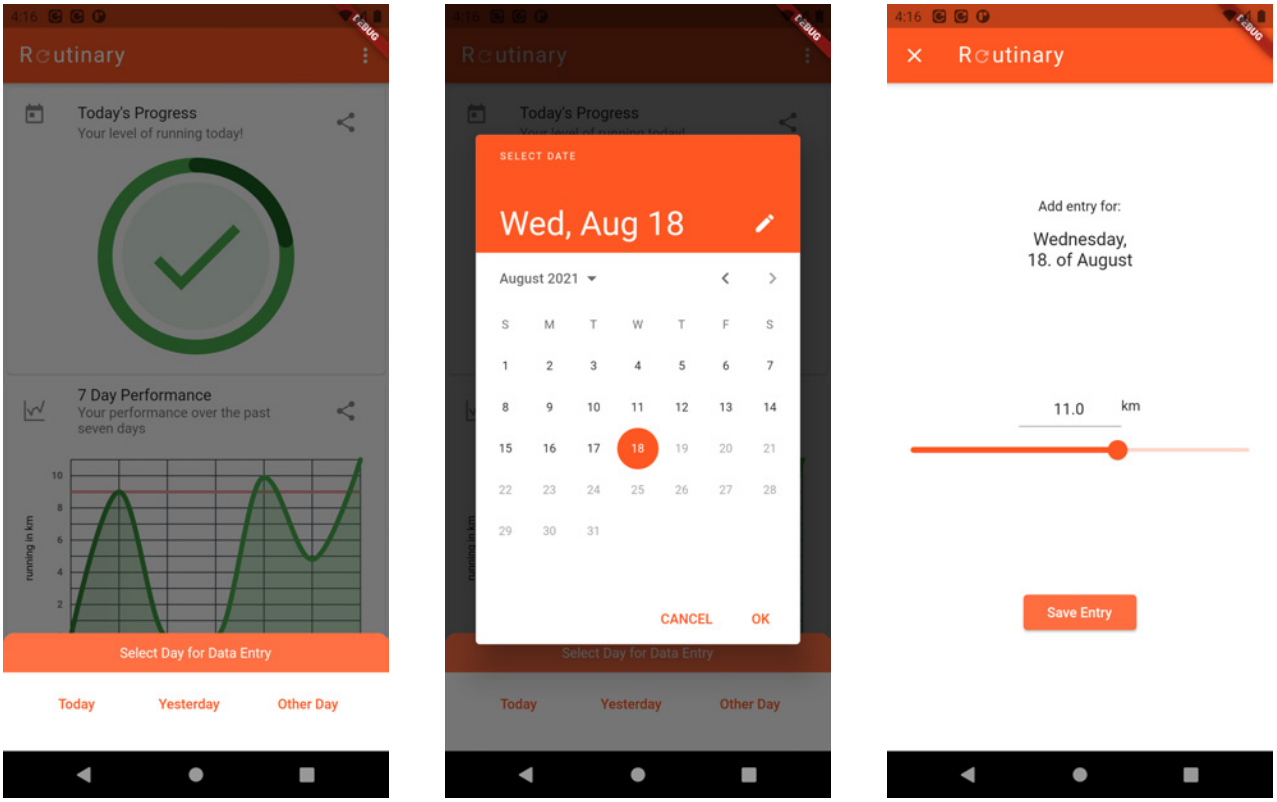


Notes. (a) The Stats screen has the fixed cards of Five Day Chart, Total Success and 7 Day Performance. (b) The Social screen, including Personal Motivation and the Accountabiliter.

Figure A.3

Data Entry Process

(a) *Simple Date Selection* **(b)** *Extensive Date Selection* **(c)** *Data Entry*

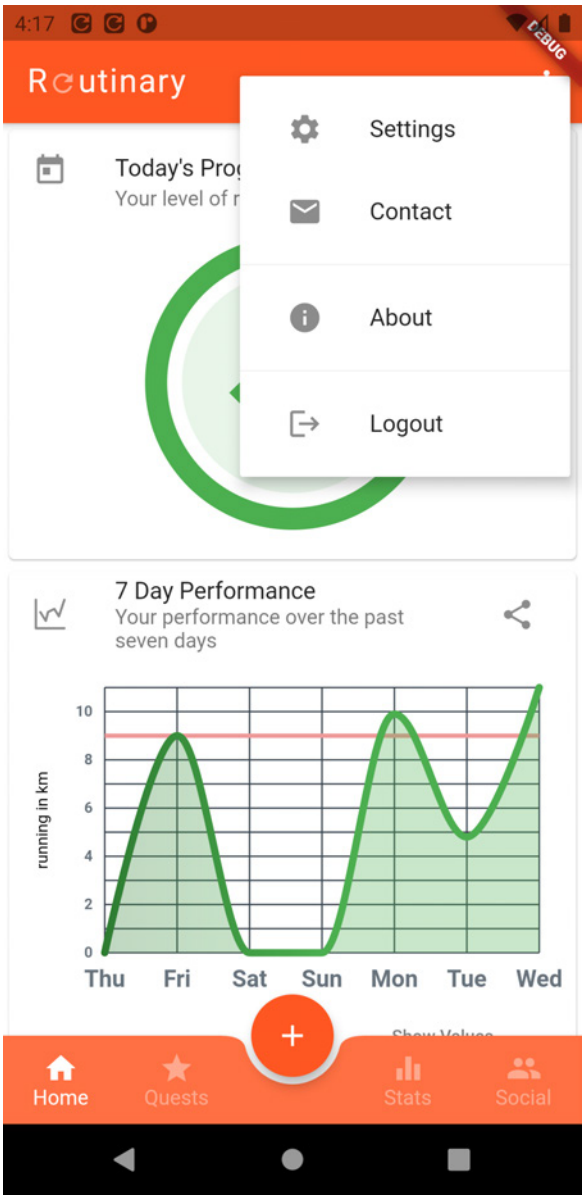


Notes. (a) When the plus-button on the main screen is pressed, a simple selection menu pops up, choosing between whether to enter the data for the same day, the previous day or any other day. (b) In case "Other Day" is selected, a date picker opens, where the user can specify the correct date and continue. (c) When a date is selected, the user gets an option to enter the amount of activity done for the specified date.

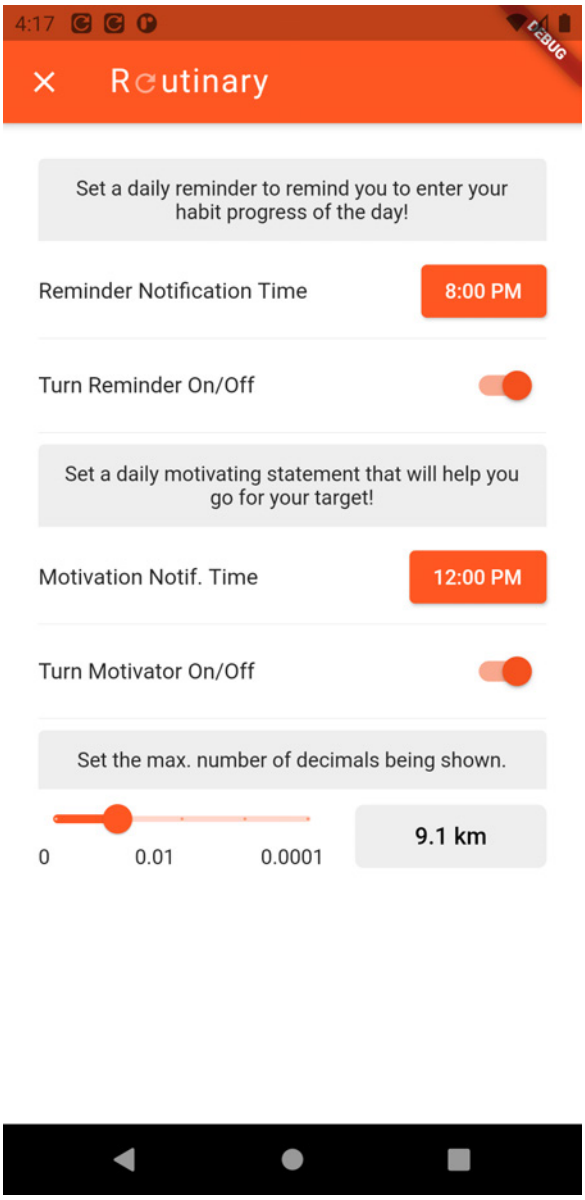
Figure A.4

Options & Settings

(a) *Options Menu*



(b) *Settings Screen*

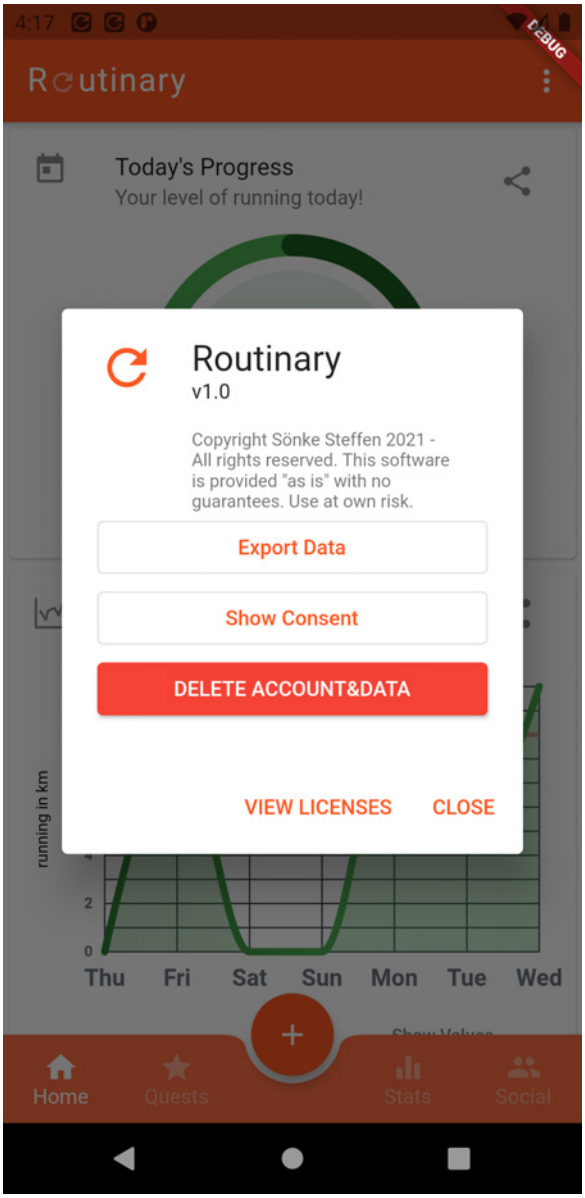


Notes. (a) When the "Navigation Drawer" is pressed, an extended options menu opens. It allows to open the settings screen, send an email via Contact, open the About page, or immediately Logout of the App. (b) The settings screen gives the user options to modify the notifications, as also done in the setup of the experiment. It further allows to change the number of decimals shown in the app, in case this wasn't properly done during the setup. This, however, does not affect the goal.

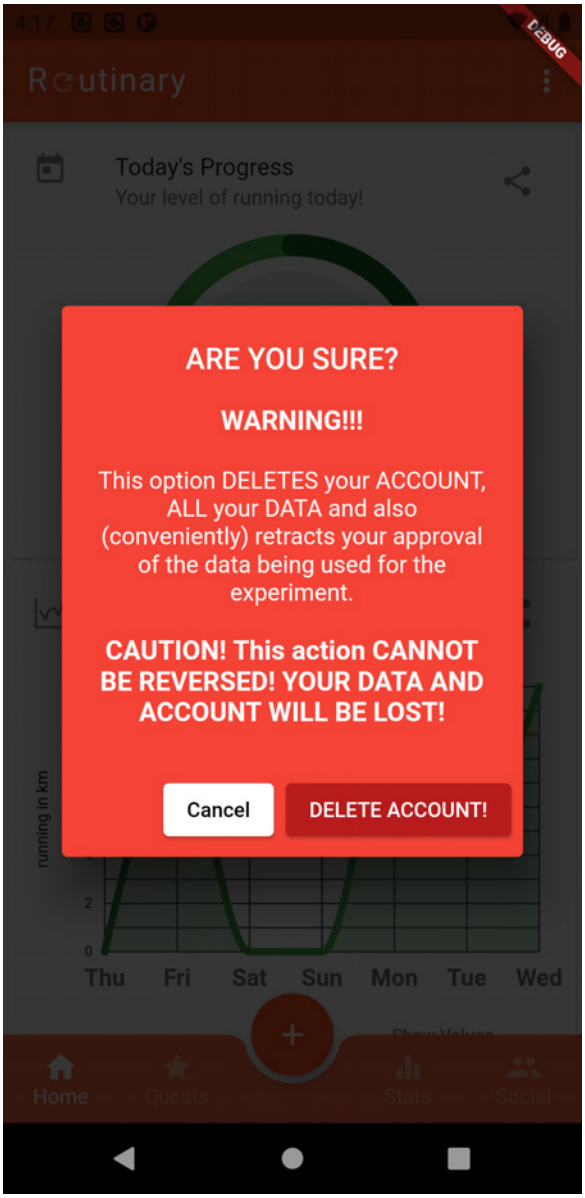
Figure A.5

About & Delete Account Popups

(a) *About Popup*



(b) *Delete Account Popup*

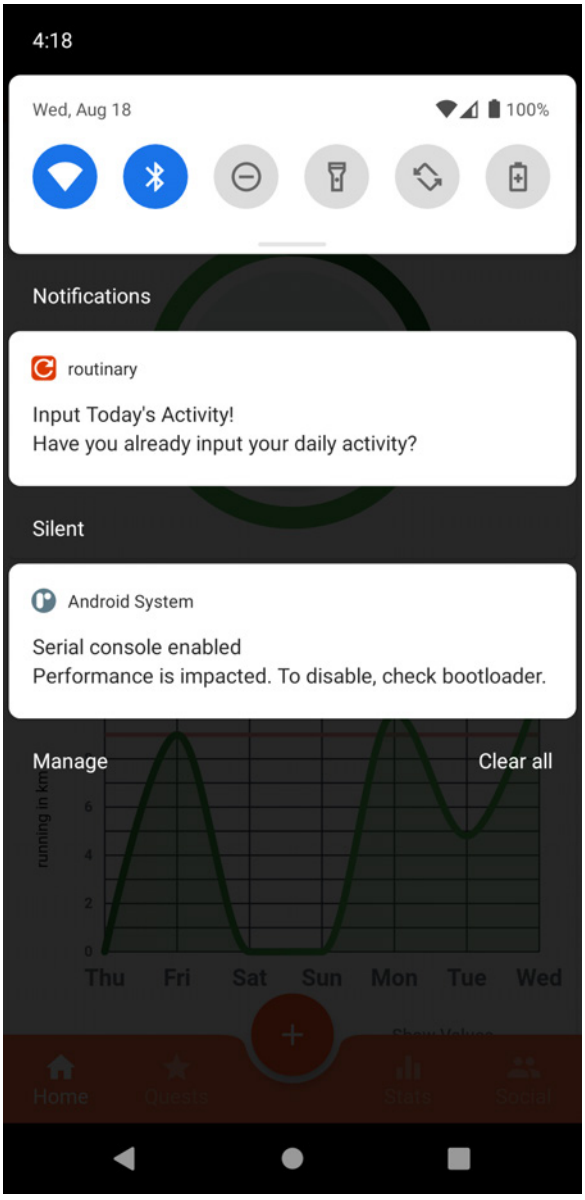


Notes. (a) The About popup allows the user to access information on the app itself. Some further options include to export the user data, show the informed consent, which is identical to the one presented in the setup of the app. The user can also view all licenses. These include the licenses from a lot of different features and libraries that were utilized for the creation of Routinary. (b) Finally, the user has the option to delete the account and data,, which opens this very attention-grabbing window, warning the users of all consequences of deleting the account.

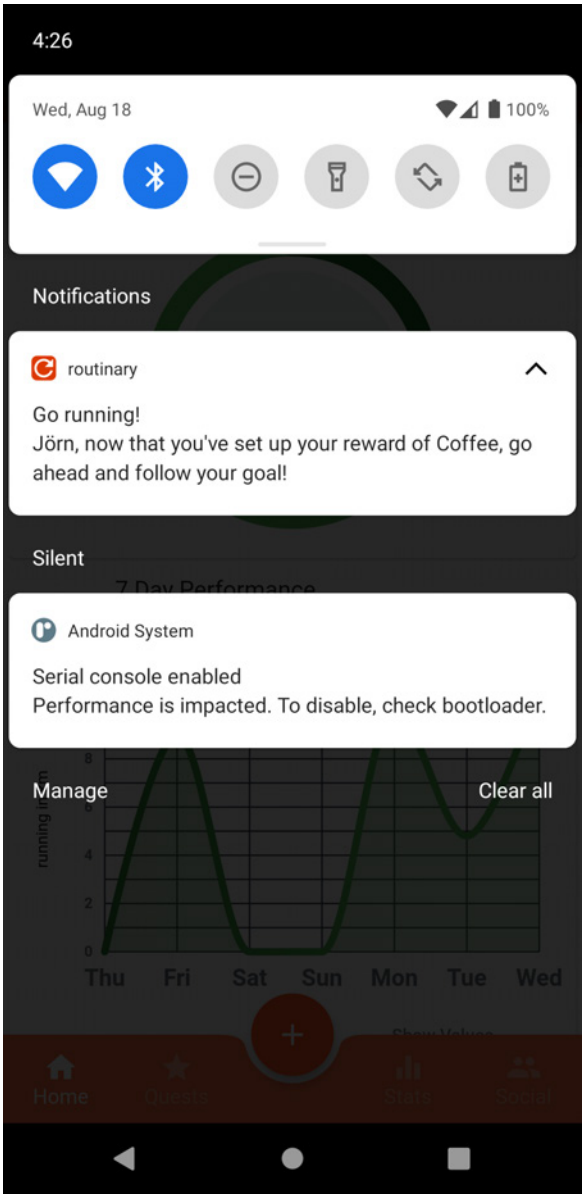
Figure A.6

Notifications

(a) *Reminder Notification*



(b) *Motivation Notification*



Notes. Examples of how the notifications are shown. (a) The simple reminder notification. (b) An example of a motivational notification.

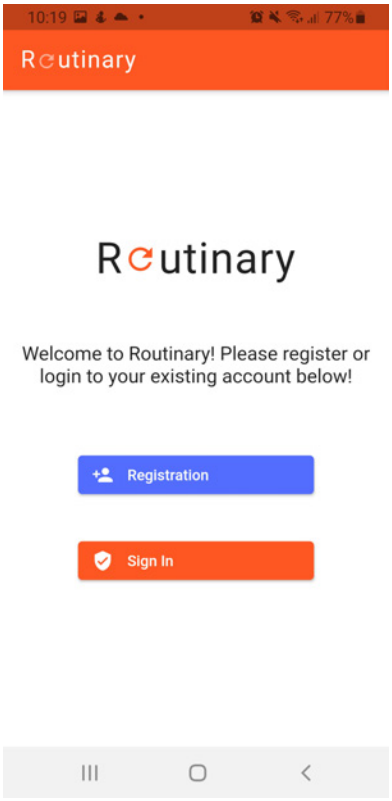
B Setup Process in Routinary

This Appendix includes detailed and uncommented screenshots/images of all steps of the login and setup process. The order is according to the succession participants go through. Slight differences in the layout and content are caused by the screenshots being taken from different devices and different habits. As the content of the card ratings is tailored to the previously input habit and goal, the content of the example cards would be adjusted accordingly. Screenshots are both from the Android Emulator and a Samsung Galaxy S9 phone, the visual appearance on other devices might differ.

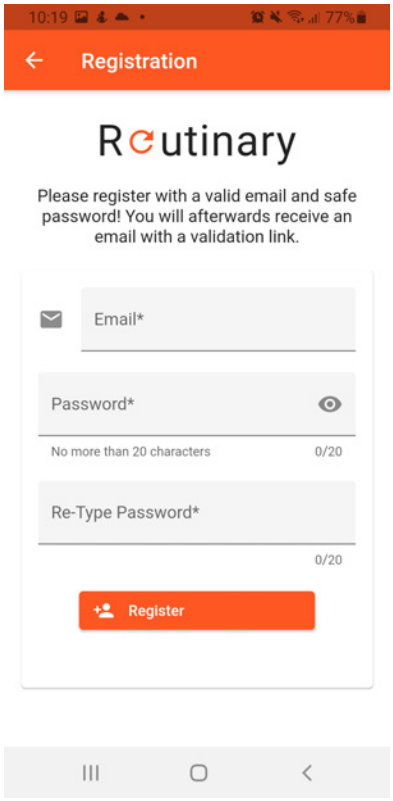
Figure B.1

Start, Register and Login

(a) *Window at first Startup*



(b) *Registration Window*



(c) *Login Window*

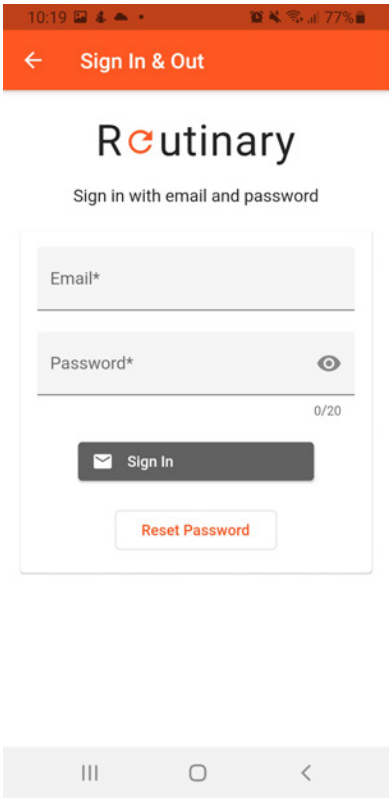
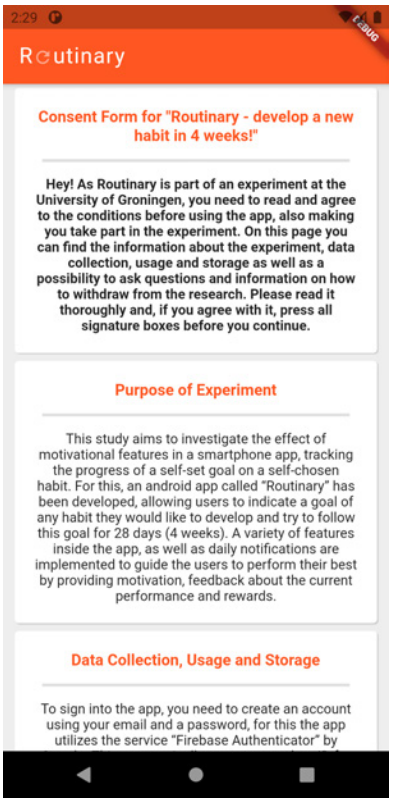


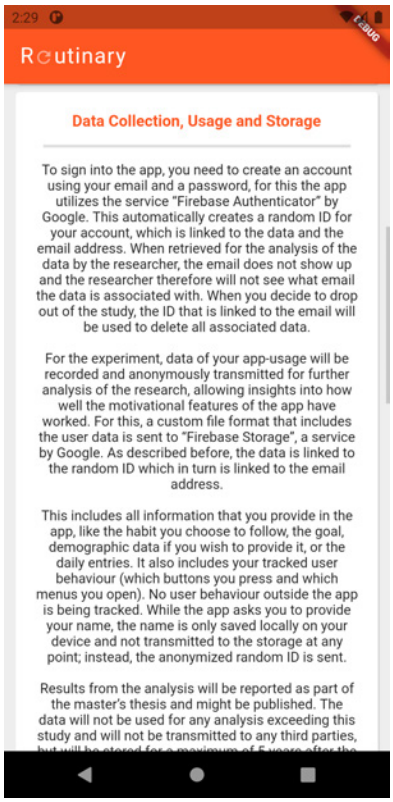
Figure B.2

Consent Form

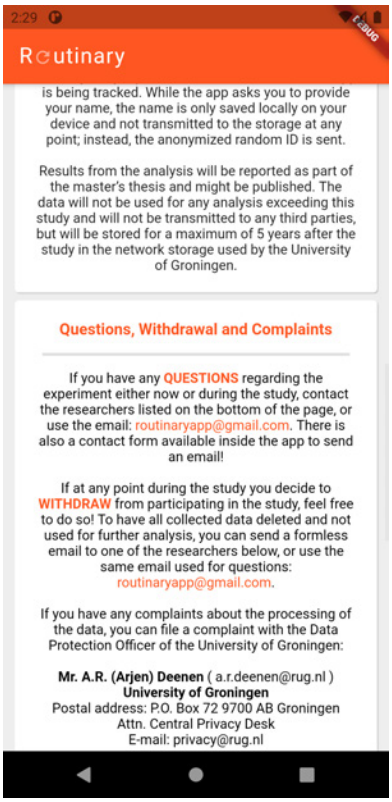
(a) *Intro & Experiment Purpose*



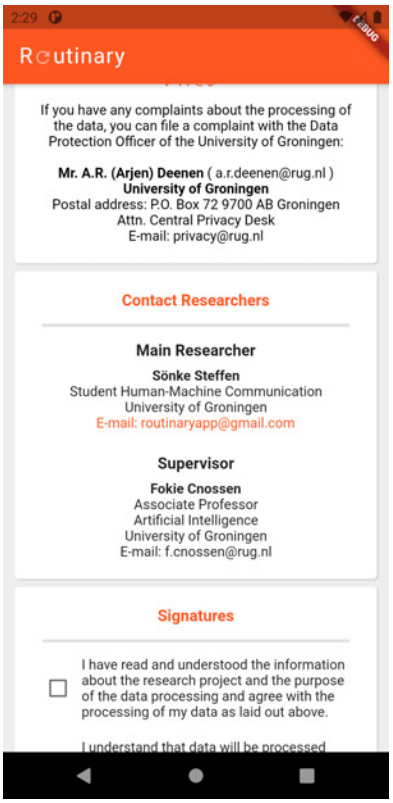
(b) *Data Policy*



(c) *Participant Actions*



(d) *Contact Researchers*



(e) *Signatures*

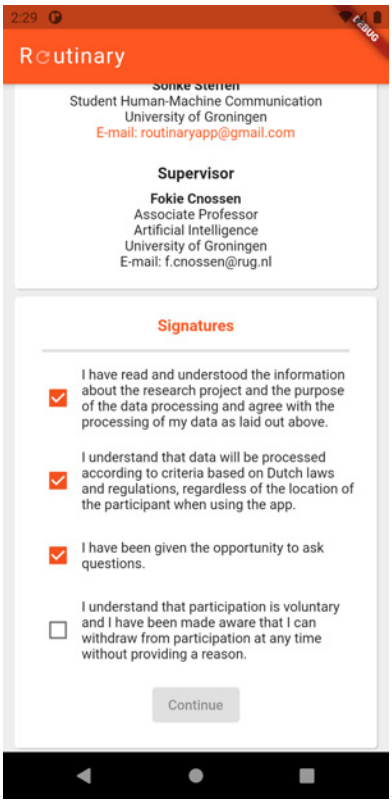
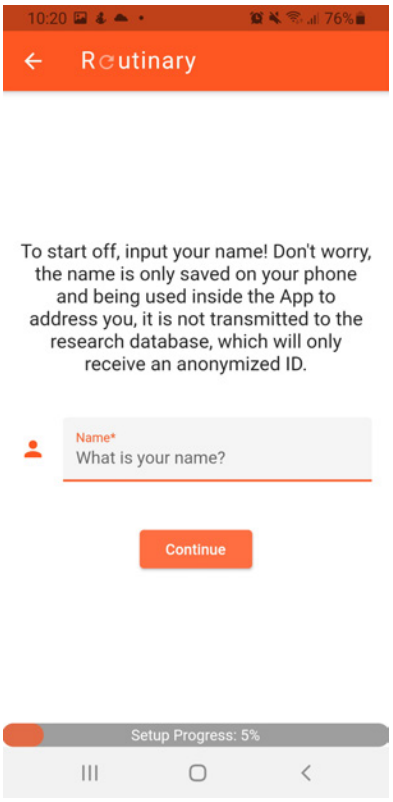


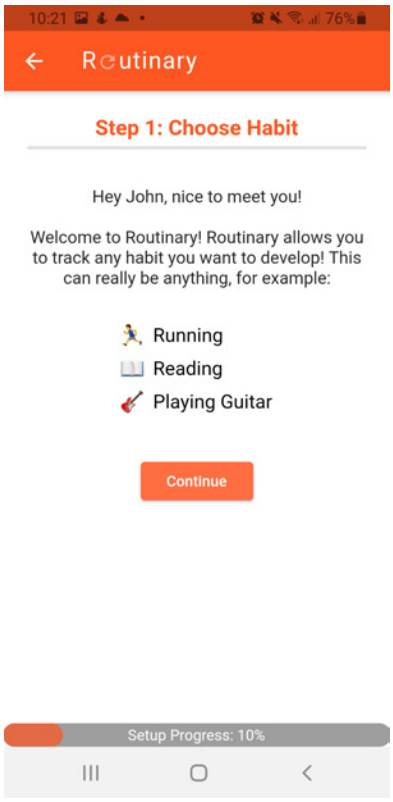
Figure B.3

Part 1: Introduction to App

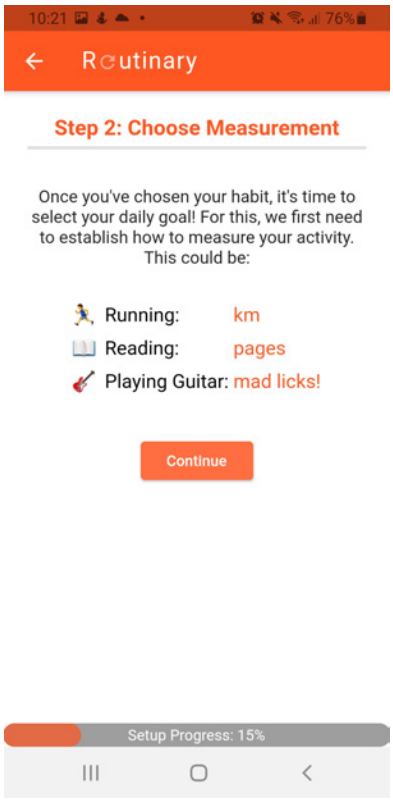
(a) *Entering the Name*



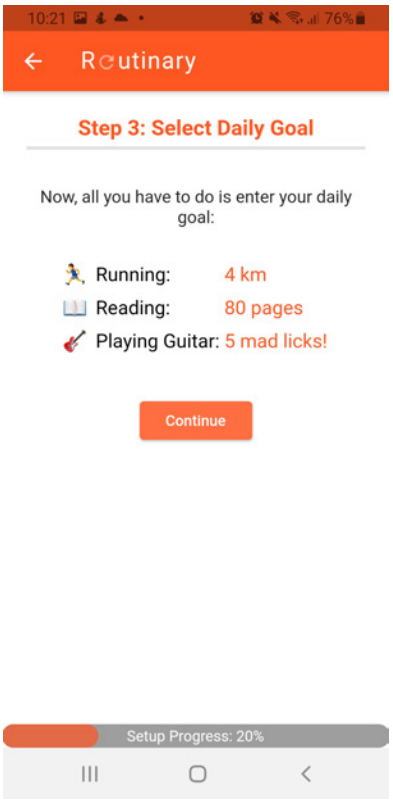
(b) *Habit Tutorial: S1*



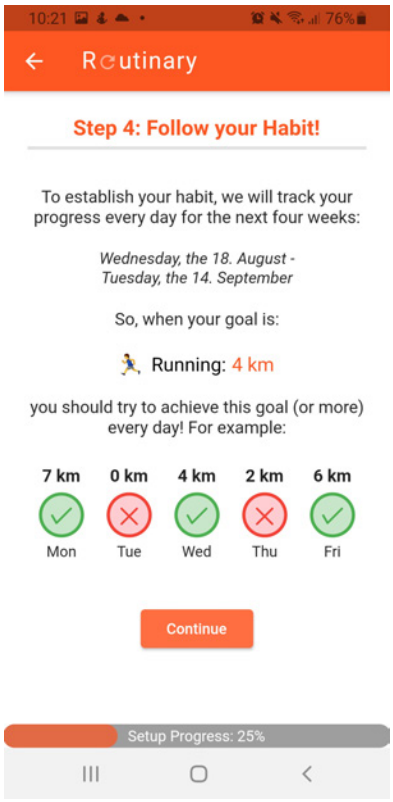
(c) *Habit Tutorial: S2*



(d) *Habit Tutorial: S3*



(e) *Habit Tutorial: S4*



(f) *Demographic Data*

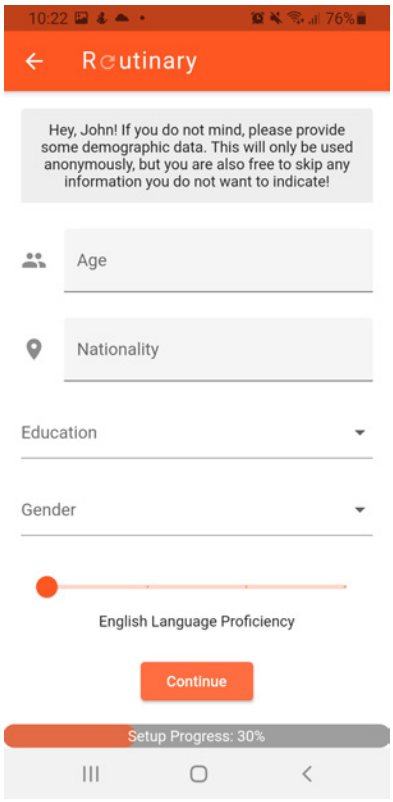
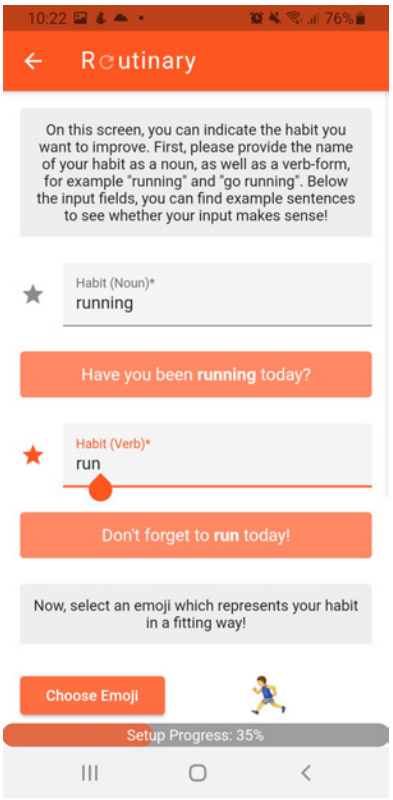


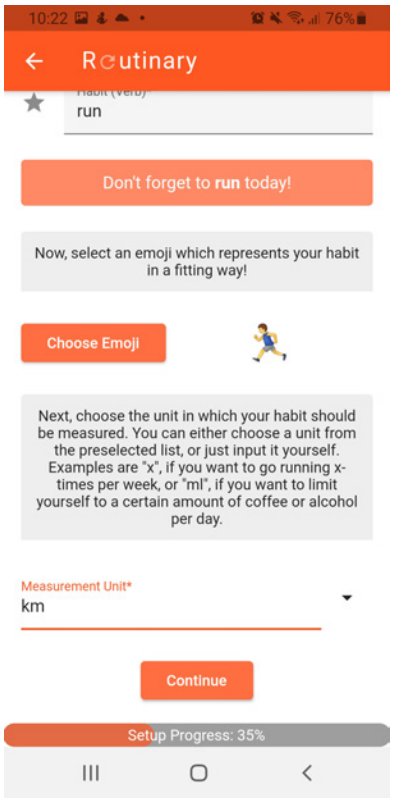
Figure B.4

Part 2: Setup of Habit and Goal

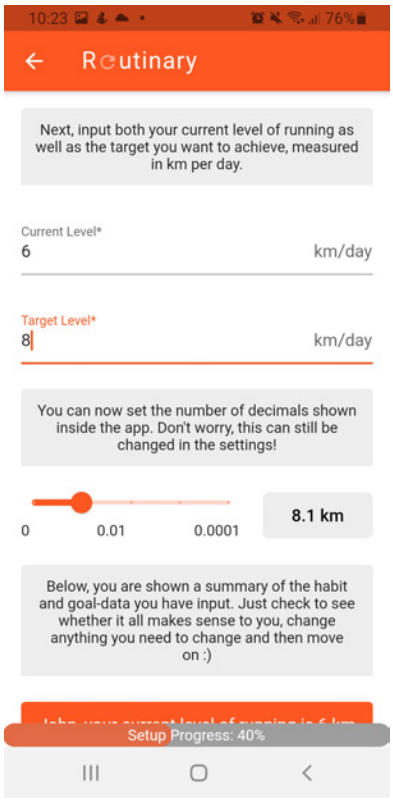
(a) *Habit: Noun & Verb*



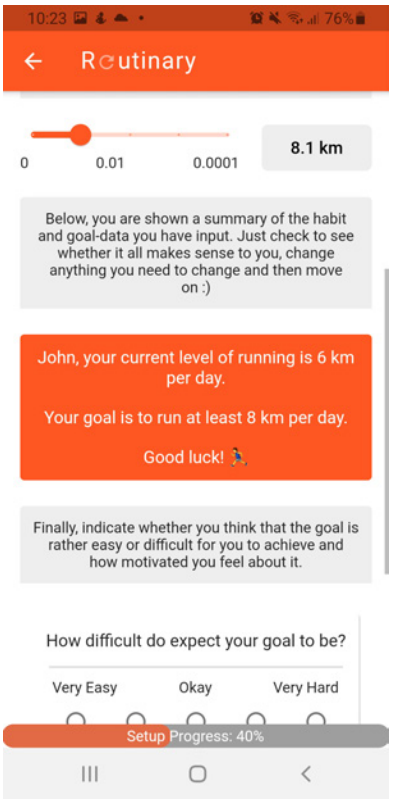
(b) *Habit: Emoji & Measure*



(c) *Goal: Target & Precision*



(d) *Goal: Summary*



(e) *Goal: Difficulty & Motivation*

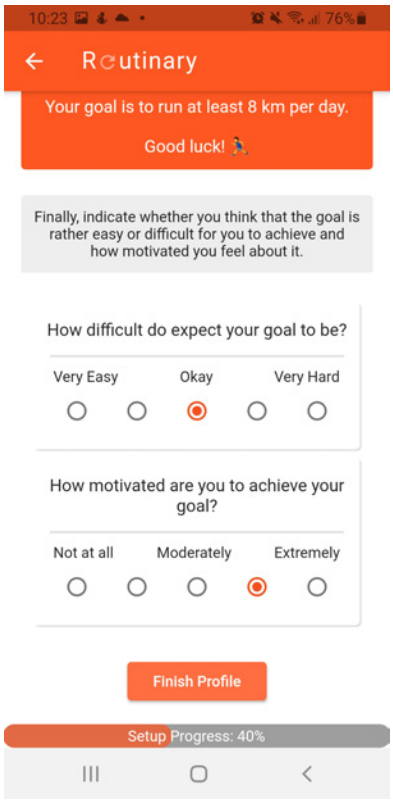
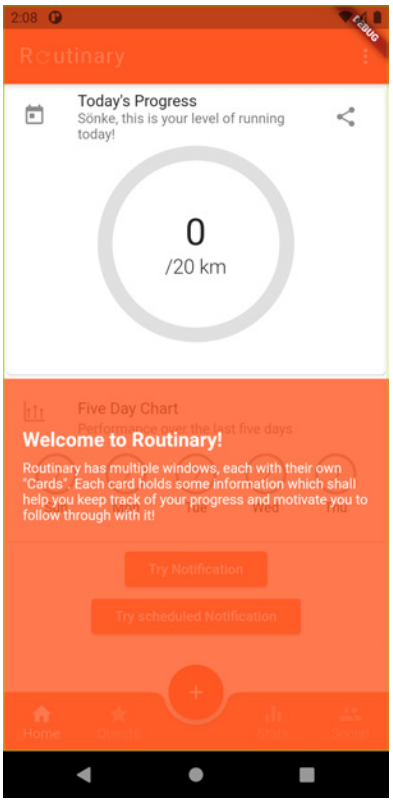


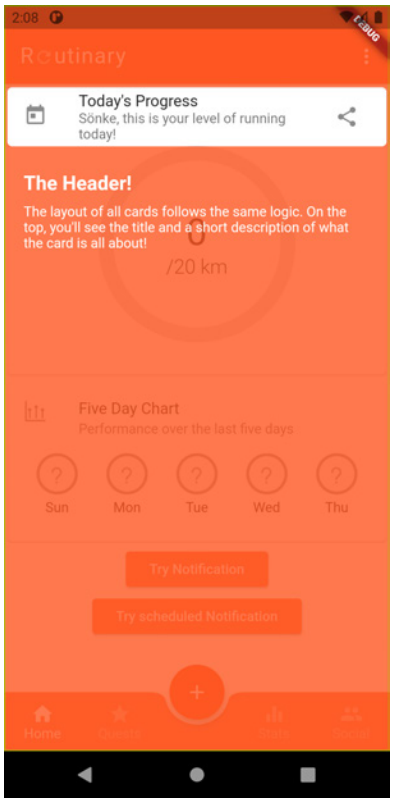
Figure B.5

Part 3: In App Tutorial

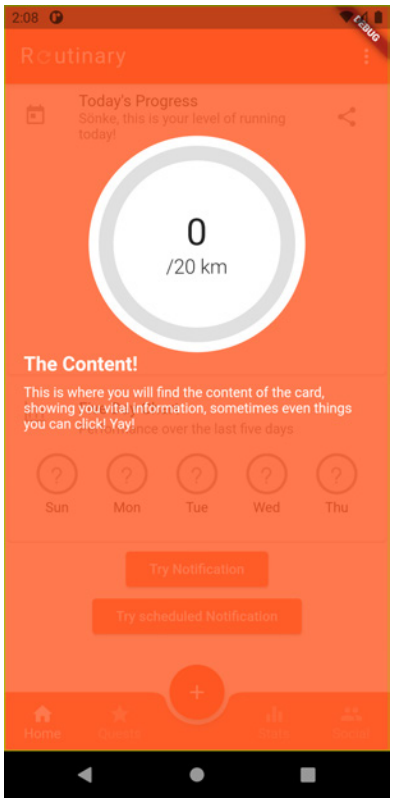
(a) *Welcome & Card Intro*



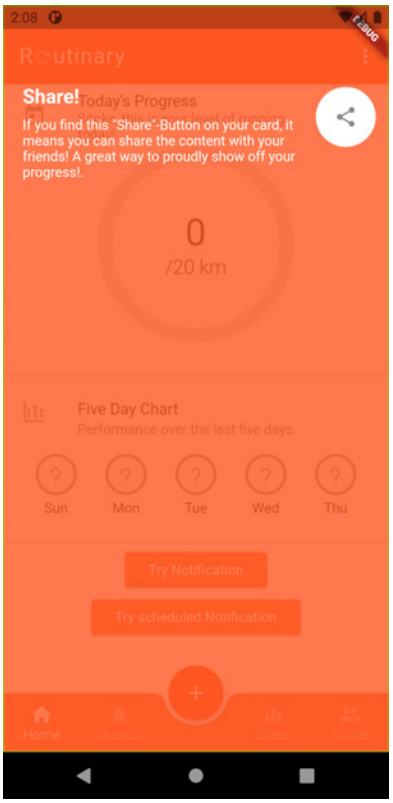
(b) *Card Header*



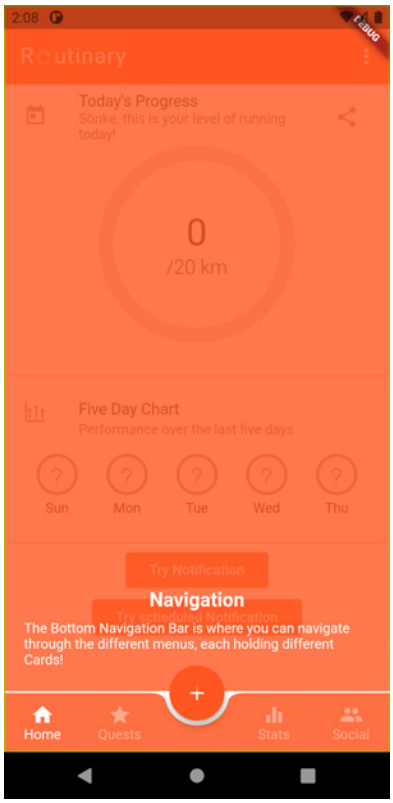
(c) *Card Content*



(d) *Card Sharing*



(e) *Navigation Bar*



(f) *Add Entry*

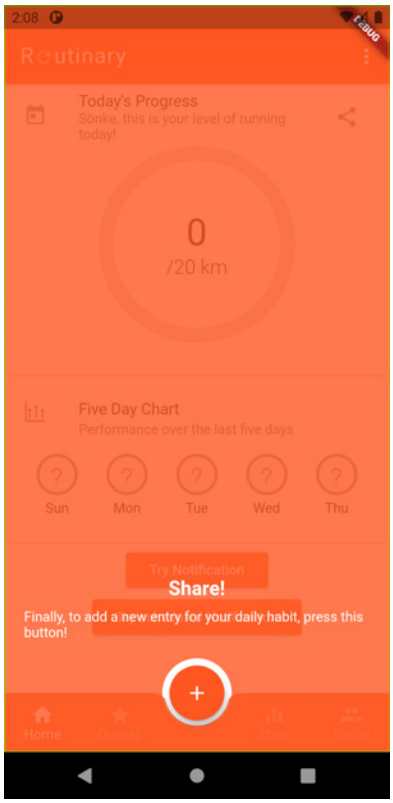
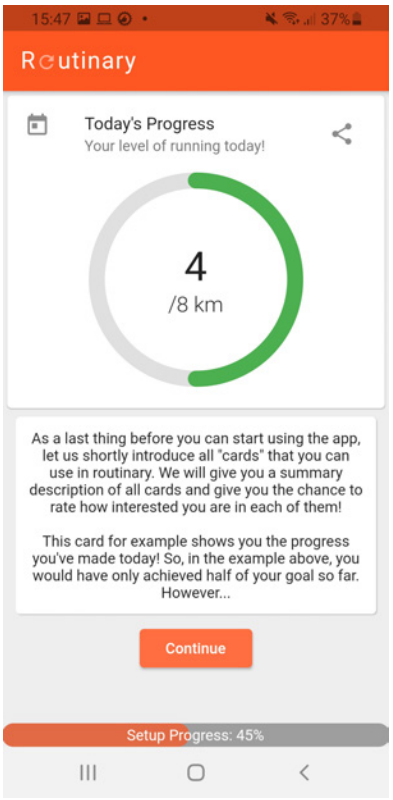


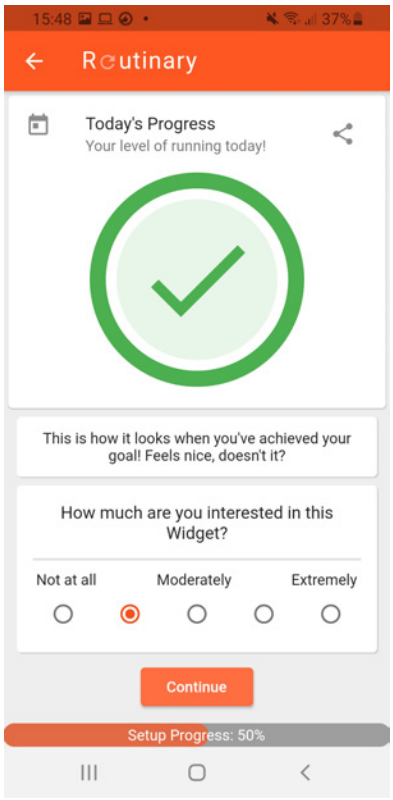
Figure B.6

Part 4: Card Ratings 1

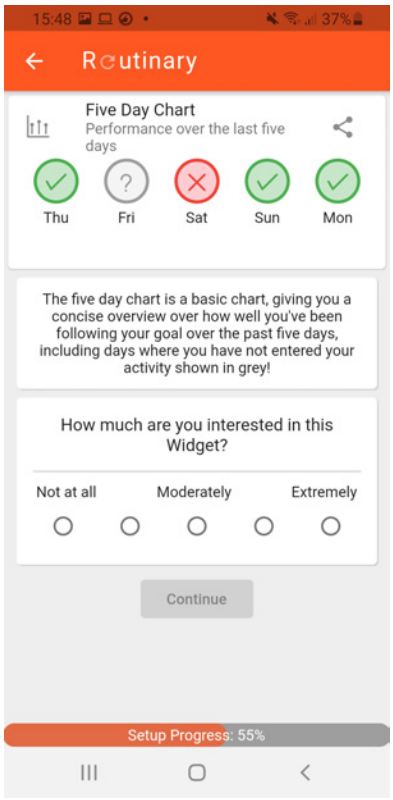
(a) *Intro Questionnaire*



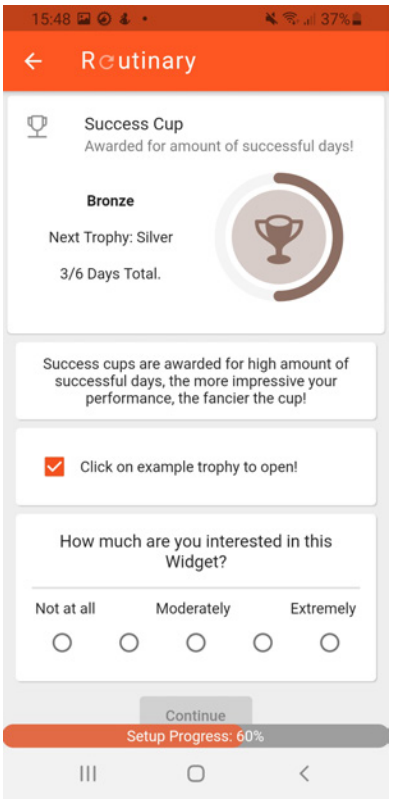
(b) *Today's Progress*



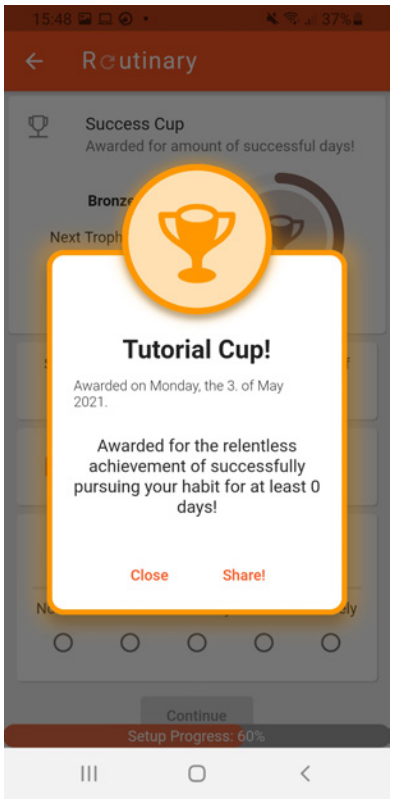
(c) *Five Day Chart*



(d) *Success Cup*



(e) *Example Success Cup*



(f) *Achievements*

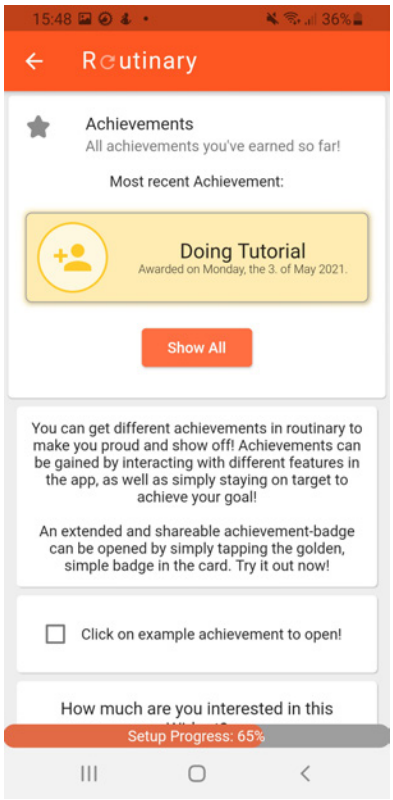
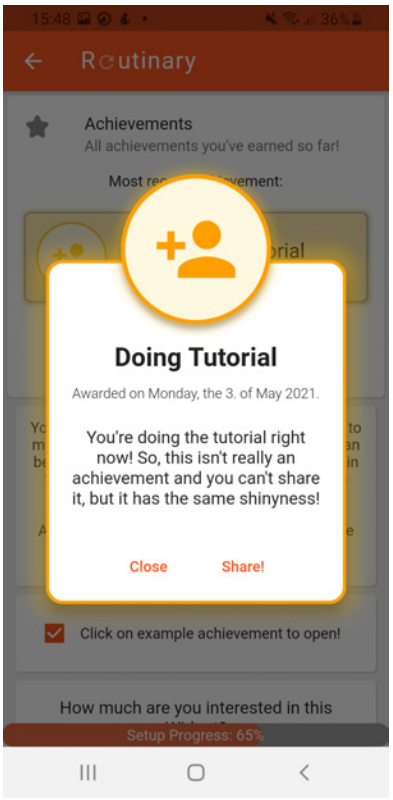


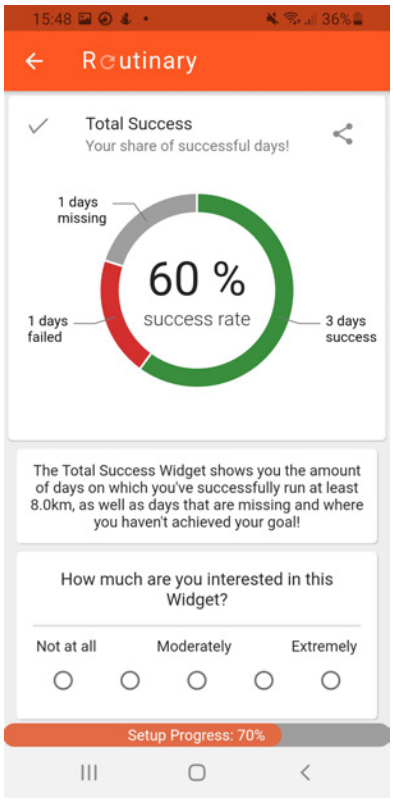
Figure B.7

Part 5: Card Ratings 2

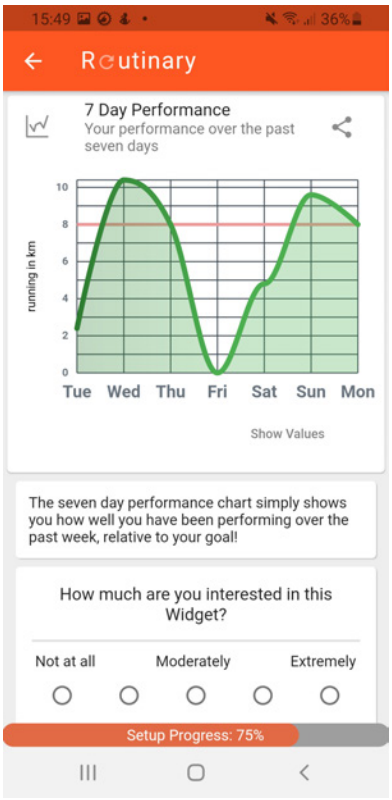
(a) *Example Achievement*



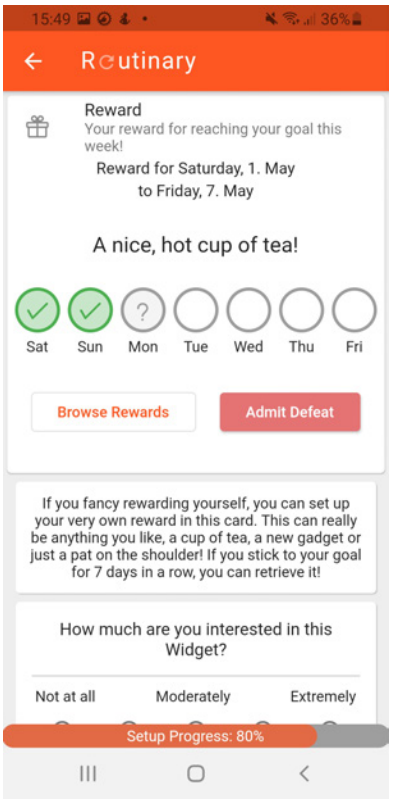
(b) *Total Success*



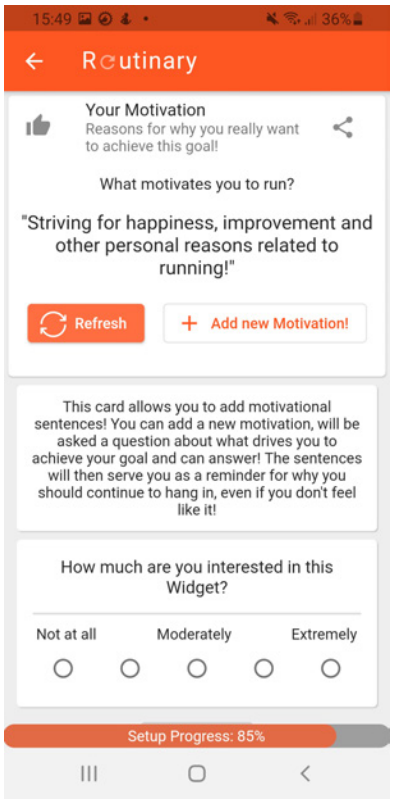
(c) *7 Day Performance*



(d) *Personal Reward*



(e) *Personal Motivation*



(f) *Accountabiliter*

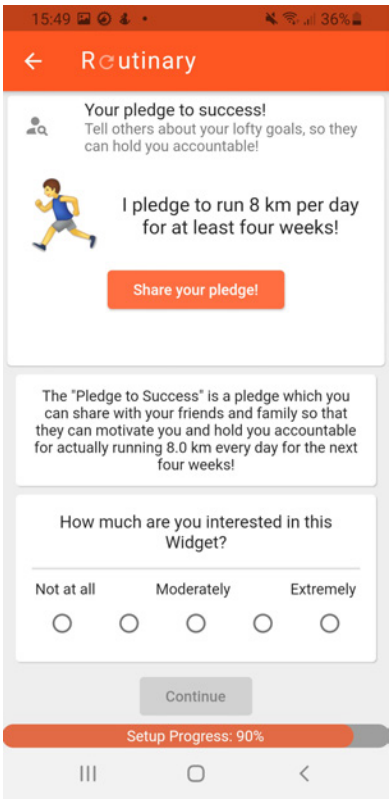
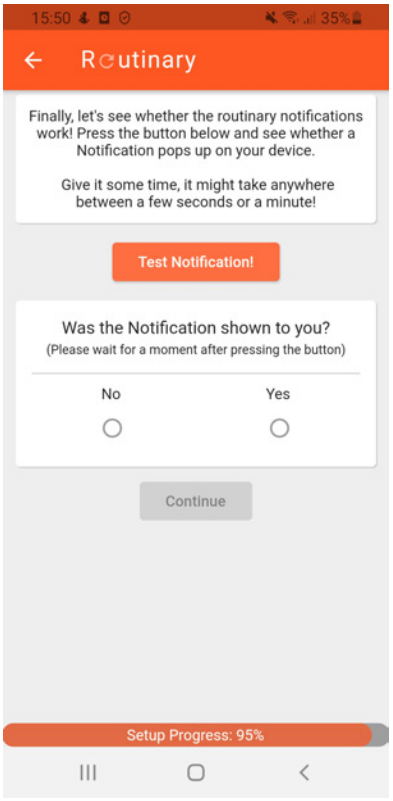


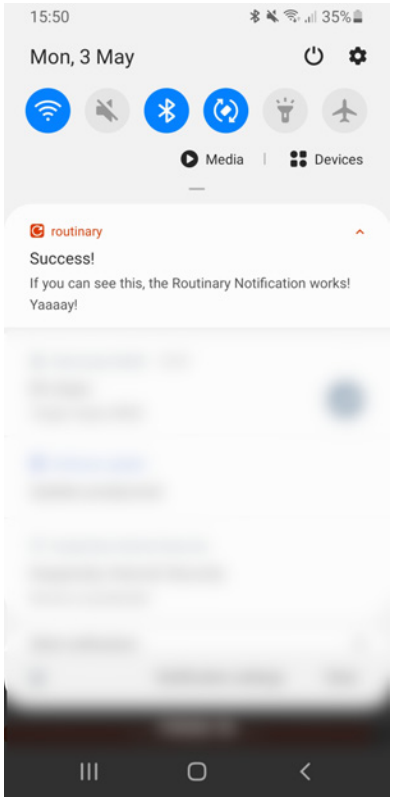
Figure B.8

Part 6: Notification Test & Settings

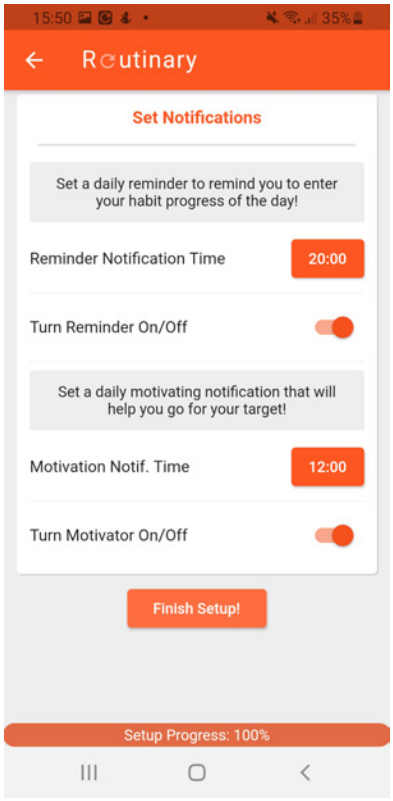
(a) *Notification Test*



(b) *Successful Notification*



(c) *Notification Settings*

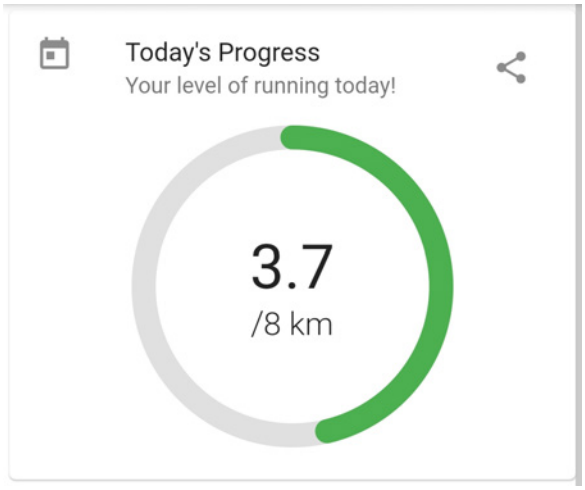


C Motivational Cards

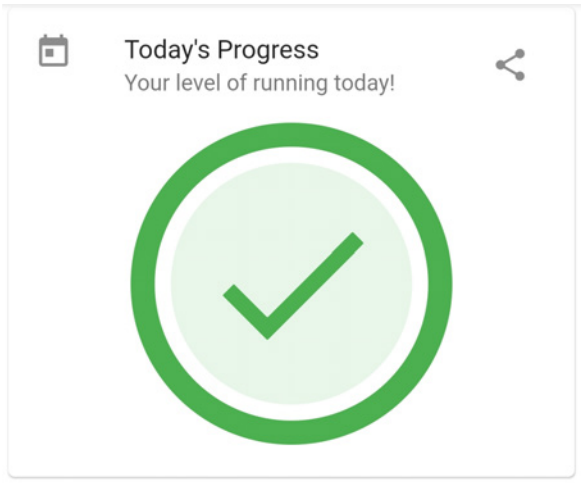
Figure C.1

Today's Progress

(a) *Card when Goal not Reached*



(b) *Card when Goal Success*

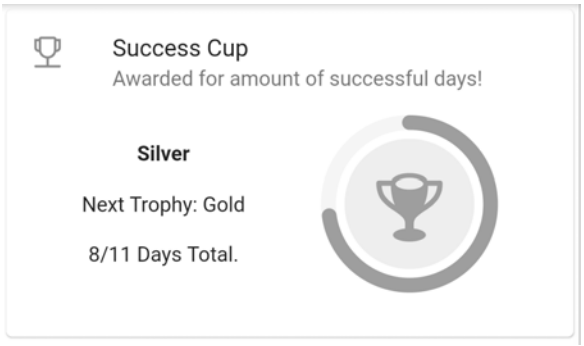


Notes. Today's Progress is only shown on the home screen, it is not used for any of the tailoring! It does show the habit progress of the current day relative to the goal. (a) When the goal hasn't been reached yet, there is both a circular progress bar and numerical representation of the progress relative to the goal. (b) In case the goal has been reached, the center of the circle is colored green including a green checkmark, indicating that the goal has been reached.

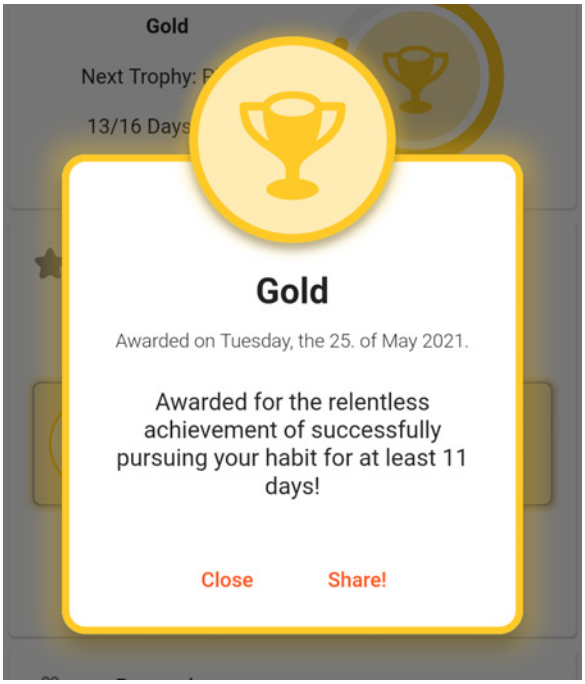
Figure C.2

Success Cup

(a) *Success Cup Card*



(b) *Popup of Success Cup*

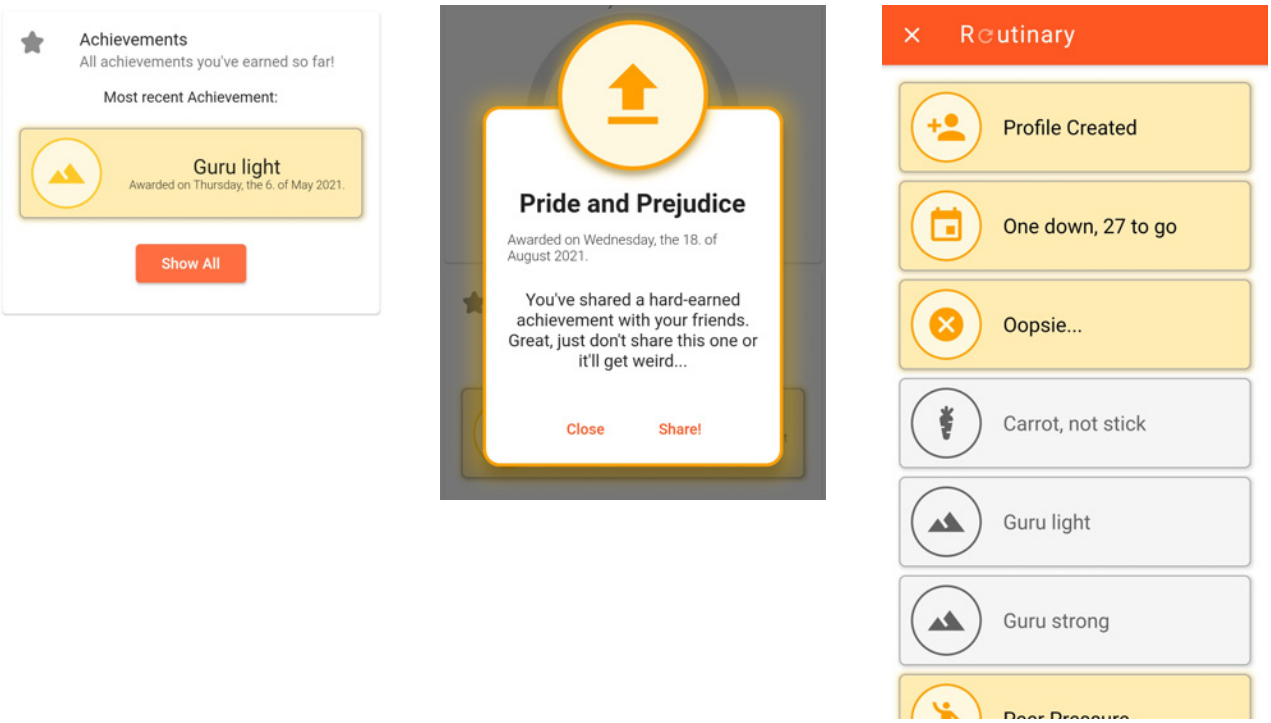


Notes. The Success Cup is an artificial reward for the number of successful days (in total, not in a row) the user has achieved the goal. This is represented by a ”cup” which increases in ”material worth”, so it will be Bronze, Silver, Gold etc. with increased success. The color is adjusted accordingly. (a) The card itself, including information and a visual representation on the progress achieved so far, the current trophy and the next possible trophy. (b) The popup shown when a new success cup has been achieved. It can also be opened by pressing on the cup itself on the motivational card.

Figure C.3

Achievements

(a) *Achievement Card* **(b)** *Popup of Achievement* **(c)** *List of Retrieved Achievements*

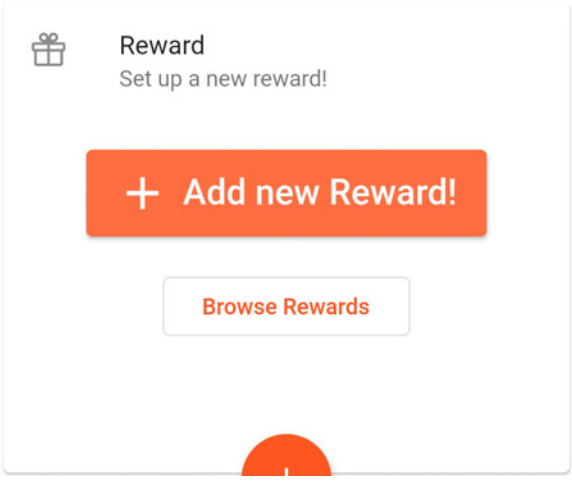


Notes. Achievements are gained by successfully fulfilling certain conditions. These vary heavily in nature, there is for example an achievement for the first entry made, for adding motivational statements or for achieving the self-set reward. (a) The card itself showing the most recent achievement, which can be opened by pressing. (b) Shows a popup of an achievement. (c) The list of achievements, with those that the users has gained shown in a golden color. This list can be accessed by pressing ”Show All” on the motivational card.

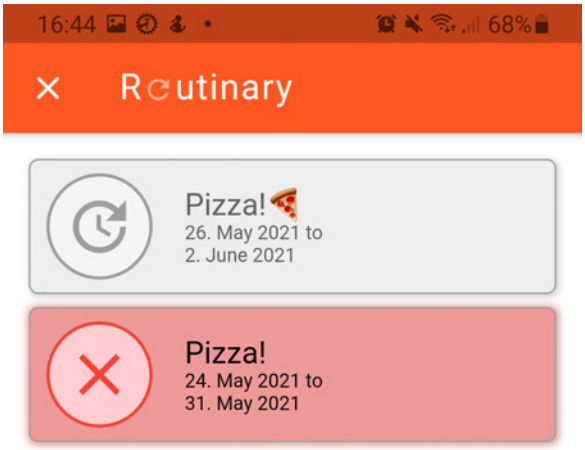
Figure C.4

Personal Reward

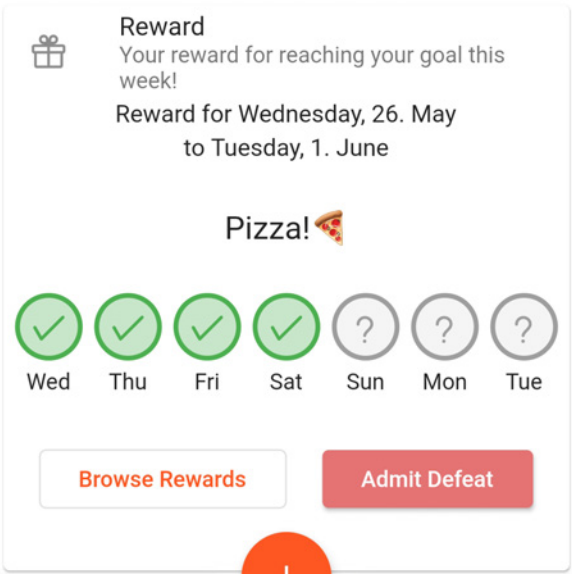
(a) *Reward Card with no Reward Setup*



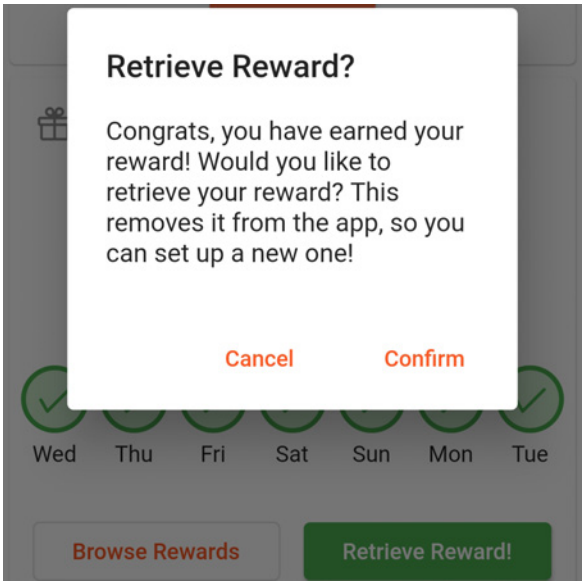
(b) *List of Pursued Rewards*



(c) *Reward during Pursue*



(d) *Retrieving the Reward*

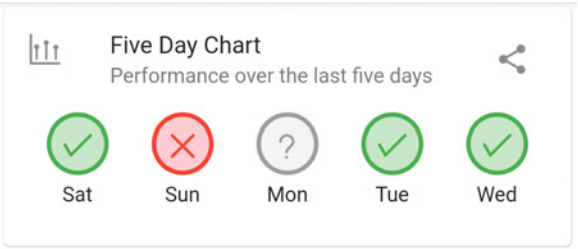


Notes. The Personal Reward allows the user to setup any reward they want as a reward for being able to pursue the goal for seven days in a row. (a) When no reward is set up, the user can add a new one by pressing the button. (b) Alternatively, past rewards can be browsed by pressing the respective button. It not only shows the pursued reward, but also the status. This can either be ongoing in grey, forfeited in red or succeeded in green. (c) When a reward has been setup, the progress can be observed on the card itself, together with an option to forfeit the reward in case the user has failed. (d) If successful, the user can retrieve the reward, resetting the card to the status shown in (a).

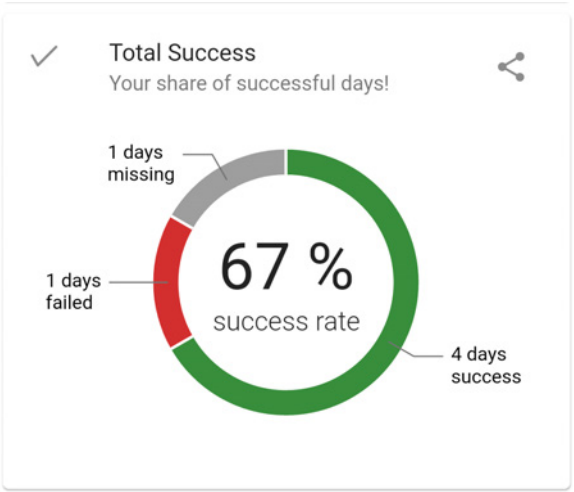
Figure C.5

Five Day Chart & Total Success

(a) *Five Day Chart*



(b) *Total Success*

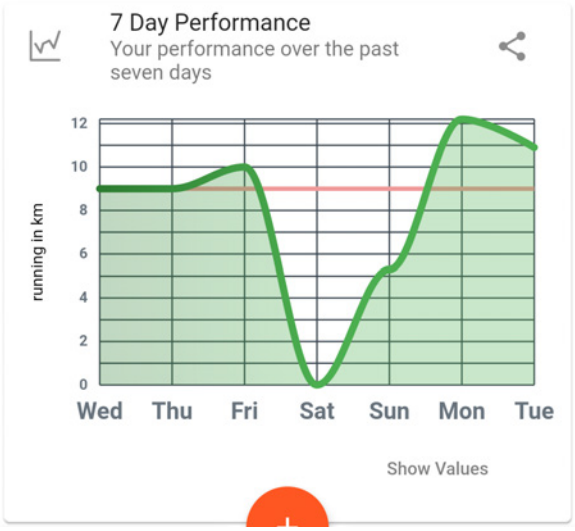


Notes. (a) The Five Day Chart just presents the users with a simple visual depiction of whether they achieved their goal or not over the past five days, with a green checkmark if that is the case, a red cross if it is not and a grey questionmark if the entry is missing. (b) The Total Success card Shows the relative success of the user over the course of the experiment. It shows a colored ring with the fractional representation and label amounting to the number of successful, unsuccessful and missing daily entries. The fraction of success is numerically shown in the center of the circle.

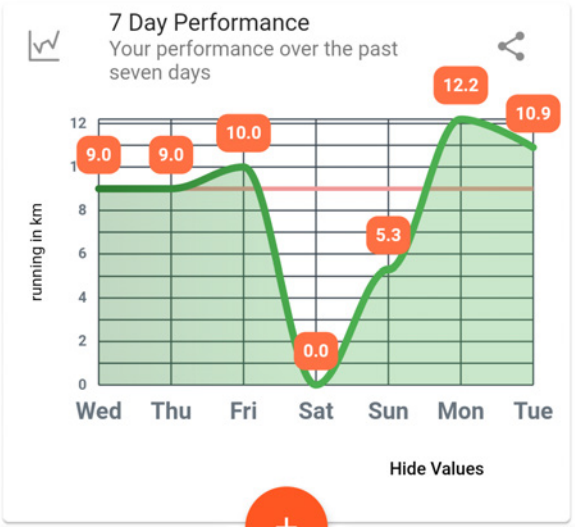
Figure C.6

Seven Day Performance

(a) *Default Card without Values*



(b) *Card with Values Shown*

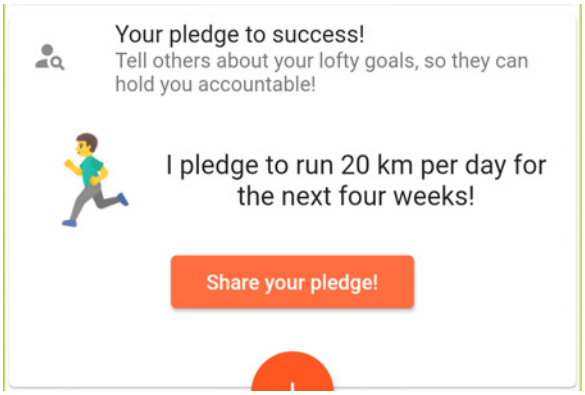


Notes. (a) The Seven Day Performance card shows a smoothed graph of the entered performance over the most recent seven days. It includes a horizontal thicker red line representing the user's goal. (b) If the "Show Values" button is pressed, the daily values are shown above the line.

Figure C.7

Accountabiliter

(a) *Accountability Card*



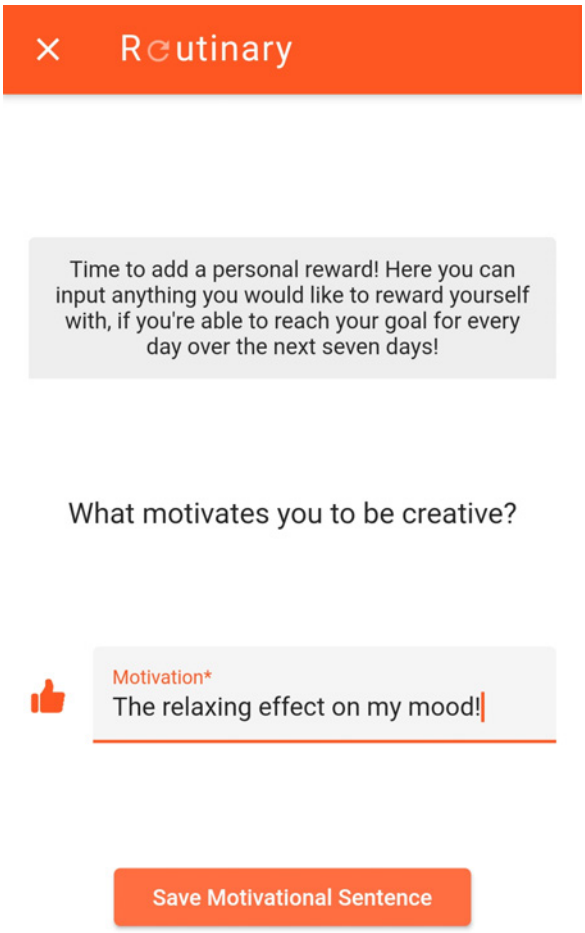
(b) *Shareable Pledge*

Notes. (a) The accountability card is shown as a "pledge to success" and contains information based on the goal, activity and timespan of the experiment, as well as the symbol which the user has chosen in the setup process. (b) The shareable pledge which opens when "Share your pledge!" is pressed. This is slightly more stylized and includes an instruction to the recipient to hold the user accountable. By pressing "Share The Pledge!" the actual OS-own sharing option is triggered and a screenshot of the pledge is shared.

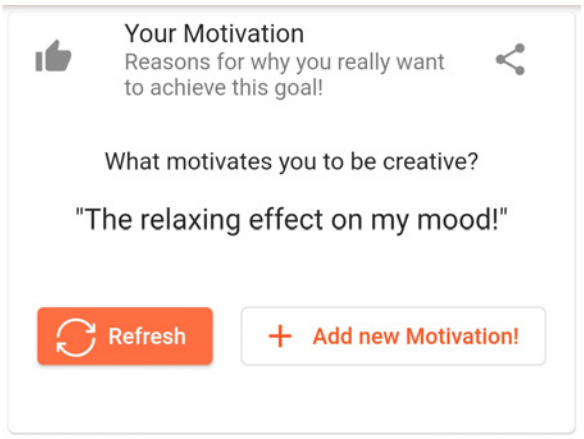
Figure C.8

Personal Motivation

(a) *Adding New Motivational Statement*



(b) *Personal Motivation Card*

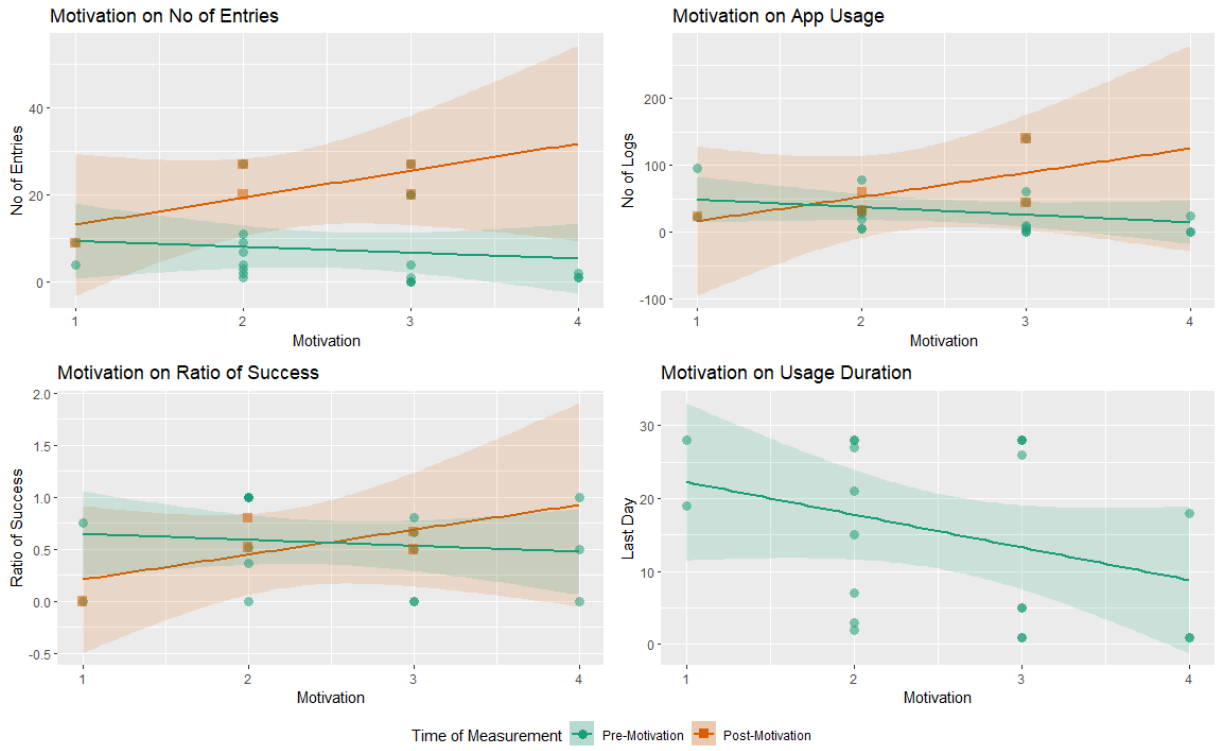


Notes. The Personal Motivation Card offers users with an ability to add motivational sentences by answering questions about what motivates them. (a) When a new statement is added, one of multiple questions is asked randomly. The user can then enter the motivation in the response box and save it. It is also visible that the instruction on top is erroneously from the Personal Reward card, but this did not seem to affect users, as the card is quite self-explanatory. (b) The card itself, which shows one of the randomly entered sentences, including question and response. A different one can be shown by pressing ”refresh” and a new one can be added as well.

D Additional Results

Figure D.1

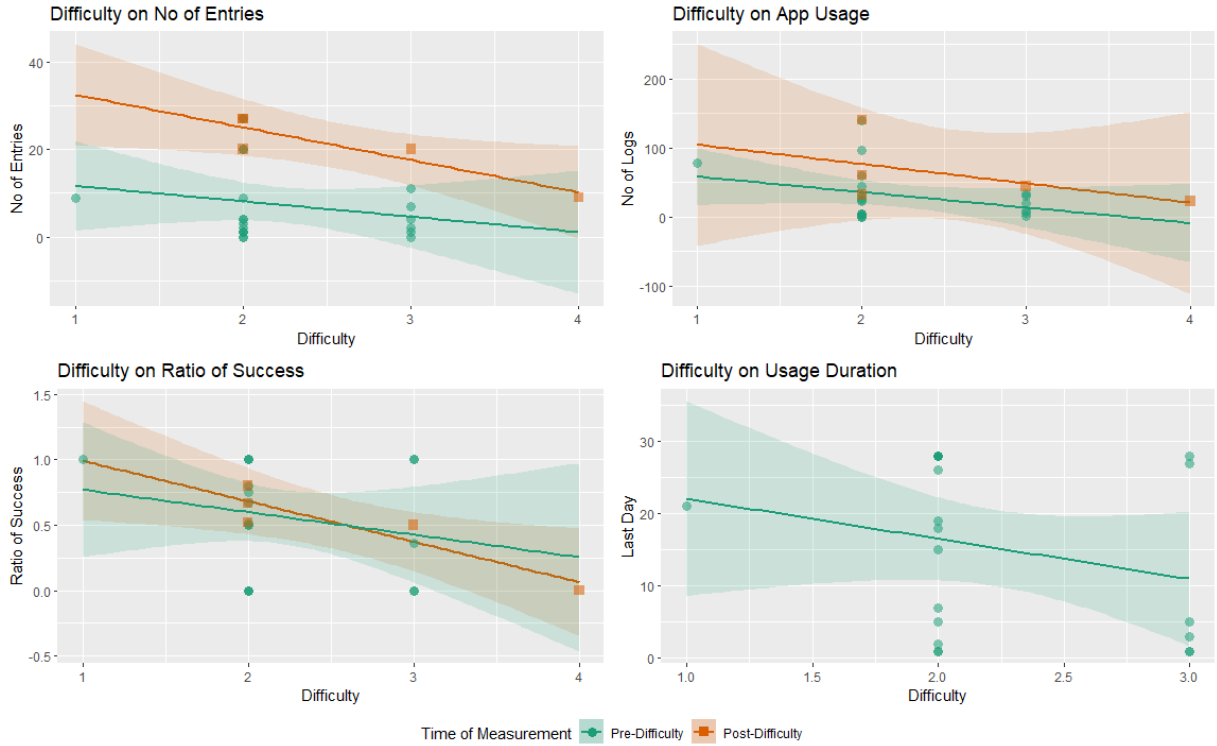
Linear Models of Motivation on Outcome



Notes. Linear models of motivation pre-experiment (n = 21) and post-experiment (n = 5) on the outcome variables a) No of Entries, b) No of Logs, c) Ratio of Success and d) Last day of app usage. The relation between post-experiment motivation in d) is left out due to the implied requirement of having made it to day 28 to be able to indicate the post-measures. See Table 3 for the confounding relations between the different outcome variables.

Figure D.2

Linear Models of Difficulty on Outcome



Notes. Linear models of difficulty, both pre-experiment (n = 21) and post-experiment (n = 5) on the outcome variables a) No of Entries, b) No of Logs, c) Ratio of Success and d) Last day of app usage. The same limitations as in Figure D.1 apply.

Table D.1

Effect of Pre-Experiment Difficulty on Outcome

	Dependent variable:			
	noOfEntries	noOfLogs	ratioOfSuccess	lastDay
	(1)	(2)	(3)	(4)
difficulty	−3.5	−22.3	−0.2	−5.5
	(3.7)	(14.6)	(0.2)	(4.8)
Constant	15.2*	80.5**	0.9**	27.6**
	(8.4)	(33.6)	(0.4)	(11.0)
Observations	21	21	18	21
R^2	0.05	0.1	0.1	0.1
Adjusted R^2	−0.004	0.1	−0.01	0.02
Residual Std. Error	8.8 (df = 19)	35.2 (df = 19)	0.4 (df = 16)	11.5 (df = 19)
F Statistic	0.9 (df = 1; 19)	2.3 (df = 1; 19)	0.9 (df = 1; 16)	1.3 (df = 1; 19)

Note: *p<0.1; **p<0.05; ***p<0.01

Table D.2

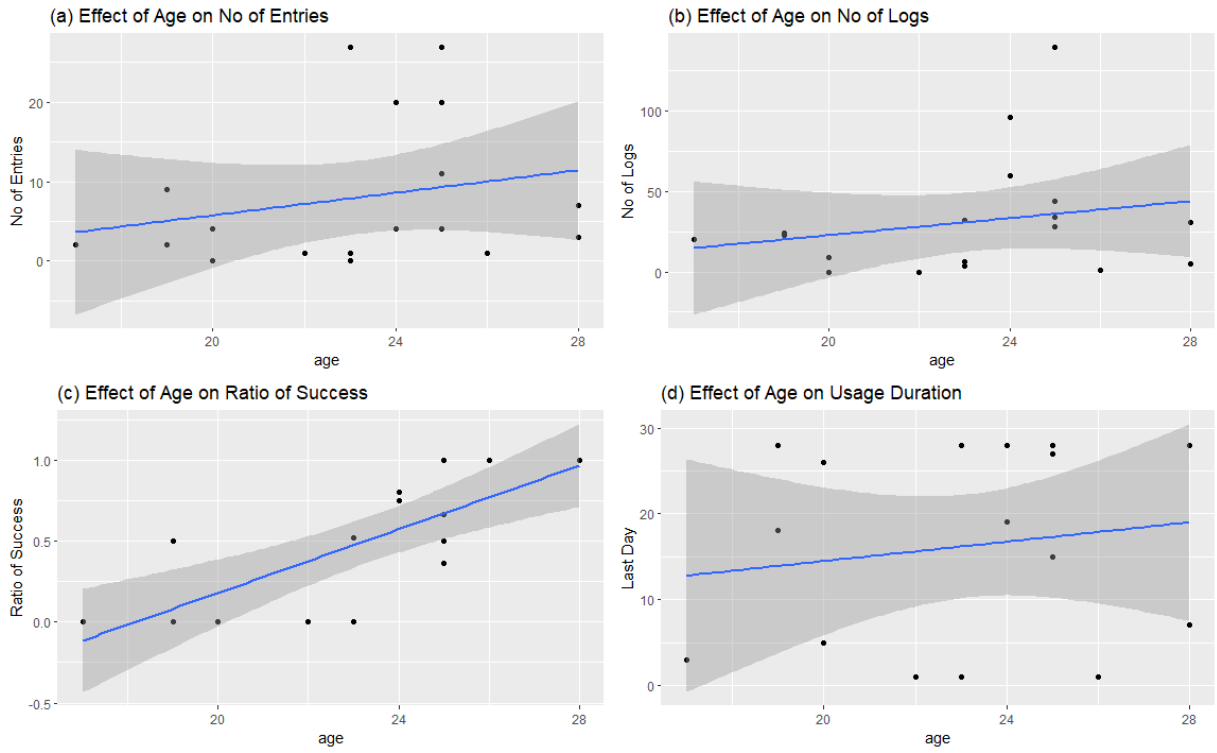
Effect of Post-Experiment Difficulty on Outcome

	Dependent variable:			
	noOfEntries	noOfLogs	ratioOfSuccess	lastDay
	(1)	(2)	(3)	(4)
postDifficulty	-7.4^{**}	-27.9	-0.3^{**}	0.0
	(2.0)	(25.6)	(0.1)	(0.0)
Constant	39.9^{***}	132.4	1.3^{***}	28.0^{***}
	(5.6)	(69.7)	(0.2)	(0.0)
Observations	5	5	5	5
R^2	0.8	0.3	0.8	
Adjusted R^2	0.8	0.04	0.8	
Residual Std. Error (df = 3)	3.7	45.9	0.1	0.0
F Statistic (df = 1; 3)	13.2^{**}	1.2	15.1^{**}	

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Figure D.3

Effect of Age on Outcome Measures



Notes. Linear models of age on the outcome variables. One participant was excluded from the analysis due to being an outlying age of 62, therefore leaving $n = 18$ participants for **(a)**, **(b)** and **(d)**, as well as $n = 16$ for **(c)**. Age ranged between 16 and 28 years.

Table D.3

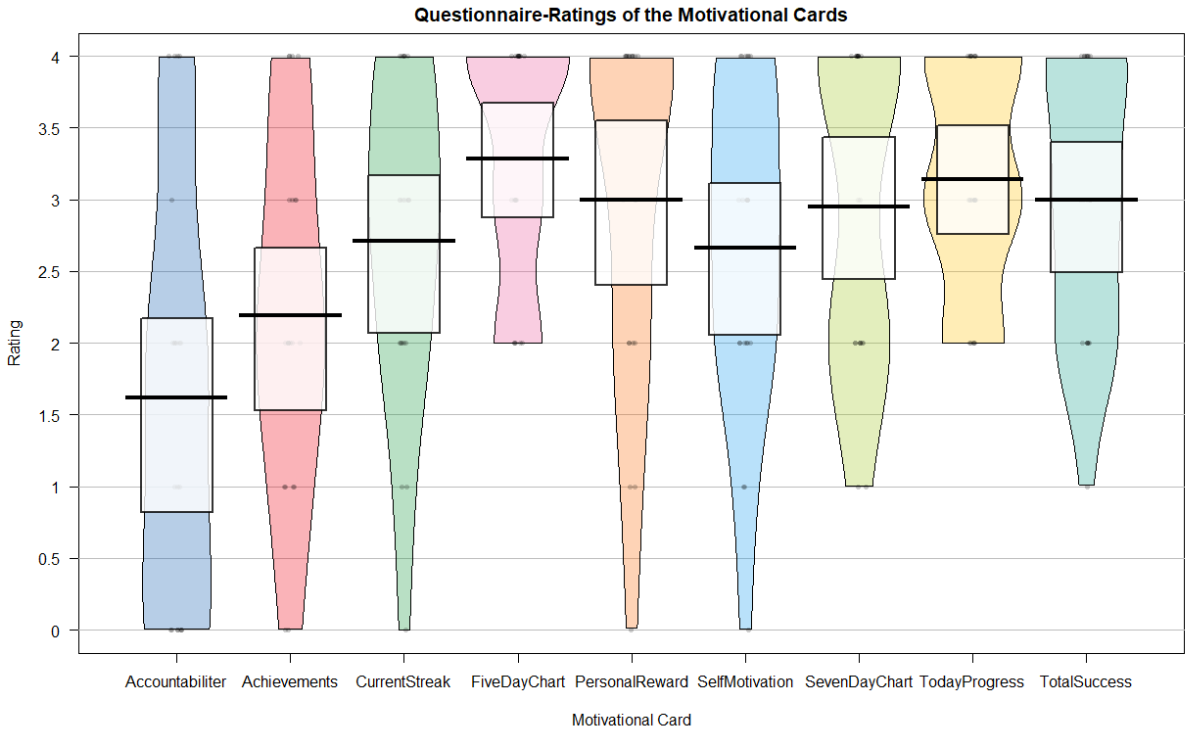
Effect of Age on all Outcome Measures

	Dependent variable:			
	noOfEntries	noOfLogs	ratioOfSuccess	lastDay
	(1)	(2)	(3)	(4)
age	0.71	2.65	0.10***	0.57
	(0.72)	(2.87)	(0.02)	(0.94)
Constant	−8.44	−30.31	−1.78***	3.16
	(16.78)	(66.84)	(0.50)	(21.92)
Observations	18	18	16	18
R^2	0.06	0.05	0.61	0.02
Adjusted R^2	−0.002	−0.01	0.58	−0.04
Residual Std. Error	9.21 (df = 16)	36.70 (df = 16)	0.26 (df = 14)	12.03 (df = 16)
F Statistic	0.97 (df = 1; 16)	0.85 (df = 1; 16)	21.47*** (df = 1; 14)	0.36 (df = 1; 16)

Note: *p<0.1; **p<0.05; ***p<0.01

Figure D.4

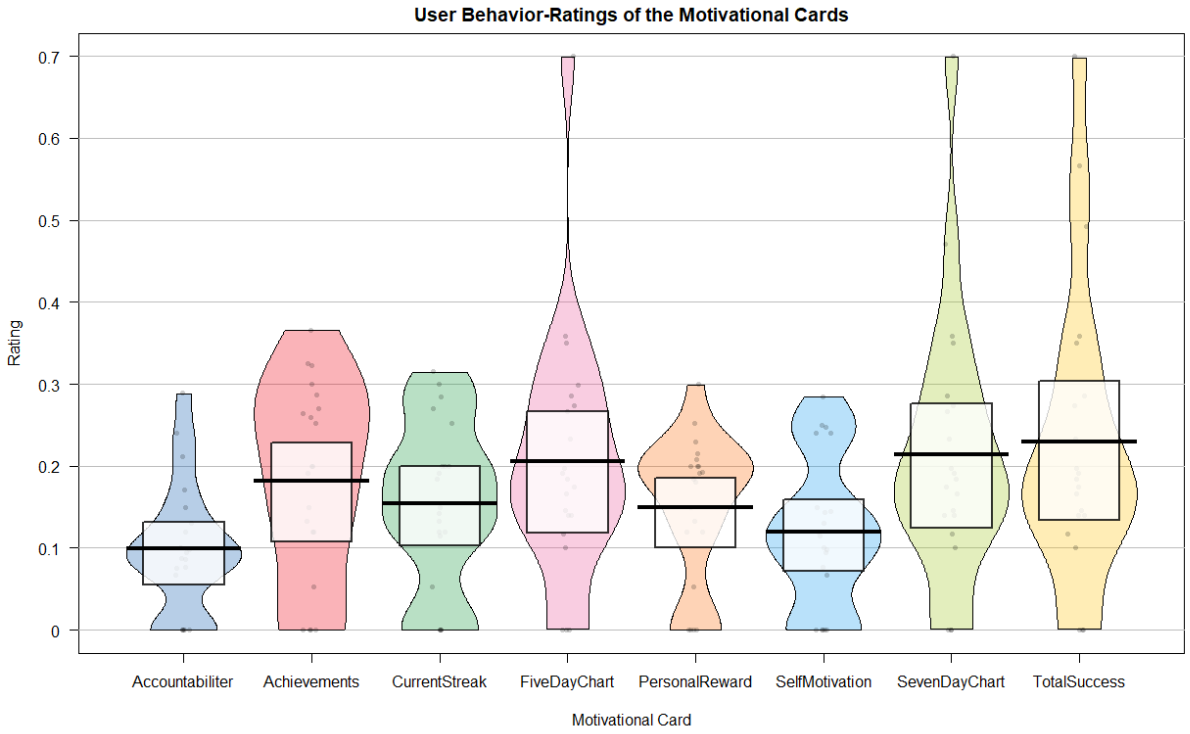
Questionnaire-Ratings of Motivational Cards



Notes. Ratings provided between 0 (Not at all) and 4 (Extremely). $n = 21$.

Figure D.5

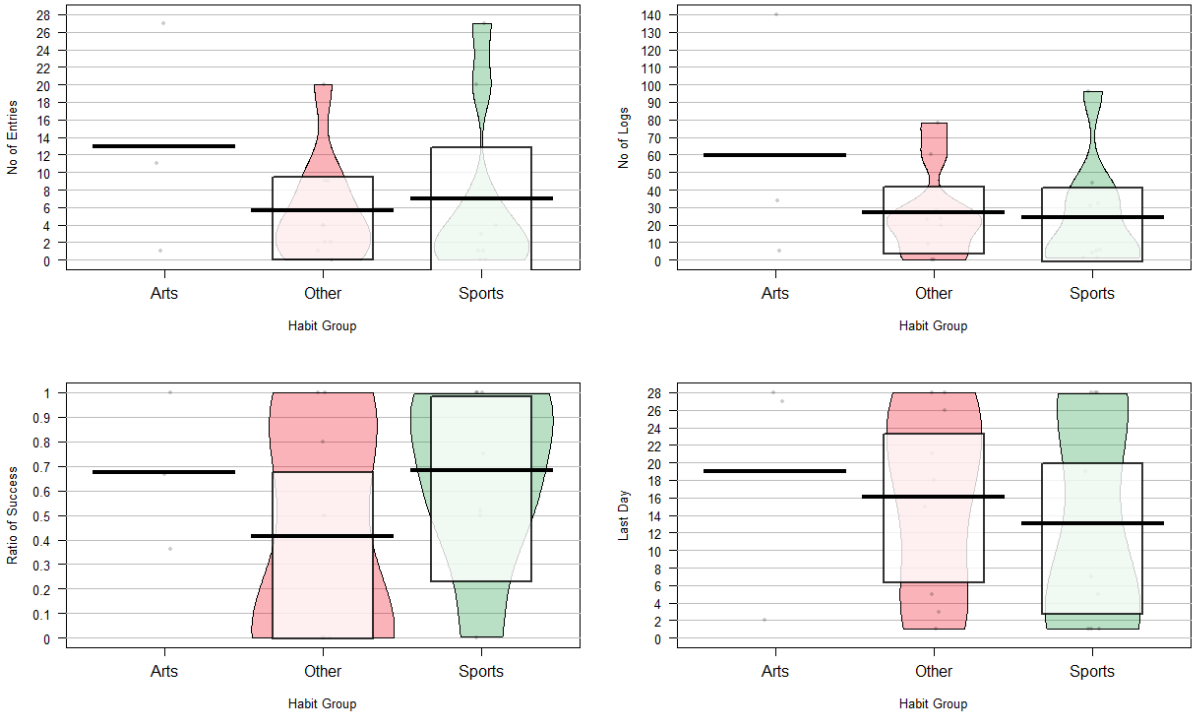
User Behavior-Ratings of Motivational Cards



Notes. $n = 21$.

Figure D.6

Outcome Differences between Habit Groups



Notes. Groups were selected post-hoc. Sample size for the groups Sports and Others was $n = 9$ each. The number of participants pursuing an art habit was too low ($n = 3$), to gain a representation of the SD.

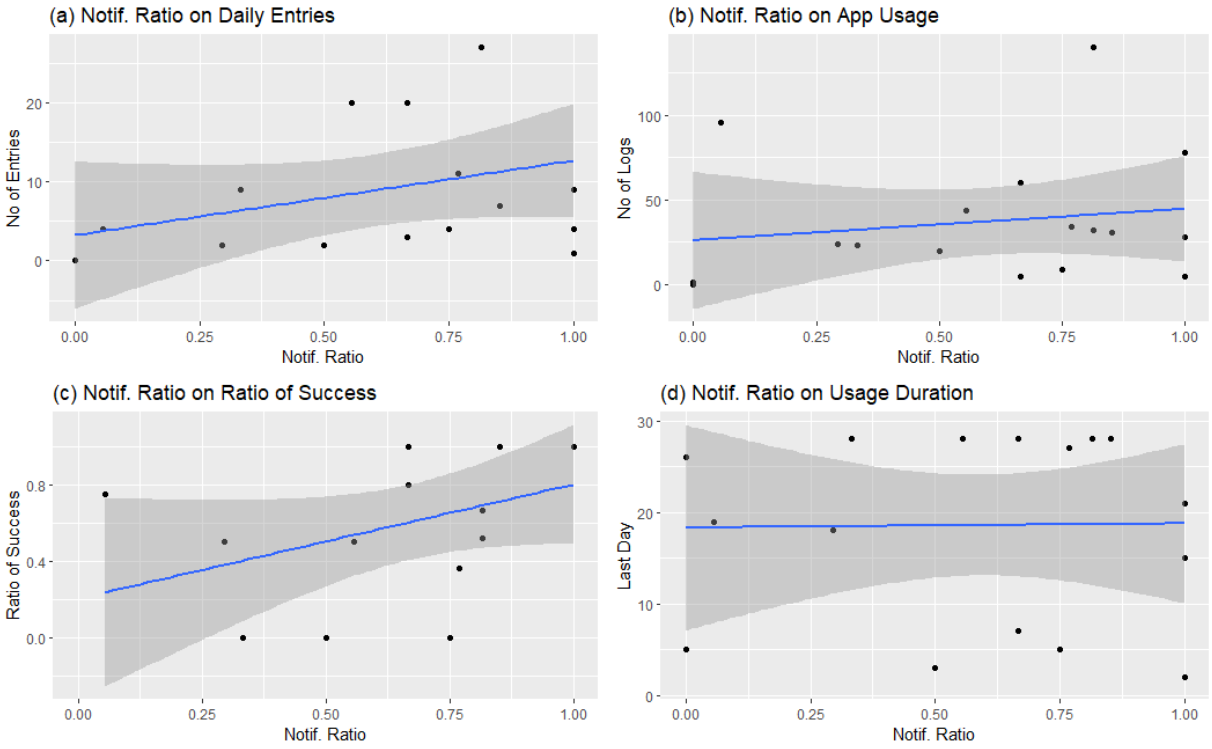
Table D.4

Kruskal-Wallis Comparison of Habit Groups on Outcomes

	No of Entries	No of Logs	Ratio of Success	Last Day
Kruskal-Wallis K	1.12	1.12	1.55	0.45
Observations	2	2	2	2
Asymp. Sig.	0.57	0.57	0.46	0.8

Figure D.7

Effects of the Ratio of Shown Notifications on Outcome



Notes. Linear fitting of the effect of the ratio of shown notifiers on all outcome measures, with **(a), (b), (d)** having $n = 17$ and **(c)** having $n = 14$.

Table D.5

Effects of the Ratio of Shown Notifications on Outcome

	Dependent variable:			
	noOfEntries	noOfLogs	ratioOfSuccess	lastDay
	(1)	(2)	(3)	(4)
notifRatio	9.43	18.55	0.60	0.45
	(6.41)	(27.87)	(0.34)	(7.74)
Constant	3.24	26.07	0.20	18.32***
	(4.36)	(18.95)	(0.25)	(5.26)
Observations	17	17	15	17
R^2	0.13	0.03	0.19	0.0002
Adjusted R^2	0.07	−0.04	0.13	−0.07
Residual Std. Error	8.81 (df = 15)	38.30 (df = 15)	0.35 (df = 13)	10.64 (df = 15)
F Statistic	2.16 (df = 1; 15)	0.44 (df = 1; 15)	3.11 (df = 1; 13)	0.003 (df = 1; 15)

Note: *p<0.1; **p<0.05; ***p<0.01