Countering Concurrent Login Attacks in "Just Tap" Push-based Authentication: A Redesign and Usability Evaluations

Jay Prakash[1, 2], Clarice Chua Qing Yu [1, 2], Tanvi Ravindra Thombre[2], Andrei Bytes[2], Mohammed Jubur[3], Nitesh Saxena[3], Lucienne Blessing[2], Jianying Zhou[2], and Tony Q.S Quek[2]

[1] Silence Laboratories, Singapore
[2] Singapore University of Technology and Design, Singapore
[3]University of Alabama at Birmingham, USA

***Abstract*—In this paper, we highlight a fundamental vulnerability associated with the widely adopted "Just Tap" push-based authentication in the face of a concurrency attack, and propose the method REPLICATE , a redesign to counter this vulnerability. In the concurrency attack, the attacker launches the login session at the same time the user initiates a session, and the user may be fooled, with high likelihood, into accepting the push notification which corresponds to the attacker's session, thinking it is their own. The attack stems from the fact that the login notification is not explicitly mapped to the login session running on the browser in the Just Tap approach. REPLICATE attempts to address this fundamental flaw by having the user approve the login attempt by replicating the information presented on the browser session over to the login notification, such as by moving a key in a particular direction, choosing a particular shape, etc. We report on the design and a systematic usability study of REPLICATE. Even without being aware of the vulnerability, in general, participants placed multiple variants of REPLICATE in competition to the Just Tap and fairly above PIN-based authentication.**

1. Introduction

Push notification based authentication, such as seen in solutions like, Duo-Push [1] or Authy [2], has witnessed a sharp rise in adoption in the past few years. It has been deployed as second-factor authentication (TFA) or password-less authentication. A device is first enrolled as a token device and associated with an *(account, service)* pair. Next, whenever a user attempts to log in to an application or web-service, and enters the correct credentials, the token device receives a push notification. When the user opens/taps on the notification, a screen overlay requests if the user wants to approve or deny the login attempt (Figure 1). The usability pain point is well relieved by this "Just Tap" push-based authentication compared to traditional one-time PIN (OTP) based TFA as there is no need to copy the PIN code from the device to the login terminal/browser. Hence, being more usable than OTP-based TFA, push notification assisted authentication has witnessed growing user adoption as reflected in the success of Duo Security and commercial adoption by software and service giants like Twitter, Yahoo, Google [1], [3], [4] and academic entities.

However, Just Tap push-based authentication has a fundamental and easy-to-exploit vulnerability, which we call the "concurrency attack." In this attack, the malicious actor launches the login session at the same time as the user. The user may then be fooled, with high likelihood, into accepting the push notification corresponding to the attacker's session, allowing the attacker to successfully access the user's account. This will break the second-factor security offered by Just Tap TFA, assuming the attacker has already compromised the first factor, the password (e.g. via hacked password databases). In the case of Just Tap password-less authentication, the attacker does not need to compromise the user's password. In the concurrency attack, the attacker's goal is to confuse the legitimate user with two or more similar push notifications. As represented in Figure 3, the push notification prompts ask the user to tap on the "Yes" or "No" button. The only differentiating information from login attempt of an attacker is the location name, usually given as coordinates. However, this information is too coarse and gives ample scope of the attack. Also, the attacker can spoof the location. The legitimate user would have no definite way to identify the correct push notification or even the possibility of an adversarial login, and they will most likely approve the attacker's notification.

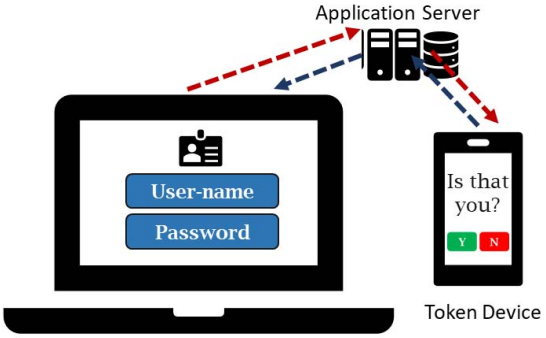

Figure 1. Conventional push-based, Just Tap to authenticate, TFA

To better understand this attack context, we simulated the situation of concurrent logins with 75 pairs of legitimate users and the attacker. The study set consisted of diverse personas, including undergraduate students of different streams, faculties, corporate employees from both business and information technology(IT) domains, and retired and old users. Statistically, only 5% of people (mostly comprising of IT employees and a few university students) raised doubts when receiving such notifications in the concurrency attack. Most of the people approved

2021 IEEE European Symposium on Security and Privacy (EuroS&P) | 978-1-6654-1491-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/EUROSP51992.2021.00013

DOI 10.1109/EuroSP51992.2021.00013

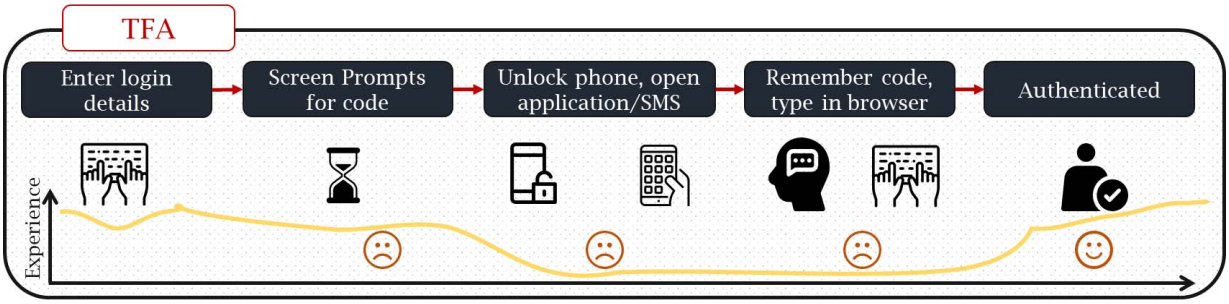

Figure 2. UJM, Experience vs time, of traditional OTP and TOTP based TFA suggests falls in user experience while attempting to authenticate a login.

both *approve* attempts of the push notification. We concluded that their action was due to the consistency in the user-interface (UI). Regular users of this form of authentication have formed a habitual reflex to tap *"Approve"* without looking at any other details. This habituation effect is further corroborated by studies conducted in [5], [6] where researchers suggest that most security notifications gets un-noticed due to the habituation of users. Some users find it normal to receive multiple near-concurrent push notifications as they are habituated to being stuck in such loops due to mobile network discrepancies and delays. We also conducted automated tests to investigate concurrency attacks on deployed platforms. We noted that the concurrency attacks affect DuoPush, Dropbox, Facebook, LastPass, TransferWise, Authy, Okta and many more services. Fig. 3 and Fig. 4 demonstrate concurrency attack scenarios for Dropbox, Duo-push and TransferWise.

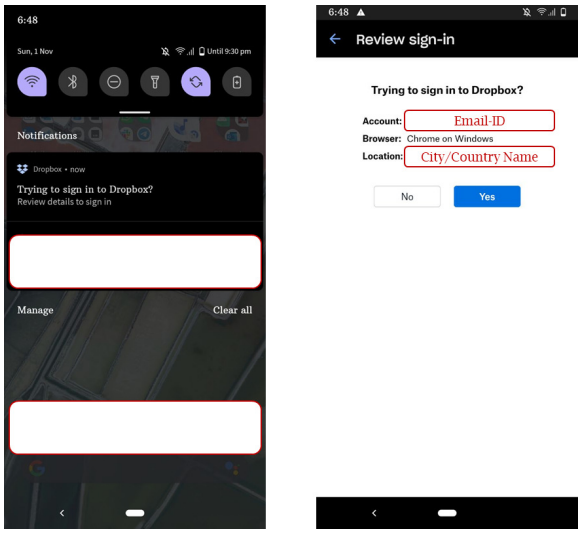

Figure 3. Response to concurrency attack on Dropbox a) push notification prompt and b) the push notification UI: the push notifications for two logins do not have vivid differences for across login attempts

The concurrency attack stems from the fact that the login notification is not explicitly linked to the login session running on the browser in the Just Tap approach to push authentication. Therefore, if the user receives two notifications simultaneously, there is no way to tell which of the two notifications actually correspond to the browser session that the user initiated. In order to address this basic design flaw with Just Tap push-based authentication while still retaining the underlying usability of the approach, we

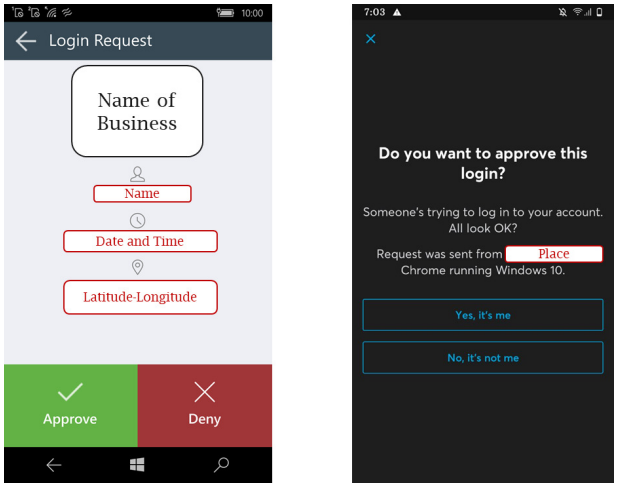

Figure 4. Concurrency attack on Duo and TranferWise: the differentiating information, place and location co-ordinates, are too coarse; latitude-longitude information is not comprehensible enough as well

propose a new design called REPLICATE. REPLICATE explicitly binds the user session with the push notification by having the user quickly approve the login attempt via replicating the randomized information presented on the browser session over to the login notification, such as by moving a key in a particular direction, choosing a particular shape, etc. The user interaction in REPLICATE is still very simple and easy. Yet, it serves to prevent the concurrency attack as the interaction required to approve the user session will not match with the interaction required to approve the attacker's session.

We designed REPLICATE and conducted a usability study evaluating its effectiveness compared to the Just Tap approach. We also developed a new table-top measure of user experience: a derivative of the emotion chart, and hybridized emoticons and emotion ranges for the user journey map. Our quantitative and qualitative results suggest that, even without disclosing the vulnerability in question, in general, participants of our study placed multiple variants of REPLICATE in competition with the Just Tap authentication. The analysis also suggests that participants liked the randomization and active engagements underlying the REPLICATE approach.

Main contributions: The research is based on the premise that the "Just-Tap-to-Authenticate" method is an extremely popular, lowest-effort scheme, and adopted by Duo-Security, Authy etc. Hence, it is imperative to im-

prove its security while maintaining similar usability. Our key-contributions:

- We propose that simple and usable design interventions can solve this vulnerability, and suggest a set of possible designs.
- We empirically demonstrate that the proposed designs stand close in usability and higher in terms of security when compared with "Just-Tap" and PIN-based authentication.
- We do a quantitative and qualitative evaluation of proposed designs using a combination of system usability scale (SUS) and user-journey mapping methodology.
- We introduce steps for remote-usability-study.
- We propose and use a hybrid way to study user-experience through a fusion of emojis and a color-intensity-based emotion chart.

2. Our Design for Concurrency Resilience

While designing REPLICATE, we faced a twofold challenge: a) to enhance the security of push based two-factor authentication against concurrency attacks and b) to ensure that we still maintain high usability. We propose a set of novel security-inclusive UI interventions as part of the REPLICATE family. As we shall observe in later sections, they do better than the existing push-based TFA as far as security is concerned and maintains almost a similar profile across usability metrics.

2.1. System Assumptions

The underlying assumption governing the design of REPLICATE, which are congruent with existing TFA designs as well, are as follows:

- The legitimate user, who is attempting to login to a service or application, has a registered token device with himself/herself if the login device is different from the token.
- The token should have either a camera or a keypad to support the registration process.
- The user can see the login page, and cognitively respond to the instructions shown at the login prompt.
- The attacker can monitor the legitimate user remotely or while being in proximity and should not gain any advantage.

2.2. Threat Model

REPLICATE's threat model assumes that the attacker has access to the login credentials (username and password) of the legitimate user and aims to gain access to the account. The credentials can be learned from leaked password databases or other methods. While such motives can be challenged by the state-of-the-art TFAs to a certain extent, REPLICATE assumes a powerful attacker who is also aware of the time window in which the legitimate user would attempt login (using contextual information, physical proximity to the user or by monitoring network traffic) and hence aims to exploit the vulnerability associated with concurrent login in push-based Just Tap TFA. As such, the attacker cannot have access to the token device.

2.3. Trivial but not so Usable Solutions

A trivial solution to counter concurrency attack can be to avoid multiple logins at the same time and keep track of the locations of logins. It turns out that such mechanisms would not be feasible where network connection at the end of the user is not good or there is deliberate logins in same time window, i.e., before the previous push becomes invalid. A tighter implementation will lead to loss of services to the companies at the cost of user experience. Also, Duo and other push-based TFAs suggest the login location on the notification prompt -location's name along with latitude and longitude - on the push screen. While such information can be spoofed [7], it is too demanding to expect users to remember their geographical co-ordinates with such high granularity.

The usability of our security interventions, REPLICATE, would undoubtedly be a crucial factor in determining its eventual adoption by users. We have, therefore carefully considered the user experience and accessibility. A human-centric design process was applied in the steps of initial ideation of the ways in which the user could give an input based on the login prompts, and then in the steps of evaluating and deciding the best method. In the following subsections, we will be going through the process followed in the development of the concepts and evaluation for the proposed solutions. We embarked on a series of needs understanding, competitor benchmarking, feature rationalization and iterations to determine intuitive prototype designs which would deter possibilities of concurrency attacks.

3. Proposed Design and Security Intervention

Abiding by the nature of the vulnerabilities in push notification based TFA, we propose design and security interventions in the form of a closed-loop feedback-based push TFA- REPLICATE. REPLICATE proposes a set of designs and aims to explore them through the lens of security and usability. The key idea behind REPLICATE is based on the basic premise that a user finds it difficult to distinguish between the legitimate and the compromised push. We believe that the main reason behind the inability to separate a legitimate push message from the attack is two-fold - invariance in the UI and repetition of the exact same steps in each successive login attempts by the user [5].

REPLICATE proposes to adopt design interventions wherein users would be expected to perform an interaction on the token device which should match the interaction suggested by the login prompt of the web-browser or in-app application. Since the suggested interactions are randomly selected for each login attempt, the expected way in which a user will interact with the push notification's UI prompt at the token device will vary for the legitimate and attacker's login trials. The core philosophy is based on the fact that only a legitimate user can simultaneously see the login screen and push overlay on the registered token device, and an attacker has no control over what activity is chosen by the REPLICATE service. In the next subsections, we will describe the list of proposed variants of push notification in Push-Auth. These are expected to replace the existing tap to "*Approve*" or "*Deny*", which

pops up on the registered token device on a successful entry of the login details.

3.1. Key-Drag

As shown in Fig.5, Drag-to-Auth is an on-screen drag-assisted TFA where the browser or login screen suggests to the user to drag an icon (a key in this case) on the screen overlay of the push notification, appearing on the registered token device. As shown in Fig. 5, the browser suggests to drag the key in the *upward* direction, randomly chosen out of multiple directions, and the user is expected to do the same on the token device (on the right hand side). As shown in Fig. 6, the REPLICATE service selects one of the directions from 8 possible directions of dragging.

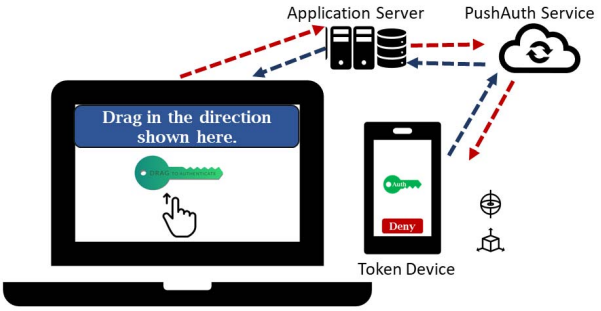

Figure 5. Login side of Drag-to-auth would suggest a direction to drag the key. The user is supposed to drag the key icon, which comes as a screen overlay in the push notification at the token device, in the suggested direction.

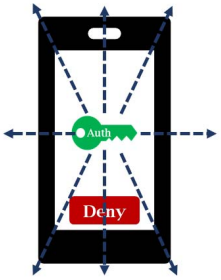

Figure 6. REPLICATE service's drag-to-auth mode randomly suggests one of these directions to drag the key

3.2. Move-a-Shape

In the second variation, users are randomly shown a shape on the login screen, (Fig 7 (a)), along with the direction in which they have to drag it on the push notification prompt. But unlike key dragging, we have raised the bar for user engagement by suggesting multiple shapes on the authentication overlay. Users will see 4 shapes on their phone (Fig 7(b)) and have to drag the correct shape in the correct direction.

3.3. Randomized-Keypad

In the second variation, REPLICATE service randomly generates a numeric key which the user is expected to key in using the numeric keypad which comes as screen overlay. As shown in Fig. 8(a), the keypad in the notification is also randomized by purpose. Randomized keypad shows users a 2-digit number on the laptop screen (Fig 8 (a)), and users have to type out the number on their phone (Fig 8 (b)).

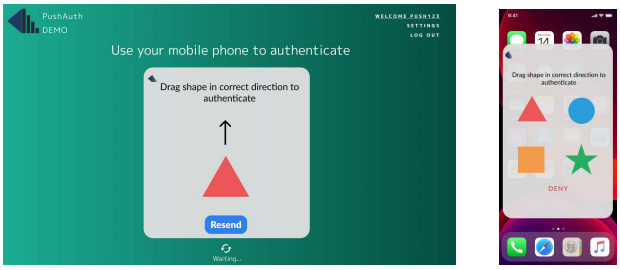

Figure 7. Move a Shape to Auth (a) Login window display and (b) Phone authentication push screen overlay

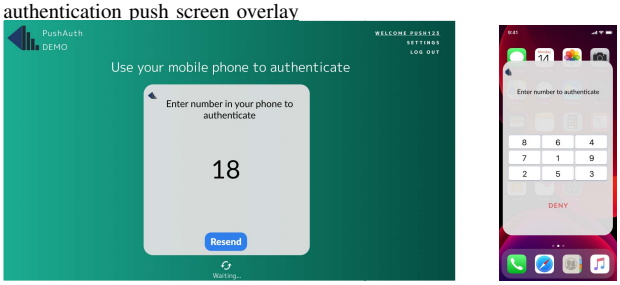

Figure 8. Randomized Keypad to Auth (a) Login window display and (b) Phone authentication push screen overlay

3.4. Choose-a-Colored-Button

Colored Button instructs users to tap on the button of a certain color, Fig 9(a). To authenticate, the user will have to tap on the button, in the notification prompt of the phone, of suggested color on their phone, Fig 9(b).

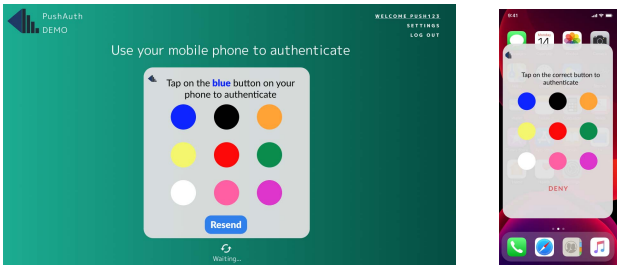

Figure 9. Choose a Colored Button to Auth (a) Login window display and (b) Phone authentication push screen overlay

3.5. Tap-on-Black-Button

In another version of REPLICATE, the user would be suggested to tap at the suggested location of a black colored button. Users would be expected to tap on the location on their phone (Fig 10(a)) which should conform to the black button suggested on their laptop, (Fig 10(b)).

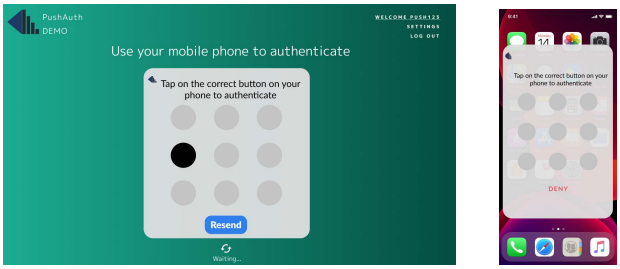

Figure 10. Tap on Black Button (a) Login window display and (b) Phone authentication push screen overlay

3.6. Draw-Shape

Draw Shape approach shows users an unlock pattern, similar to Android phone's unlock screen, on the login screen (Fig 11(a)). The user is expected to draw the correct pattern on their phones (Fig 11).

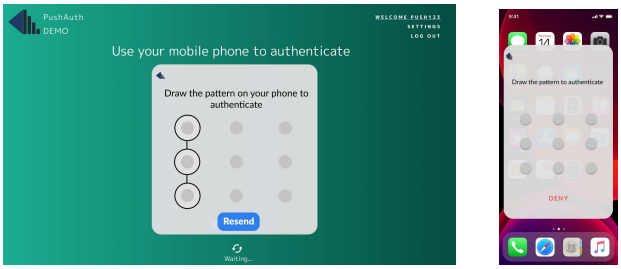

Figure 11. Draw Unlock Pattern (a) Login window display and (b) Phone authentication push screen overlay

4. Architecture

Fig. 12 shows the steps the REPLICATE system goes through. The system comprises of a user application (or browser), application server, REPLICATE service and the registered token device. When a user types in the correct login credentials, the application server commands the REPLICATE service to initiate the second-factor authentication for the corresponding user. The service responds by sending a push command to the token device, which is reflected as an interactive push overlay on the screen of the token device. At the same time, the browser is prompted to show an activity, chosen randomly from an *activity set*, on the screen. Each *activity-set*, comprising of finite possibilities, corresponds to one of the multiple prototypes discussed in REPLICATE. The browser suggests the user to replicate that activity in the push overlay on the token device. If responded within an acceptable time window, the token device transmits the Activity ID to the Browser and device sensor readings to the REPLICATE service. If the Activity ID matches with what was shown in the browser/application screen, the service instructs the application server to further grant the access. In the subsequent sections, we will be describing multiple versions of the protocol wherein different activity sets are suggested and discussed through the lens of security and usability. It is to be noted that the proposed architecture falls in line with the Just Tap method and can be adopted by the introduction of the random number, matched to the random activity-set, generated on the server.

4.1. Resilience to Concurrency attack

Overall the security of REPLICATE comes from *different* push overlays for concurrent logins. Since expected interactions are randomized, the legitimate user will be suggested to interact with the push in a different manner from that of the attacker's push. If we randomize the selection of action and show it on the login screen, the legitimate user will have two advantages: a) he/she will know what to do with the push prompt on the token device b) if there are concurrent logins, attack, she/he will only perform what is shown to the screen before her/him and not the one at attacker's login screen.

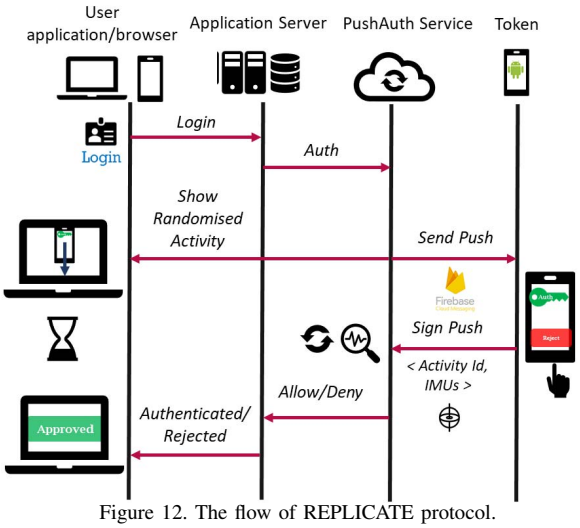

Figure 12. The flow of REPLICATE protocol.

5. Usability Study Preliminaries

We conducted a month-long study on variants of REPLICATE at our university. The research objective of the study was to compare the usability of the proposed substitutes of existing push-based second-factor authentication mechanisms. The objective of the study was to find out:

- if people understand the authentication instructions given,
- if people are able to carry out the authentication process successfully without much help,
- if people find our proposed authentication methods more user-friendly than existing authentication solutions, and
- which authentication method (among those that we have proposed) results in the most positive user experience?

Due to limitations of social distancing during the COVID-19 pandemic, the participants were invited for remote usability testing over a period of 3 weeks. Each online session lasted for approximately 1 hour, and there were a total of 40 participants.

5.1. Remote Study Setup Design

This subsection will discuss details of the study setup, components and what settings were used.

A. Prototype of Login Portal and Authenticator Application: We simulated a bank balance details inquiry wherein the users would log in to a prototype webpage with pre-shared username and password pairs. An interactive web portal, as shown in Fig. 13, was designed using Figma [8]. To emulate authenticator apps, which are expected to receive the push notifications and the designed UI prompts in REPLICATE, Figma prototypes for smartphones were designed. For randomization of the order, the prototypes were put in queues, and the participants were given the prototype with an index suggested by a random-number-generator.

B. Rationale Behind Using Figma: The study was fully remote. We chose Figma because it allows sharing prototypes with participants without needing them to create an account. Participants were not required to install any

application as we could share links to the prototypes and mirror the experience of interacting with an application on the participant's side. Also, since the Figma prototypes are inherently hosted on the cloud, we could monitor the participants' activities and interactions in real time, i.e. we could observe participants remotely via the Observation Mode. Figma is a sufficiently close alternative because the required gestures, tapping-on-screen and dragging-in-a-straight-line, were supported by Figma. Also, for the key dragging gestures, we did preliminary testing and ensured that it works well on all phones, especially noting the margins.

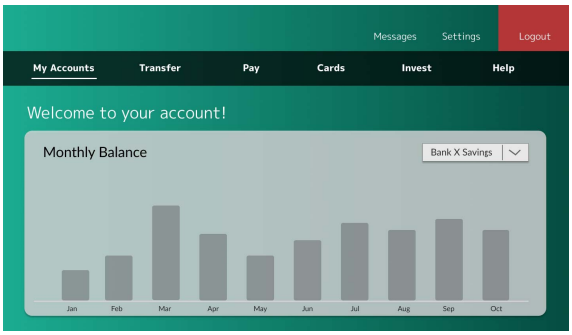

Figure 13. An Interactive Web Portal, emulating a banking service, was designed.

C. Preparation: The study, involving a facilitator/observer and the participants, was done through Zoom [9]-based video conference. The following materials and facilities were prepared before conference calls:

1) Creation of pre-test and post-test survey questions, overall post-test interview questions
2) As this was a remote study, we reached out to potential participants and ensured that they have laptops/desktops with Zoom installed, a smartphone with Telegram messenger installed and stable internet connection.
3) Scheduling a Zoom session with each participant.

Before describing the details of the processes, we will proceed with the description of the content of the surveys, questions and a novel technique to map user's experience through the authentication journey.

D. Pre-Test Survey: A short questionnaire was shared through Google form before the start of the test. It aimed to collect information about the participants, so that we could categorize them to find trends, and also to ensure that we had diverse participants. The questions can be enumerated as:

- What is your age group? a) 20 and below b) 21-30 c) 31-40 d) 41-50 e) 51 and above
- On a scale of 0 to 4, how would you rate your level of familiarity in using your laptop? (0 being the least familiar and 4 being the most familiar)
- On a scale of 0 to 4, how would you rate your level of familiarity in using your phone? (0 being the least familiar and 4 being the most familiar.) (Rationale: They will be using their laptop and phone to try out the prototypes. Their level of familiarity might have an impact on their speed/chances of error)
- Do you know what second-factor authentication is? a) Yes or b) No
- In which of the following areas have you used a two-factor authentication? You can select more than one option. a) I have not done a two-factor authentication before, b) Banking, c) Email, d) Social Media, and e) Others.
- Have you come across a push-based TFA? Some examples are shown in the image below. a) Yes b) No

■ **E. Post-Test Survey, Trial 1:** This questionnaire was designed in accordance with the System Usability Scale (SUS) and was shared through a Google form after the participant had tried each prototype during trial 1. SUS [10], [11] is a ten item questionnaire where participants get to choose on a five point scale from *strongly disagree* on the extreme left to *strongly agree* on the extreme right. A score out of 100 is generated by converting the user's response into a score from 0-4 for each of the 10 questions and then multiplying this score by 2.5 to convert it from a 0-40 to a 0-100 scale. This final score can be used for usability bench-marking [12], [13]. Research studies confirm the advantage of SUS in being simple and yet the most reliable (with at least 12-14 participants) in measuring user's reaction more holistically [14]. It is to be noted that our survey uses the all-positive version of SUS survey questions. Previous research work such as [15] has investigated the effect of positive and negative wordings while designing the SUS questionnaires and concluded that an all-positive version of the questionnaire can be used. In doing so, the participants are less likely to make mistakes between answering the alternating positive-negative questions and the researchers are less likely to make errors in coding as well. Furthermore, it gives scores similar to the standard SUS. In our study, the final SUS scores for the prototypes and the existing solutions are compared against each other, as well as the known benchmarks. This survey contains SUS questionnaire and a few questions to track user emotions. The questions can be enumerated as:

- I think that I would like to use this authentication method frequently.
- I found the authentication method simple.
- I found the authentication process easy.
- I think that I could use this authentication method without the support of a technical person.
- I found the various functions in this system were well integrated.
- I thought there was a lot of consistency in this system.
- I would imagine that most people would learn to use this authentication method very quickly.
- I found the authentication method very intuitive.
- I felt very confident using the authentication method.
- I could use this authentication method without having to learn anything new.
- How would you rate your overall experience with this authentication method for the laptop/desktop to phone authentication? (seven point scale from *very positive* on the extreme left to *very negative* on the extreme right)

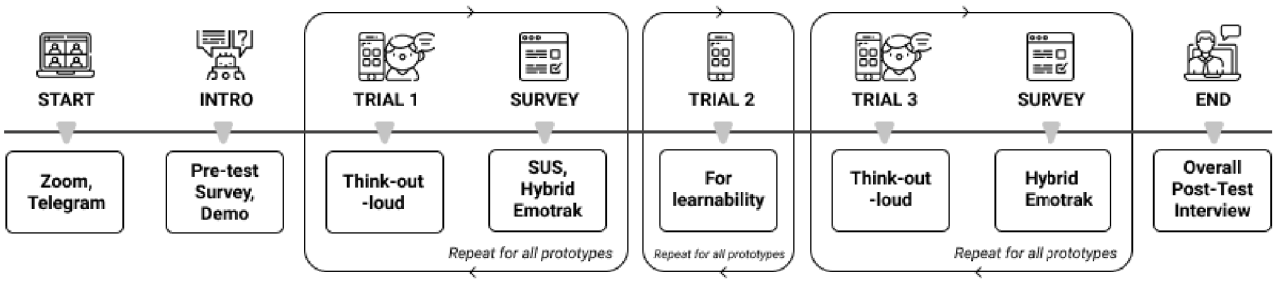

Figure 14. The designed study process for the remote user study.

- How would you rate your overall experience with this authentication method, for the phone to phone authentication? (Seven point scale from *very positive* on the extreme left to *very negative* on the extreme right)
- What did you like the most about this authentication method? Why?
- What did you like the least about this authentication method? Why?
- What, if anything, surprised you about the experience?
- Any other areas that you think can be improved on?

In order to track the emotions, the facilitator would show a chart after the participant answers the questions. The chart contains different emotions, described in Section 5.2.A. The participants have to select an emotion that best represents how they feel for a particular stage of the prototype. They can type the corresponding code of the emotion e.g. A1, A2, etc. The participants have to make this selection for three stages: a) reading and processing information, b) transitioning from laptop to phone, only in the case of login through laptops or desktops, and c) carrying out authentication action.

F. Post-Test Survey: Trial 3 This questionnaire was shared through Google form after the participant had tried each prototype for the trial 3. The question was: How would you rate your overall experience with this authentication method? (on a seven-point scale, from very positive on the extreme left to very negative on the extreme right). Again, similar to the survey for the trial 1, the participants were asked to choose emotions for three stages.

G. Overall Post-Test Interview: These questions were asked after the participant had completed all three trials. The questions can be enumerated as:

- Of all the authentication methods you have tried today, which is your favorite? (The screenshots of all the prototypes are shown to the participant in the survey form.)
- Why is that authentication method your favorite?
- Which is your least favorite?
- Why is it your least favorite?

The SUS score based approach described in previous subsections gives a good idea about overall usability scores. But SUS has its limitations due to the granularity of information about usability. One of the limits of the SUS is that while it allows usability bench-marking, it does not provide us with the reasons for the score. Such knowledge would require qualitative analysis tools which would facilitate ways to capture pains and emotions at each stage of interaction with the authentication process.

5.2. Generating User Journey Map (UJM)

In order to address the aforementioned challenges, we complimented SUS with a user journey map (UJM) because it helps in identifying pain points at each step and gives a representative picture of experiences as the steps are performed by a user. While users were trying out the prototypes, we noted their journey and also asked questions for mapping their user experience.

First of all, we conducted an internal evaluation with the help of 5 evaluators. This is in line with Nielsen's heuristic evaluation [16] which is a usability engineering method for discovering usability issues in a user interface design using a set of recognized usability principles which are also called "heuristics". Early evaluation helps in fixing potential problems through an iterative process. We found some principles that had been violated, which lead us to improve our design. These violations were fixed before the eventual usability testing with users. We identified the importance of visibility, user control and freedom. For example, the lack of visibility of any notification to open the authentication overlay on the phone-to-phone authentication could make users confused. Hence, we incorporated a pop-up notification to tell users to open the notification overlay.

A. Emotion Chart: Deriving the Experience Timeline: To obtain the graph detailing users' emotions, we conceptualized an emotional journey chart - *Hybrid Emotrak*. *Hybrid Emotrak*, Fig 15, helps to gather data about the emotional journey of a product. During the post-test survey, participants were asked to select which emotional state they were in for the stages of reading instructions, transitioning from laptop to phone, and carrying out actions. Since each box on the chart has an associated score, we calculated the average score for each stage for each prototype and plotted the graph accordingly. We set a fixed emotion for the login stage and outcome of the authentication stage as neutral and extremely delighted, respectively, for completeness of the graph. The focus of this study is on how users interact with the second-factor authentication and not on the entering of login details at the login stage. We collected data for the transitioning from laptop to phone stage as well, but our analysis suggested that it

was not representative of what participants would feel if they were in a normal (non-study and physical) settings. Hence, we fixed the emotion for this stage as neutral.

B. Rationale for Using an Emotional Journey Chart: We note that participants might face difficulties expressing how they feel. Asking questions to the participants, like "how do you feel?" will result in vague answers that do not bring much value to the study. Previous research studies have shown that participants are more reflective when they are picking emotions from a chart [17].

C. Designing Hybrid Emotrak: In order to counter difficulties associated with capturing emotions of the users, an agile self-reporting method, Emotrak, was proposed by UE group in [18]. Another popular emotional engagement tool is the Plutchik's Wheel of Emotions [19]. The wheel comprises of 32 different emotions with different levels of intensities. Although there are more emotions for participants to choose from, some of the emotions like "grief" or "sadness" might not be too relevant for our study. Furthermore, it is more challenging to quantify the data in an accurate manner, given the subjective nature of the emotions on the wheel. We chose Emotrak as a base template to design a hybrid version mainly because it allows us to quantify the data easily and accurately, and also because it makes the experience of selecting an emotion simpler without compromising on the nuances of emotions. We find that the eight selected emotions with different levels of intensities are comprehensive enough for us to yield meaningful insights and are yet very simple for participants to use.

We build on the top of Emotrak by adopting the use of emojis along with colors and names of the emotions. We chose to make a hybrid chart because numerous studies have shown how graphical representation can be useful when trying to understand users' emotions. We believe that the modified chart will be useful as people who are familiar with the English terms used in the chart can still rely heavily on the original UE Group's tool, while people who have difficulties understanding the words can make use of the graphics. Multiple studies [17], [20] have support the use of emojis for emotion's capture. *EmojiGrid* was presented in [17] for studying food elicited emotions. In addition, emojis can perhaps also help people who are comfortable with words to pick their emotion on the hybrid version of UE Group's chart faster. Also, a large section of the population is familiar with emojis, and there is quite a diverse range of emotions to choose from. Emojis are also "standardized", hence, eliminating any potential biases [17]. The emoji for each emotion on the chart was selected after we completed 3 surveys involving at least 30 people each wherein participant were requested to select emoji, which is the most suitable for a particular stated emotion. The options presented were narrowed down after each survey.

D. Hybrid Emotrak: The emotion chart in Fig 15 is a derivative of the emotional journey chart proposed by UE Group called Emotrak [18]. The original chart by UE Group has a spectrum of eight emotions from positive to negative, with different levels of intensities as well as a neutral column. In the proposed chart, colors are being associated with the emotions to help participants identify how they feel more quickly. Scale values are assigned to each emotion, but these values are not revealed to participants. In our modified chart, emojis have been added to the lowest and highest intensity box of each emotion, as seen in Fig. 15.

5.3. Study Process

Each session involved a facilitator and a participant. Throughout the session that lasted for approximately one hour, the participant was in a Zoom call on the laptop/desktop with the facilitator. Each participant was added to a Telegram messaging [21] application group chat where the links to all the surveys and prototypes done on Figma [8] have been pre-sent. Participants were told to click on the links at appropriate times, and each link was deleted by the facilitator upon completion to avoid confusion. Surveys were done by participants remotely on their phones without the observer observing via Zoom as studies have shown that participants tend to give more positive results if being observed [22].

In total 8 pairs of prototypes (login portal and Figma application) were studied. This included the commercial of "tap-to-approve" approach as well for comparison purposes. We created 2 versions - one for TFA involving mobile phones only, and the other for TFA involving both the laptop/desktop and phones. 30 participants tested the Laptop to Phone (L2P) prototypes, while 10 tested the Phone to Phone (P2P) prototypes. For phone-to-phone TFA , participants would open the links on their phones and try out the prototypes on their phones. For laptop-to-phone TFA, the observer would screen share the laptop component using Zoom and click the login button on the prototype on the participants' behalf. Upon logging in and reading the authentication instructions on the laptop component, the participant would be prompted to proceed with the phone component on their phones, which has the Figma prototype ready. Using Figma's observation mode, the observer would observe the participant's interaction with the prototype. Both the Zoom call and interactions with the prototypes on Figma were recorded.

There were 3 trials for each prototype (see section 6), and participants were asked to articulate their thoughts or whatever came to their mind as they tried out the prototype. This is in accordance with *think-out-loud protocol* [23]. After which, the participant was asked some questions regarding their articulated thoughts. Hybrid Emotrak was also shown to all participants for them to pick out the emotions that they felt at the respective stages of the prototype. The comments gathered during the *think-out-loud* process and the input gathered from the emotions chart were necessary for enabling us to analyze the user journey better. Participants were also asked to complete a post-test survey immediately after trying each prototype for the first and third trial. There was also a pre-test survey at the start of the study and an overall post-test interview right before the study ends.

For L2P, in each trial for each prototype, the participant would look at one Figma screen on the laptop and one Figma screen on their phone, whereas for the P2P case, the participant only looked at one Figma screen on their phone.

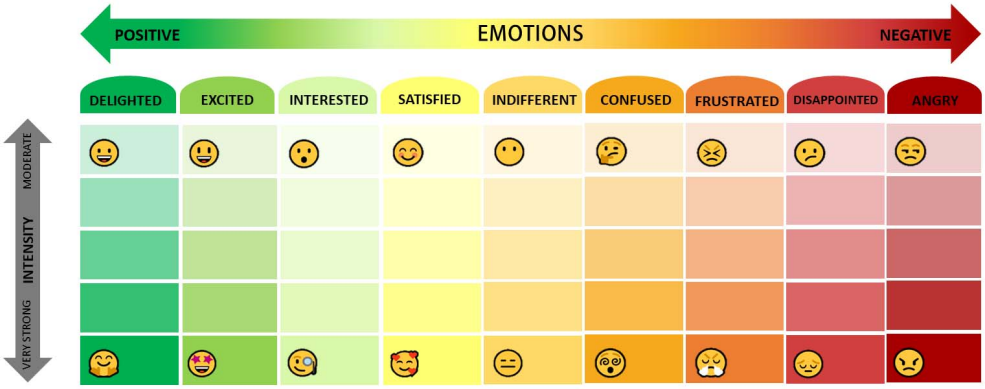

Figure 15. Hybrid Emotrak

A. Study Outline: Each session comprised of a observer and a participant. The flow of remote usability testing can be enumerated as follows:

1) Observer opens Figma on their laptop (A) and prepares for screen recording in order to capture the participant's interactions via Figma's observation mode.
2) Observer initiates Zoom call on another laptop (B) with the participant in the session.
3) Observer starts Zoom recording.
4) Observer adds the participant to the Telegram group chat with links to the prototypes and the surveys. Observer requests the participant to do a pre-test survey, remotely.
5) Observer explains the scenario, expected steps of the tasks and directs the participant to the correct link of the Figma prototypes.
6) Observer instructs the participant to open the smartphone prototype but not to look at it until the participant has completed the expected interactions with the laptop.
7) Observer performs a demo of the prototype for the participant to enhance familiarity.
8) Observer gets the participant to try trial 1 of the first prototype while conducting the thinking-out-loud process [23] before conducting the trial 1 post-test survey. In the post-test survey, the observer screen-shares the emotion chart for the participant to pick a square on the chart that represents how the participant feels for each of the identified stages. This process is repeated for all the prototypes.
9) The study proceeds with trial 2 for all the prototypes and then trial 3 and trial 3 Post-test survey for all prototypes. Participants were also asked to think-out-loud for trial 3.
10) Observer gets the participant to do the overall post-test survey, remotely.
11) After completing all 3 trials for all the variants of the prototypes, the observer conducts the overall post-test interview with the participant.
12) The participant is debriefed and the study is concluded.

5.4. Demographics

As the recruitment was done on a university's unofficial group chat, most participants were young and tech-savvy. 40% (16/40) of the participants were female, 75% (30/40) of the participants were between age 21-30, 22.5% (9/40) were 20 and below, and 2.5% (1/40) participants were between 31-40 years old. All participants indicated in the pre-test survey that they know what TFA is and have used a TFA before. This might be because the participants came from a technology University and had engineering backgrounds. The most common platforms where participants carried out TFA was banking (100%, 40/40), government app/website (77.5%, 31/40), social media (47.5%, 19/40) and email (35%, 14/40). 82.5% (33/40) of participants have come across a push-based, Just Tap to authenticate, TFA before.

6. Quantitative Analysis

In this section, the rationale for having 3 trials and the measurable aspects of user experience will be covered.

6.1. Rationale for 3 Trials:

One of our goals of this study is to find out if REPLICATE results in a more positive user experience than existing push based TFA solutions. Most participants are familiar with existing push based TFA solutions, but are unfamiliar with REPLICATE. Comparing results based on participants' first attempt with REPLICATE to results based on participants' many tries with existing solutions may thus lead to unfair results. Having multiple trials for REPLICATE will at least acquaint participants with REPLICATE and reduce unfairness.

In addition, given that the participants will most probably use TFA more than once and that many participants will not get a choice in deciding whether or not to adopt TFA, we believe that learnability is an important factor that should not be ignored. Even if participants' initial exposure to a method is less positive than another method, that does not necessarily mean that the subsequent trials will result in the same outcome. The participants' first encounter with the method is less significant as compared to the participants' future encounters - what we are aiming for is the best possible user experience in the long run. We need to have multiple trials in order to study this. With the

above two reasons, we decided on three trials - the minimum recommended number of trials to study learnability [22] - as we were concerned about participants' fatigue. To make our study as realistic as possible, participants did not attempt the same prototype successively. Participants completed all prototypes for one trial before moving on to the next trial. In the subsequent parts of this section, we chose to focus mainly on the results of trial 3, and only bring in results from trials 1 and 2 where meaningful comparisons can be made.

6.2. Metrics Used

We track 5 parameters: a) task success, b) task time, c) task efficiency, d) SUS ratings and e) combined analysis.

A. Task Success: We define task success as the participant managing to reach the "Unlocked" stage of our prototype in one try without any external help. A score of 1 was assigned if the participant completed the authentication successfully and 0 otherwise. The task success rate of each prototype was calculated as the percentage of the participants who successfully completed the trial 3 of each prototype. As our sample size is relatively small, we also determined the confidence intervals using the Adjusted Wald method. Our goal was to achieve a task success rate of at least 90%. The results can be seen in Table 1.

Authentication Method	Task Success Rate (in %)	95% Confidence Interval (Adjusted Wald Method)
Just Tap	100	(90,100)
Choose Number	100	(90,100)
Randomized Keypad	93	(76,99)
Key Drag	100	(90,100)
Shapes	100	(90,100)
Colored Button	97	(82,100)
Black Button	100	(90,100)
Draw Unlock Pattern	100	(89,100)

TABLE 1. TASK SUCCESS RATES WITH 95% CI (L2P)

All of our proposed solutions reached a task success rate of close to 100%. After applying the Adjusted Wald method and considering the 95% confidence interval (CI), Key Drag, Shapes, and Black Button will still have a task success of at least 90%. For methods with the same average task success rate, the variation in CI exists due to the different number of data taken into consideration. Erroneous data was removed. Considering the profile of our participants, the actual task success rate in the general population might be lower.

B. Task Time: We define task time as the time taken (in seconds) for the participant to authenticate successfully; task time = end time - start time. The start time is defined as the time the facilitator clicks on the login button, and the end time is the time the participant reaches task success. Since we recorded the whole process, we were able to obtain the task time by checking the recordings. We only recorded task times of successful attempts. For each prototype, the average task time was calculated. Our goal was to find an alternative solution that is competitive in time.

Ideally, to prove that one of our solutions is comparable to existing solutions, we should define acceptable criteria for the difference in task time and conduct a power analysis to determine the number of participants needed for a statistically significant result for that effect size. Then, we can conclude with a certain confidence level that we would have found the effect if it was there and hence we think it is not different. However, the number of participants needed is very high and unrealistic for this stage of work. Instead, we conducted the Shapiro-Wilk normality test and obtained a p-value of $0.524e-15 < 0.05$, hence we conclude that the data is not normal and we have to use a non-parametric test. Thus, Friedman's test and Wilcoxon Signed-Rank Test (WSRT) were used. We conducted Friedman's test and obtained a p-value of $6.82e-8 < 0.05$ which suggests that there are significant differences in time between the different authentication methods. However, Friedman's test does not tell us which groups are significantly different. As such, we conducted pairwise WSRT and adjusting of the p-values using Bonferroni-Holm correction. Black Button has a p-value of 1.00 when paired with existing solutions like Just Tap and Choose Number, but bigger studies are needed to confirm our hypothesis that they are comparable in time.

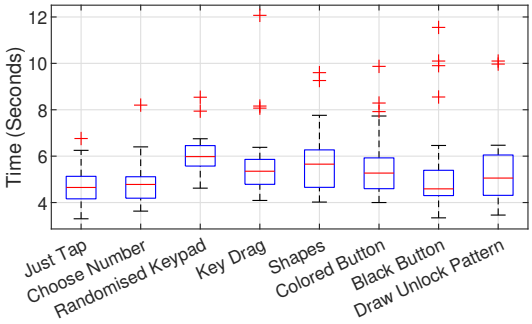

Figure 16. Time to Authenticate for Various Methods

We also plotted a graph of mean task time against the number of attempts (trial number) for each method to track learnability, as seen in Figure 17. Mean task time decreases across the trials and almost reaches a plateau by the trial 3. Taking the 95% CI into account, all of our proposed methods show some sign of plateauing from trial 2 to trial 3. Figure 18 shows how authentication time changes for Black Button across the 3 trials. The overlap in the CI of trial 2 and 3 suggests that users would have learned the method by the second trial.

Authentication Method	Q1	Mean	Median	Q3	95% Confidence Interval
Just Tap	4.19	4.69	4.65	5.12	(4.37,5.01)
Choose Number	4.24	4.87	4.78	5.08	(4.50, 5.24)
Randomized Keypad	5.60	6.02	5.98	6.44	(5.69, 6.35)
Key Drag	4.82	5.66	5.35	5.86	(5.07, 6.25)
Shapes	4.67	5.73	5.66	6.27	(5.25, 6.21)
Colored Button	4.60	5.61	5.27	5.89	(5.09,6.13)
Black Button	4.31	5.36	4.59	5.36	(4.61, 6.11)
Draw Unlock Pattern	4.35	5.41	5.06	5.99	(4.82, 6.00)

TABLE 2. STATISTICS OF AUTHENTICATION TIME (IN SECONDS)

C. Task Efficiency: We define task efficiency as the task success per unit time; task efficiency = task success / task time. The average task efficiency for each prototype is calculated, and a comparison is done to find out which prototype or existing solution is the most efficient. Figure 19 shows the respective task efficiencies with 95% confi-

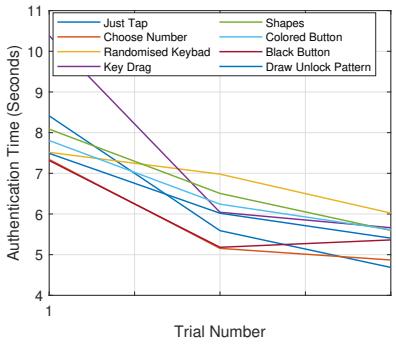

Figure 17. Task Time Over 3 Trials

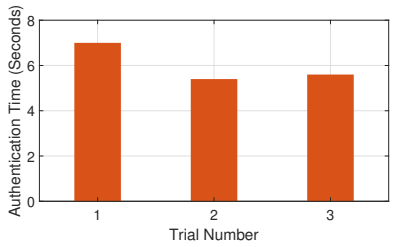

Figure 18. Method Learned by the Second Trial

dence intervals.) Black Button and Draw Unlock Pattern's task efficiency were comparable to existing solutions.

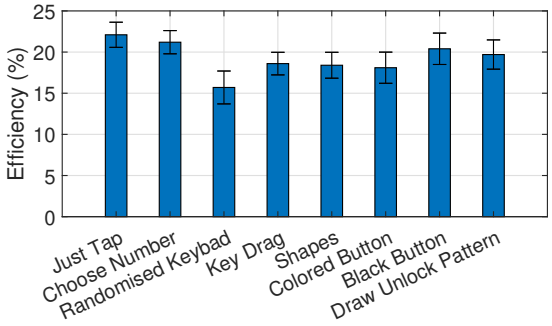

Figure 19. Efficiency of Prototypes

D. SUS Scores: The benefit of the SUS system is that due to the standardization of questions and score calculation, the scores can reliably be compared with other SUS scores that followed the same procedure. The procedure of achieving the SUS scores consists of administering a 10 question survey to users, getting a score from 0-4 for each question and multiplying the score out of 40 by 2.5, to achieve the percentile score out of 100. To increase the understanding of the survey for users, we modified the questions to give them better context, as shown in Table 6, Section 11.

As observed, the originally negative questions have also been made to be positive. The purpose of the original questionnaire alternating positive and negative questions was to reduce acquiescence bias. However, this change has been shown not to make an impact on the scoring and could even be better as "respondents are less likely to make mistakes when responding, researchers are less likely to make errors in coding, and the scores will be similar to the standard SUS. Furthermore, we contextualized the system to the "REPLICATE system", which has also been shown not to affect the scores. Ultimately, keeping the original SUS questions could allow us to compare with

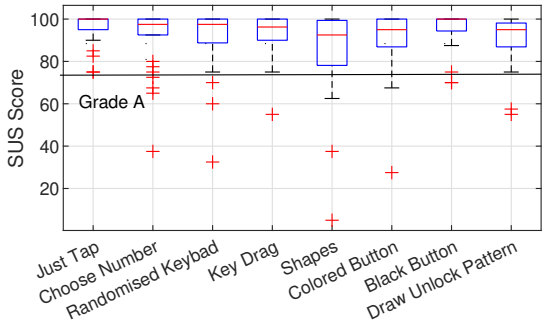

Figure 20. SUS Scores for Various Methods (L2P)

industry score benchmark with more confidence but also risks participant error. The SUS scores for all methods can be seen in Figure 20 and Table 3. All methods have attained a score in the A range. Notably, Tap-on-Black Button fared better than the existing solution Choose Number and was comparable to Just Tap TFA.

Authentication Method	Q1	Mean	Median	Q3	95% Confidence Interval
Just Tap	95.6	95.7	100.0	100.0	(93.0,98.4)
Choose Number	92.5	91.8	97.5	100.0	(86.7, 96.9)
Randomized Keypad	89.4	90.5	97.5	100.0	(84.5, 96.5)
Key Drag	90.0	91.8	96.3	100.0	(87.9, 95.7)
Shapes	78.8	84.7	93.0	98.8	(76.1, 93.3)
Colored Button	87.5	90.1	95.0	100.0	(84.5, 95.7)
Black Button	95.0	95.0	100.0	100.0	(91.6, 98.4)
Draw Unlock Pattern	87.5	90.3	95.0	97.5	(85.8, 94.8)

TABLE 3. SUS SCORE SUMMARY STATISTICS

E. Combined Analysis: Taking all the metrics used into account, a comparison of the authentication methods was made. Two additional items were being taken into consideration - favorite count and least favorite count. These were the frequency in which participants named the method as their favorite or least favorite during the Overall Post-test interview. We ranked the methods from 1 to 8 for each metric, and tabulated a score for each method by summing up the method's rank for each metric. The results can be seen in Table 4. Of all proposed methods, Black Button seems the most promising - being comparable to the existing solution, Just Tap, and better than the Choosing Number method.

Authentication Method	Task Success	Time	Task Efficiency	SUS Score	Favourite Count	Less Favourite Count	Score
Just Tap	1	2	1	1	1	4	10
Choose Number	1	3	2	3	4	1	14
Randomised Keypad	8	8	8	3	4	8	39
Key Drag	1	6	5	5	4	5	26
Shapes	1	7	6	8	7	7	36
Coloured Button	7	5	7	6	3	6	34
Black Button	1	1	3	1	2	2	10
Draw Pattern	1	4	4	6	7	2	24

TABLE 4. COMBINED ANALYSIS FOR EACH METHOD

6.3. Comparison of best REPLICATE with PIN based method of TFA

In order to study how REPLICATE stands in comparison to the PIN-based TFA, a further study involving 24 participants was done to compare our most promising solution, Black Button, with the existing TFA solution, PIN. The survey was repeated with the same participants as in the main study. In terms of time to authenticate

and efficiency, the PIN method fared worse than all other solutions, as seen in Figures 21(a) and 21(b). When participants were asked which method they preferred among the two, 19 out of 24 preferred Black Button, 3 out of 24 preferred PIN, and 2 had no preferences. Almost all participants who preferred Black Button preferred it for its speed and simplicity, while almost all participants who preferred PIN preferred it because they were habituated and felt that it is more secure. Also, the authentication time for the PIN method was not up to par with REPLICATE. We found that PIN-based TFA has p-value < 0.05 when paired with all other methods, suggesting that PIN takes a significantly longer time. The task success rate of PIN method approached 91% compared to the perfect score of REPLICATE. The tables comparing the PIN-based TFA with other methods of REPLICATE are in Section 10, Appendix. The SUS scores of PIN is fairly low, 83.8, compared to the black button method of TFA.

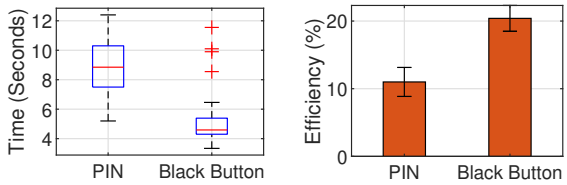

Figure 21. (a) Time to authenticate for Black Button (REPLICATE) is much lesser than that of PIN based TFA and (b) Efficiency of Black Button is higher than PIN method.

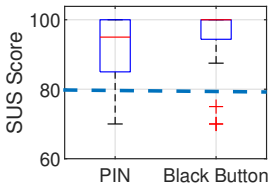

Figure 22. Comparison between SUS scores for PIN and Black Button suggests a clear preference of the REPLICATE's Black Button method over the prior.

Based on the previous studies such as [24], [25], PIN-TFA is more tedious as opposed to just a small amount of interaction in our designs.

7. Qualitative Analysis

Through the open-ended survey questions and the post-test interviews, we gain some insights that could help explain the quantitative results. These insights also allow us to understand the likes and dislikes of participants to further improve our proposed solutions. In this subsection, we will analyze the methods based on the common feedback given by participants.

7.1. Ease of Execution and Cognitive Effort

The ease of execution of a method, or how simple it is to carry it out, is the most common reason quoted by participants to explain why they like or dislike a particular method. It is very often mentioned together with the amount of cognitive effort required. For the proposed method with the highest SUS Score (Black Button), 3 out of 7 (43%) participants who chose the method as their favorite cited simplicity as one of the reasons why they liked it the most. In addition, 22 out of 30 participants (73%) cited simplicity when asked to state what they liked most about the method.

P11: *"This is the easiest to understand. By far the best for understanding and carrying out."*

Some participants went on to further explain which aspects of Black Button made it simple.

P14: *"It is easy enough for you to just press 1 time. The rest requires dragging and more actions. This is the fastest."*

P30: *"The buttons were all grey, and only the one I had to press was in black. The difference in the coloring was fun and helped me easily identify which button to press. Also, after I pressed the correct button, then the exact color in the instructions showed on my phone, which helped me to recognize that I did it correctly."*

In contrast, for the method with the lowest SUS Score (Shapes), 5 out of 8 participants (63%) chose it as their least favorite prototype because it was difficult to execute or required too much cognitive effort.

P21: *"Many things to consider, different shapes and directions."*

P30: *"I have to look out for 2 things - the shape and the direction, so it is not very intuitive."*

Participants clearly stated better usability of the Black Button over PIN-based TFA.

P29: *"I prefer the Black Button based method because it is so much easier and so much faster. The PIN-based method needs memorization of codes, and I will have to refer back and forth."*

P3: *"I prefer the Black Button method. For PIN-based method, I had to look at the screen at least 3 times to ensure I was doing it right. For Black button, I didn't have to refer back."*

7.2. Security

Security is another aspect that was commonly brought up by participants. While the existing solution, Just Tap, has the highest SUS score and the most number of participants (10 out of 39, 26%) citing it as their favorite method, it is the least secure. 13 out of 30 participants (43%) cited the lack of security as their least favorite thing about the method.

P5: *"I don't see what's the security function. If someone hacks your phone, then they can tap it."*

P34: *"Too easy, I don't feel secure at all."*

Participants found the Black Button method of TFA to be more secure.

P9: *"It is easy but secure. Just Tap is easy but not secure. For Black Button you only have 1/9 chance to randomly click on it. Easiest and more safe."*

P34: *"Simple but makes you think twice. Thinking twice is important because I want to be aware of what I am doing."*

Interestingly, although Shapes has the lowest SUS Score, 12 out of 30 participants (40%) gave positive comments regarding the security it provides.

P25: *"I like that it is fairly simple but relatively much more secure than current simple TFA methods."*

While some participants highlighted how they did not like Shapes because having to consider the shape and direction requires too much effort, other participants liked that there were two things to consider because it gave them a sense of security.

P7: *"It is a 2 step process. I have to locate which shape and direction. A bit harder but feels safer. A bit*

more troublesome, but that is why I like it because it is more secure."

7.3. Inclusivity

Some participants provided feedback by thinking from the perspective of others. For example, 15 out of 30 participants' (50%) least favorite thing about Colored Button was the fact that people with color blindness might not be able to use the method.

P12: *"In consideration of a minority of people that are colorblind, this method won't be feasible."*

Other participants thought from the perspective of the elderly. One participant commented about Randomized Keypad:

P20: *"I don't like this, very troublesome. This is especially bad for old people because they are not used to looking at numbers in another manner."*

Another participant commented about Draw Unlock Pattern:

P37: *"Troublesome, especially for old people. This is okay for unlocking phone because we are used to the same pattern, but in this case, it is always a different pattern."*

7.3.1. Level of Engagement:. The engagement level, or how fun a prototype is, was not the most significant consideration of most participants. As one participant puts it:

P4: *"In terms of user experience, I like Colored Button because it is more fun. In terms of security, Randomized Keypad and Draw Unlock Pattern feel quite secure. Security is more important than fun if I have to choose."*

However, participants prefer the more engaging or fun method if everything else is comparable. Key Drag, Shapes and Colored Button were deemed fun by some participants. One participant commented about Key Drag:

P13: *"Once in a while, it will be a very fun way. Looking at the key was more fun. The shapes look simple, but the key has a nicer display. It is a unique way to unlock because it literally has a key to unlock."*

Another participant commented about shapes:

P33: *"I like this. Usual OTP methods are very boring. This method is like a small puzzle. Nice shapes. More fun and less boring, I like this."*

One participant found Black Button method boring as compared to other methods.

P40: *"This is on the boring side after seeing the more colorful ones like Shapes."*

7.4. Inference from User Journey Map (UJM)

A UJM was generated for each prototype, using inputs from the emotions chart and comments/feedback from the *think-out-loud* process. Our UJM for trial 3 of all methods can be seen in Fig. 23, Section 11 Appendix. We decided to include the user persona section in our UJM since different types of users have different levels of experience with TFA and different levels of technical knowledge and hence may have very different experiences when trying our methods. The user's goals and expectations section in our UJM allows us to focus on the needs of the users, and breaking the journey into stages allows us to zoom in on the part of the journey that has the most potential for improvement. The user persona and verbatim comments also bring the experience to life and allow us to relate better to the users. In general, the participants' inputs for Hybrid Emotrak matched their comments for the prototypes. When participants gave a more positive comment for one prototype as compared to another, this was reflected as more positive emotions (higher scores) in the Hybrid Emotrak tool. Although we wanted to compare the average scores of the prototypes to see which prototype gave the users the most positive experience initially, we realized that the average would give roughly the same scores across the prototypes. This was perhaps due to the subjective nature of the method - for example, some participants might choose "satisfied" when they are neutral about the prototype, while others might choose "indifferent". Thus, it would be more meaningful to look at how individual participant's responses changed across different prototypes.

Furthermore, while we had expected that most participants would face a pain point during the Transitioning from Laptop to Phone Stage, most participants rated that stage equal to or better than the other stages. We believe that this might be due to the circumstances of our study, where the participant already had his or her phone in hand with the phone component of the prototype loaded. The transitioning stage was thus effortless as compared to other stages as participants simply had to glance down at their phones. This rationale was confirmed by a few participants. To avoid presenting an inaccurate UJM, we set the score for the transitioning from Laptop to Phone Stage at a default, 0.

7.5. Comparison of Replicate with other TFA

We compared REPLICATE with both Just Tap and PIN-based methods of TFA, Table 5, using the framework of Bonneau et al. [26]. The framework of Bonneau et al. considers 25 evaluation parameters, termed as "benefits", derived from the perspective of usability, deployability, and security that an authentication scheme should ideally provide.

Usability: While TFA is not the exact or the sole reason, each of the schemes requires users to carry a token device, a registered smartphone, to receive a push notification or generate OTP codes. None of the schemes are memory-wise effortless as they are part of the second-factor authentication ecosystem which requires username and password pair to begin with. Just Tap and REPLICATE are more efficient and have much lesser chance of errors compared to PIN-based method.

Deployability: While none of the methods is server compatible, all are browser compatible. The accessibility of Just Tap and REPLICATE is higher than the PIN-based TFAs. We would like to highlight that REPLICATE can be deployed on top of the Just Tap method after small updates in the users' application and changes to incorporate randomization at the server. We believe that REPLICATE will eventually become mature with time owing to its usability and gain in security.

Security: As highlighted in this research, REPLICATE offers better security than Just Tap to authenticate in the context of targeted attacks such as concurrent logins. Random guessing is not possible for both methods- Just Tap and REPLICATE. A special case of multiple attacks

Schemes	Usability								Deployability						Security											
	Memorywise-Effortless	Scalable-for-Users	Nothing-to-Carry	Physically Effortless	Easy-to-Learn	Efficient-to-Use	Infrequent-Errors	Easy-Recovery-from-Loss	Accessible	Negligible-Cost-per-User	Server-Compatible	Browser-Compatible	Mature	Non-Proprietary	Resilient-to-Physical-Observation	Resilient-to-Targeted-Impersonation	Resilient-to-Throttled-Guessing	Resilient-to-Unthrottled-Guessing	Resilient-to-Internal-Observation	Resilient-to-Leaks-from-Other-Verifiers	Resilient-to-Phishing	Resilient-to-Theft	No-Trusted-Third-Party	Requiring-Explicit-Consent	Unlinkable
PIN			+		*	+	+	+	+	*		*	*	*	+	+	*	*		*	*	*	*	*	
Just Tap			+		*	*		+	*	*		*	*	*	*		*	*		*	*	*	*	*	
REPLICATE			+		*	*		+	*	*		*		*	*	+	*	*		*	*	*	*	*	

TABLE 5. COMPARING REPLICATE AGAINST PIN AND TAP-TO-AUTHENTICATE USING THE FRAMEWORK OF BONNEAU ET AL. [26]. '*' REPRESENTS THAT THE SCHEME "OFFERS" THE BENEFIT AND '+' REPRESENTS THAT THE SCHEME "SOMEWHAT OFFER" THE BENEFIT.

and its countermeasures have been discussed in Section 8.1. Based on this analysis, we believe that the usability of REPLICATE is higher than the PIN method and close to that of Just Tap TFA, while its security is higher than Just Tap.

8. Related Works and Discussion

While there has been substantial work in the direction of proposing a usable and yet secure TFA, there has not been a dedicated study on exposing and solving vulnerabilities with the Just Tap TFA solution. [3] did a rigorous analysis of the usability of popular TFA solutions. All these studies prove that the Just Tap method of TFA is largely popular due to its high usability. This necessitates a study on possible vulnerabilities of the Just Tap method. Vulnerabilities associated with concurrent logins, by an adversary, have been raised by previous works such as [27] as well. While attempts to solve this problem have been made by the likes of Microsoft (*compare-to-confirm* methods [28]), studies have shown that users tend to avoid informed comparisons due to the earlier-mentioned habituation of tapping at the same place or repeating the same task multiple times [29].

■ **Assisting NBU applications:** In order to capture the Next Billion Users (NBU) markets [30]–[34], many companies and services have started to let users login into services by just typing in their mobile number, and the received OTPs. REPLICATE can be used in such cases wherein users can respond to screen overlays in place of SMS OTPs to login.

■ **Seamless and password-less Login:** REPLICATE can support password-less logins as well. For each login attempt post-registration, REPLICATE would send an interactive push to the token device in place of typical *Approve* and *Deny* UI which appear for Microsoft Authenticator [35] and Authy [2]. REPLICATE could be the forerunner in the passwordless login.

8.1. Limitations and Discussions

While the methods studied in REPLICATE surely provide high security (both actual security and as perceived by the users) when compared to Push-based TFAs, some users did raise concern about the simplicity of methods such as Black Button or Key Drag. Inherently, the Push-based TFAs are secure at the network layer [3], [36] and a push would only arrive to the registered token device in a similar fashion as a PIN would be generated only on the registered token device. While such users could be educated, a larger fraction of users who are habituated with Push-based TFAs, would find REPLICATE, Black Button, both secure and usable. Although we did not observe wrong clicks for authentication, the spacing between the black-buttons would have to be automatically adjusted as per the resolutions of the screen of the token device to rule out the possibility of collisions in approval by mistake. One of the limitations of our study is that user-testing was carried out online through Zoom to reduce the risks of contact during the COVID-19 pandemic. However, measures were carried out to understand the users' true feedback which have been outlined in the study design. We agree that a study with diverse users representing an average sample should be conducted for conclusive argument.

■ **The Case of Multiple Attacks:** We agree that launching several attempts will create multiple push notifications and requests for interactions at the user side. Hence, the probability of collision might increase. But this case can be resolved by selecting and randomizing between 2-3 best prototypes (based on our study) of REPLICATE and their variants. The study is a first attempt at showing that usable substitutes can reduce the vulnerability associated with the Just Tap based method of TFA. The studied substitutes can be fused together for better security. It is to be noted that the usability might be affected as users would be interacting with different methods in successive attempts. We understand that users would be habituated if permutations are from few variants of REPLICATE.

9. Conclusion

REPLICATE proposes a new set of UI and security interventions to resolve security issues, concurrency attacks, with existing push-notification based Just Tap method of TFA. The randomized interactions with notifications in the forms of screen overlays negate the attack. We design a remote usability testing suite to comment on the possible adoption of the proposed methods. In doing so, we propose new charts to evaluate user emotions and capture journeys. Rigorous analysis using prototypes suggests that REPLICATE is engaging, brings an element of fun, control and feeling of being in the loop of the authentication process.

10. Acknowledgments

We sincerely thank the anonymous shepherd and reviewers for their insightful comments and suggestions.

References

[1] "Inc. duo push," 2017, https://www.duosecurity.com/product/methods/duo-mobile.

[2] "Authy two factor authentication," https://authy.com/.

[3] K. Reese, T. Smith, J. Dutson, J. Armknecht, J. Cameron, and K. Seamons, "A usability study of five two-factor authentication methods," in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. Santa Clara, CA: USENIX Association, Aug. 2019. [Online]. Available: https://www.usenix.org/conference/soups2019/presentation/reese

[4] "Google 2-step verification," 2017, https://www.google.com/landing/2step/.

[5] B. B. Anderson, C. B. Kirwan, J. L. Jenkins, D. Eargle, S. Howard, and A. Vance, "How polymorphic warnings reduce habituation in the brain: Insights from an fmri study," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2883–2892.

[6] M. Jubur, P. Shrestha, N. Saxena, and J. Prakash, "Bypassing push-based second factor and passwordless authentication with human-indistinguishable notifications," in *Proceedings of the 16th ACM Symposium on Information, Computer and Communications Security*.

[7] I. Khokhlov and L. Reznik, "Data security evaluation for mobile android devices," in *2017 20th Conference of Open Innovations Association (FRUCT)*. IEEE, 2017, pp. 154–160.

[8] Figma. [Online]. Available: www.figma.com

[9] Zoom. [Online]. Available: https://zoom.us/

[10] M. R. Drew, B. Falcone, and W. L. Baccus, "What does the system usability scale (SUS) measure?" in *International Conference of Design, User Experience, and Usability*. Springer, 2018, pp. 356–366.

[11] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[12] K. Grindrod, H. Khan, U. Hengartner, S. Ong, A. G. Logan, D. Vogel, R. Gebotys, and J. Yang, "Evaluating authentication options for mobile health applications in younger and older adults," *PloS one*, vol. 13, no. 1, p. e0189048, 2018.

[13] T. Nguyen and N. Memon, "Tap-based user authentication for smartwatches," *Computers & Security*, vol. 78, pp. 174–186, 2018.

[14] T. S. Tullis and J. N. Stetson, "A comparison of questionnaires for assessing website usability," in *Usability professional association conference*, vol. 1. Minneapolis, USA, 2004, pp. 1–12.

[15] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?" in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 2215–2224.

[16] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '90. New York, NY, USA: Association for Computing Machinery, 1990, p. 249–256. [Online]. Available: https://doi.org/10.1145/97243.97281

[17] A. Toet, D. Kaneko, S. Ushiama, S. Hoving, I. de Kruijf, A.-M. Brouwer, V. Kallen, and J. B. Van Erp, "Emojigrid: A 2d pictorial scale for the assessment of food elicited emotions," *Frontiers in psychology*, vol. 9, p. 2396, 2018.

[18] S. E. Garcia and L. M. Hammond, "Capturing & measuring emotions in ux," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 777–785.

[19] K. Terada, A. Yamauchi, and A. Ito, "Artificial emotion expression for a robot by dynamic color change," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 314–321.

[20] D. De Angeli, R. M. Kelly, and E. O'Neill, "Beyond happy-or-not: Using emoji to capture visitors' emotional experience," *Curator: The Museum Journal*, vol. 63, no. 2, pp. 167–191, 2020.

[21] T. F. L. T. M. Inc., "Telegram," version 7.1.1. [Online]. Available: https://telegram.org/

[22] T. Tullis and B. Albert, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, 2nd ed., ser. Interactive Technologies. 340 Pine Street, Sixth Floor, San Francisco,CA, United States: Morgan Kaufmann Publishers Inc., 2013.

[23] T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *IEEE transactions on professional communication*, vol. 43, no. 3, pp. 261–278, 2000.

[24] N. Karapanos, C. Marforio, C. Soriente, and S. Capkun, "Sound-proof: Usable two-factor authentication based on ambient sound," in *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, 2015, pp. 483–498. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/karapanos

[25] B. Shrestha, M. Shirvanian, P. Shrestha, and N. Saxena, "The sounds of the phones: Dangers of zero-effort second factor login based on ambient audio," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 908–919. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978328

[26] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *2012 IEEE Symposium on Security and Privacy*, 2012, pp. 553–567.

[27] S. Jarecki and N. Saxena, "Authenticated key agreement with key re-use in the short authenticated strings model," in *International Conference on Security and Cryptography for Networks*. Springer, 2010, pp. 253–270.

[28] EWAND, "Enabling 2fa for msa, 2017."

[29] M. Shirvanian and N. Saxena, "On the security and usability of crypto phones," in *Proceedings of the 31st Annual Computer Security Applications Conference*, 2015, pp. 21–30.

[30] "How insights from user research help us build for the next billion." [Online]. Available: https://www.blog.google/technology/next-billion-users/how-insights-user-research-help-us-build-next-billion-users/

[31] J. O'Neill, K. Toyama, J. Chen, B. Tate, and A. Siddique, "The increasing sophistication of mobile media sharing in lower-middle-class bangalore," in *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, 2016, pp. 1–11.

[32] "Who are the next billion users and what do they want/." [Online]. Available: https://techcrunch.com/2019/03/08/who-are-the-next-billion-users-and-what-do-they-want/

[33] "The next billion users are the future of the internet." [Online]. Available: https://www.blog.google/technology/next-billion-users/next-billion-users-are-future-internet/

[34] T. N. Smyth, S. Kumar, I. Medhi, and K. Toyama, "Where there's a will there's a way: Mobile media sharing in urban india," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 753–762. [Online]. Available: https://doi.org/10.1145/1753326.1753436

[35] MICROSOFT, "Enable passwordless sign-in with the microsoft authenticator app." [Online]. Available: https://docs.microsoft.com/en-us/azure/active-directory/authentication/howto-authentication-passwordless-phone

[36] J. Reynolds, N. Samarin, J. Barnes, T. Judd, J. Mason, M. Bailey, and S. Egelman, "Empirical measurement of systemic 2fa usability," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 127–143.

11. Appendix

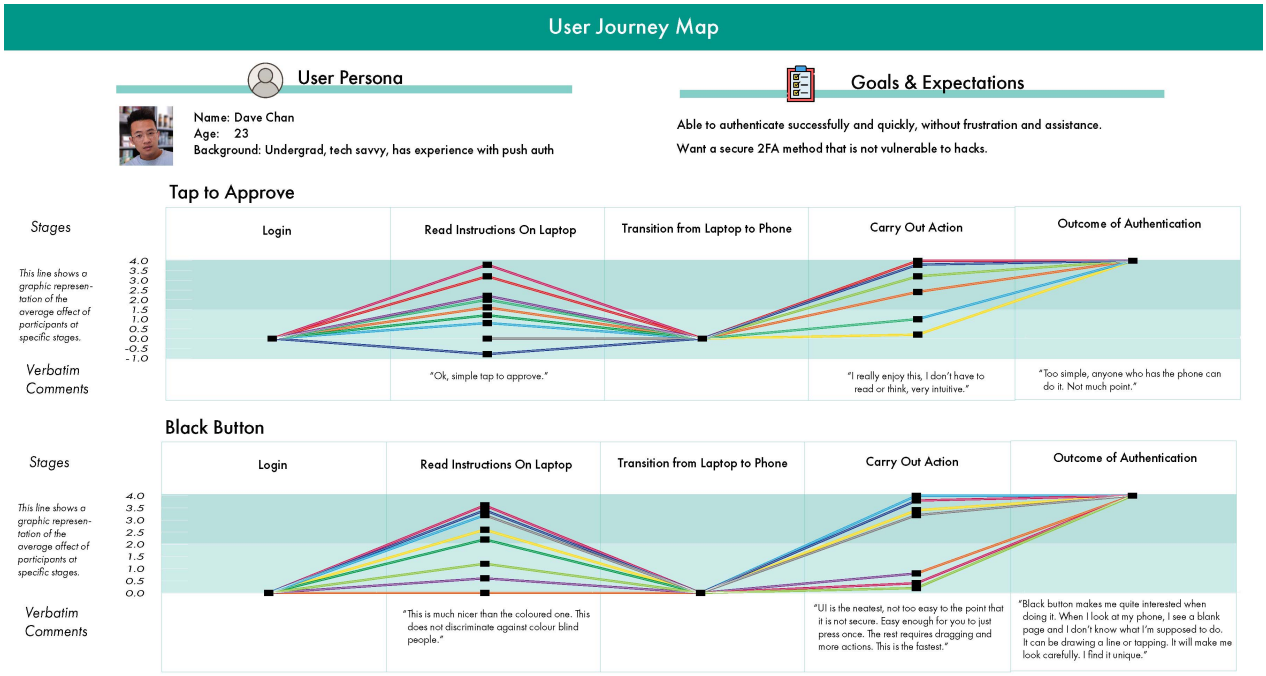

Figure 23. UJM for Just Tap and Black Button

Original	**Contextualized**
I think that I would like to use this system frequently.	I think that I would like to use REPLICATE frequently.
I found the system unnecessarily complex.	I found REPLICATE to be simple.
I thought the system was easy to use.	I thought REPLICATE was easy to use.
I think that I would need the support of a technical person to be able to use this system.	I think that I could use REPLICATE without the support of a technical person.
I found the various functions in this system were well integrated.	I found the various functions in REPLICATE were well integrated.
I thought there was too much inconsistency in this system.	I thought there was a lot of consistency in REPLICATE.
I would imagine that most people would learn to use this system very quickly.	I would imagine that most people would learn to use REPLICATE very quickly.
I found the system very cumbersome to use.	I found REPLICATE very intuitive.
I felt very confident using the system.	I felt very confident using REPLICATE.
I needed to learn a lot of things before I could get going with this system.	I could use REPLICATE without having to learn anything new.

TABLE 6. ORIGINAL AND CONTEXTUALIZED QUESTIONS