JÖNKÖPING UNIVERSITY
School of Engineering

Phishing Susceptibility and Mitigation in the 2FA Context

An Investigation of How the Interplay of Psychological and Individual Factors and UX Design Can Influence Users' Decisions to Login to a Suspicious Website

Main Subject Area: Informatics
Specialisation in: User Experience Design
Author(s): Jorina Freya Gerken
Zhaoying Wang
JÖNKÖPING 2024, June

Certificate of Completion

This final thesis has been carried out at the School of Engineering at Jönköping University within Informatics. The authors are responsible for the presented opinions, conclusions, and results.

Examiner:	Bruce Ferwerda
Supervisor:	Agnieszka Kitkowska
Scope:	15 hp (second-cycle education)
Date:	2024-06-10

Attestation of Authorship

We hereby declare that this submission is our work, based upon research that we have conducted. To the best of our knowledge and belief, it contains no material published or written by another person – except where explicitly defined in the Acknowledgements or listed in the References and properly cited. Nor does it contain any material of ours that, to a substantial extent, has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Jorina Freya Gerken
Zhaoying Wang

Acknowledgements

We want to thank our supervisor, Agnieszka Kitkowska, for her continued support and encouragement; our examiner, Bruce Ferwerda, for his constructive criticism; Kristofer Månsson for his statistics advice; Andrej Lohn and our pilot testers for their input on our study setup; and all of those who took part in the study for their participation. Thank you.

Abstract

Phishing is a form of social engineering, in which attackers attempt to trick victims with e-mails designed to look like legitimate requests (Vishwanath et al., 2011), aiming “to exploit human error or human behaviour with the objective of gaining access to information or services” (European Union Agency for Cybersecurity [ENISA], 2023, p. 7). According to the ENISA Threat Landscape 2023 report, phishing is the most prevalent form of social engineering and predicted to continue posing a significant threat to users (ENISA, 2023). In this, attackers have also already succeeded in circumventing second-factor authentication (2FA) (ENISA, 2023).

This thesis aimed to contribute to the ongoing research concerning the mitigation of social engineering attacks by investigating phishing susceptibility and a UX-based mitigation approach in the context of 2FA, which to the best of our knowledge had not been previously researched.

Based on prior research in other contexts, stress, attention, elaboration, involvement and 2FA frequency were identified as potentially relevant factors. Under consideration of these factors, a 2FA implementation was designed, combining automated URL verification with verification-based warnings. An online study ($N = 94$) was conducted to investigate how the posited susceptibility factors as well as the UX design can influence users' decisions to abort or proceed with logging in to a suspicious website. In this, a between-subject study design was used to investigate how much of an impact specifically the “opinionatedness” of a warning design, i.e. its “use of visual design cues to promote a recommended course of action” (Felt et al., 2015, p. 2893), can have in the 2FA context compared to an otherwise identical design offering a neutral choice.

In the collected sample, involvement had a significant negative effect on the likelihood to proceed with the login, in accordance with its posited influence. In addition, confidence in the decision made was discovered as another potential predictor, also showing a significant negative effect on the likelihood to proceed in the collected sample. The observed effect of the opinionated design can be seen to contradict the posited assumption that users would be more likely to go with the promoted action. However, overall, the results can be seen to suggest that taking susceptibility factors into consideration when designing 2FA implementations might be a promising approach towards phishing mitigation. Further research is needed to validate these indications, due to the insufficient sample size and use of convenience sampling in this thesis.

Keywords— phishing susceptibility, phishing mitigation, 2FA, second-factor authentication, psychological factors, involvement, confidence in the decision made, warning design, opinionated design

Contents

Certificate of Completion . . . i
Attestation of Authorship . . . ii
Acknowledgements . . . ii
Abstract . . . iii
List of Figures . . . vi
List of Tables . . . vii

1 Introduction **1**
1.1 Problem Statement . . . 1
1.2 Purpose and Research Questions . . . 3
1.3 Scope and Delimitations . . . 3
1.4 Outline . . . 4

2 Theoretical Framework **5**
2.1 Phishing Susceptibility . . . 5
2.1.1 Stress . . . 5
2.1.2 Information Processing . . . 5
2.1.3 Integrated Information Processing Model of Phishing Susceptibility (IIPM) . . . 6
2.1.4 Other Factors . . . 7
2.1.5 Resulting Model of Phishing Susceptibility at the 2FA stage . . . 7
2.2 Phishing Mitigation . . . 8
2.2.1 Automating Detection . . . 8
2.2.2 Supporting User Decision-Making . . . 8
2.2.3 Resulting Approach Investigated in This Thesis . . . 11

3 Method and Implementation **13**
3.1 Method Choice . . . 13
3.2 Artefact Creation . . . 13
3.2.1 Phishing E-Mail . . . 14
3.2.2 Login Page . . . 14
3.2.3 2FA App . . . 15
3.3 Data Collection . . . 15
3.3.1 Participants . . . 15
3.3.2 Measurements . . . 16
3.3.3 Stress Inducement . . . 19
3.3.4 Study Design . . . 20
3.3.5 Pilot Study . . . 21
3.4 Data Analysis . . . 21
3.5 Validity and Reliability . . . 21
3.6 Considerations . . . 21

4 Results **24**
4.1 Descriptive Statistics . . . 24
4.2 Scales . . . 25
4.3 Correlations . . . 30
4.4 Effect of the Stress Inducement . . . 30
4.5 Logistic Regression . . . 30
4.5.1 Introduction of Dummy Variables . . . 31
4.5.2 Testing Prerequisites for Logistic Regression . . . 31
4.5.3 Univariate Significance . . . 32

4.5.4 Results Logistic Regression . . . 32
4.6 Reasons Stated for Decision Made . . . 33

5 Discussion **35**
5.1 Results Discussion . . . 35
5.1.1 Research Question 1: Psychological Factors . . . 35
5.1.2 Research Question 2: 2FA Frequency . . . 38
5.1.3 Research Question 3: Opinionated Design . . . 39
5.2 Method Discussion . . . 40

6 Conclusions and Further Research **42**
6.1 Conclusions . . . 42
6.1.1 Practical Implications . . . 42
6.1.2 Scientific Implications . . . 43
6.2 Further Research . . . 43

7 References **45**

8 Appendices **49**
8.1 Appendix A: Full Questionnaire . . . 49
8.1.1 Initial Stress . . . 49
8.1.2 Open Question After the Interactive Task . . . 49
8.1.3 Confidence in the Decision Made . . . 49
8.1.4 Stress During the Interactive Task . . . 49
8.1.5 Attention . . . 50
8.1.6 Elaboration . . . 50
8.1.7 2FA Frequency . . . 50
8.1.8 Involvement . . . 51
8.1.9 Gender . . . 51
8.1.10 Age Group . . . 51

List of Figures

1.1 Information Flow in an Adversary in the Middle Attack 2

3.1 Phishing E-Mail 14
3.2 Initial Login Page 15
3.3 Neutral choice warning design with “confirm” on the left and “abort” on the right 16
3.4 Neutral choice warning design with “abort” on the left and “confirm” on the right 16
3.5 Opinionated warning design prompting users to abort 16
3.6 Opinionated warning design prompting users to proceed 16
3.7 Study Design 20

List of Tables

3.1 Measurement of Variables 17

4.1 Age Group Frequencies 24
4.2 Design Group Frequencies 25
4.3 2FA Frequencies 25
4.4 Component Matrix for Initial Stress 26
4.5 Component Matrix for Stress During the Interactive Task 26
4.6 Component Matrix for Attention 27
4.7 Rotated Component Matrix for the Complete Involvement Scale 27
4.8 Component Matrix for Conceptually Reduced Involvement Scale 28
4.9 Rotated Component Matrix for Elaboration 29
4.10 Pattern Matrix for Complete Scale Confidence in the Decision Made 29
4.11 Factor Matrix for Reduced Scale Confidence in the Decision Made 29
4.12 Correlations Between Variables 30
4.13 Dummy Variables for the Warning Designs 31
4.14 Dummy Variables for 2FA Frequency 31
4.15 Univariate Analysis of Independent Variables 32
4.16 Classification Table for Logistic Regression Model 33
4.17 Variables in the Logistic Regression Equation 33
4.18 Cross Tabulation of the Warning Designs and the Likelihood to Proceed 33

1 Introduction

1.1 Problem Statement

Social engineering is a cyberattack which "attempt[s] to exploit human error or human behaviour with the objective of gaining access to information or services" (ENISA, 2023, p. 7). Although it utilizes technology, it "rel[ies] on a human element to be successful" (ENISA, 2023, p. 7). To this end, it commonly employs psychological persuasion (Washo, 2021, p. 1).

One common form of social engineering is phishing, in which attackers attempt to trick victims with e-mails designed to look like legitimate requests (Vishwanath et al., 2011, p. 576). Aiming to achieve this, attackers commonly impersonate legitimate businesses (such as known shops or banks) or institutions (such as tax agencies), or relate their requests to current events (Vishwanath et al., 2011, p. 576) (such as the war in Ukraine (ENISA, 2023, p. 72)). In order to persuade their victims, they furthermore use "statements invoking fear, threat, excitement, or urgency" (Wang et al. 2009, as cited in Vishwanath et al., 2011, p. 576).

According to the ENISA Threat Landscape 2023 report, "[s]ocial engineering attacks grew significantly in 2023" (ENISA, 2023, p. 4) and "persist as a substantial threat to users" (ENISA, 2023, p. 71). In this, phishing is the most prevalent approach (ENISA, 2023, p. 4) and is expected to remain so "over the long term, as it is cheap, easy to conduct and profitable" (ENISA, 2023, p. 73) for the attackers. This is further compounded by the advent of new technologies such as AI (ENISA, 2023, p. 74), facilitating "realistic and targeted [...] attacks" (ENISA, 2023, p. 31), and Phishing as a Service (ENISA, 2023, p. 71), enabling even attackers with little expertise of their own (ENISA, 2023, p. 34).

Overall, the continuing prevalence shows that current approaches in mitigating phishing attacks have so far not been sufficiently successful. Even with awareness campaigns or trainings, people still remain susceptible (Williams et al., 2018, p. 1; Vishwanath et al., 2018, p. 1147), and attackers have proven to quickly adapt to new technological defences (ENISA, 2023, p. 33). In particular, they already succeed in circumventing existing implementations of second factor authentication (2FA) (ENISA, 2023, p. 76), a technology introduced to make systems more secure by requiring an auxiliary factor (Siadati et al., 2016, p. 65).

As the name suggests, second factor authentication requires users to provide two factors rather than just one in order to prove their identity. Each such factor can be either something the user knows (such as a password), has (such as a device), or "is" (i.e. biometric data). Most commonly a password is combined with a second factor token provided by a 2FA device (such as a card reader), SMS service or app. The resulting 2FA token is texted as a one-time code to the user or generated by the device or app and is different for each authentication attempt. Since obtaining the second factor thus requires the possession of the user's device, stealing their password is no longer sufficient to gain access to their account.

One 2FA circumvention technique - which is also included in advanced phishing kits offered on the Dark Web (ENISA, 2023, p. 78) - is the adversary-in-the-middle technique, "where the attacker positions himself between the victim and the legitimate service" (ENISA, 2023, p. 76) using a so-called reverse proxy "to intercept credentials, two-factor authentication (2FA) codes, and session cookies to hijack accounts of interest" (ENISA, 2023, p. 78), as illustrated in Figure 1.1.

In this, the user (A) is tricked into believing that they are interacting with the legitimate service (C), when in fact they are interacting with a fake version (B) set up by the attacker. When the user provides their credentials to the fake service (1), these are intercepted by the attacker. The attacker then starts a session with the legitimate service (C), using the intercepted credentials (2). This causes the legitimate service (C) to trigger a second factor authentication with the 2FA service (D) for the attacker's session (3). Depending on the type of the implementation, the 2FA service then either requests a confirmation from the user (4a) or sends them a 2FA code (4b). In the first case, the user then confirms the 2FA request (5a) causing the 2FA service to send a confirmation to the legitimate service (6a) and in turn the legitimate service to accept the authentication of the attacker (7). In the second case, the user provides the 2FA code to the fake service (5b), which the attacker intercepts and then provides to the legitimate service in their own session (6b) again

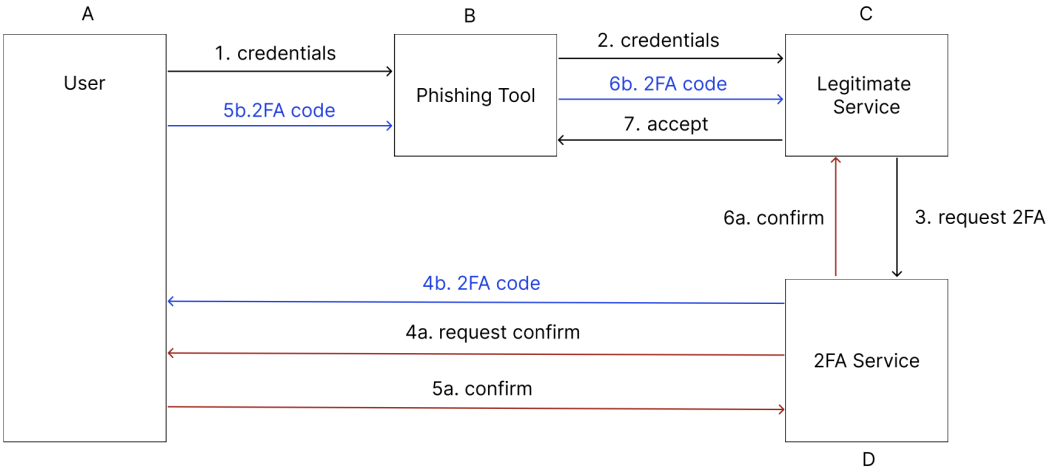

Figure 1.1: Information Flow in an Adversary in the Middle Attack

leading to the legitimate service accepting the authentication of the attacker (7). As a result, in both cases the attacker now has successfully gained access to the legitimate service.

Not only can 2FA implementations thus be circumvented, a study by Tolbert et al. (2022) can be seen to suggest that the very presence of 2FA on a website may even increase the likelihood that it is perceived as legitimate, consequently increasing users' susceptibility to 2FA-capable phishing websites (Tolbert et al., 2022, p. 197).

If a user is tricked into visiting such a website, the only contact they still have to a legitimate provider is the 2FA service, making it the last potential point for intervention. Furthermore, when users reach the 2FA stage, this arguably means they have already failed to notice any suspicious cues at the previous stages (e-mail and website), indicating that they need support in detecting the deception. The 2FA service, however, only interacts with the legitimate user and the legitimate website (steps 3-6 in Figure 1.1) and has no information about the website the user is actually interacting with ((B) in Figure 1.1). Any attempt at automatic detection therefore requires the user to provide this information, which can then be analysed and verified by the 2FA service. In order to be both usable and useful, such a solution arguably needs to balance what kind and quality of input the verification algorithm needs in order to produce sufficiently meaningful results with the resulting effort and expertise required from the user in order to provide said input. Sun et al. (2021) successfully conducted a technical proof of concept for a photo-based URL verification, requiring the user only to take a picture of their browser. They state that false positives are rare (Sun et al., 2021, p. 4), but these can arguably still impair legit use cases, which might be of high importance and urgency to users in real life. Unlike Sun et al. (2021), who propose to automatically reject any authentication attempt if the URL verification fails, we therefore argue that the automation should be used to facilitate, rather than replace user decision-making. This, of course, reintroduces the human element social engineering seeks to exploit and thus necessitates supporting the user in this decision-making.

In conclusion, attackers have proven to quickly adapt to new technological defences. Furthermore, although social engineering utilizes technology, it "rel[ies] on a human element to be successful" (ENISA, 2023, p. 7). Therefore, mitigation strategies arguably have to take into account both technological as well as psychological aspects, which is the approach that was applied in this thesis.

1.2 Purpose and Research Questions

The purpose of this thesis was to contribute to the research concerning the mitigation of social engineering attacks by investigating if and how user experience design in a 2FA app can be employed, aiming to decrease the likelihood of users logging in to a phishing website. Specifically, it was investigated how a design that combines automated URL verification (using a photo- and text-recognition-based solution as prototyped by Sun et al. (2021)) with verification-based warnings can influence user decision-making at the stage of 2FA login. The aim of the investigated design was to reduce cognitive effort for the user by automating URL verification; increase the perceived relevance of the resulting warnings by basing them on actual suspicion in the specific case rather than simply a general cautionary message; and increase the effectiveness of the warnings by applying findings from prior research as well as established warning design guidelines. Regarding the warning design, it was investigated how much of an impact specifically the "opinionatedness" of a warning design, i.e. its "use of visual design cues to promote a recommended course of action" (Felt et al., 2015, p. 2893), can have in the 2FA context compared to an otherwise identical design offering a neutral choice. In addition, it was investigated how psychological and individual factors can influence user decision-making in this context. Based on prior research, as detailed in the next chapter, stress, attention, elaboration, involvement and 2FA frequency were identified as potentially relevant factors.

Overall, this resulted in the following research questions:

In the context of automated URL verification and verification-based warnings in a 2FA app,

RQ1: how can psychological factors, such as stress, attention, elaboration and involvement, influence the user's decision to abort or proceed with logging in to a suspicious website?

RQ2: how can 2FA frequency influence the user's decision to abort or proceed with logging in to a suspicious website?

RQ3: how can opinionated design influence the user's decision to abort or proceed with logging in to a suspicious website?

To the best of our knowledge, this combined approach of automation and warning design at the stage of 2FA with consideration of psychological and individual factors had not previously been investigated.

1.3 Scope and Delimitations

This study aimed to contribute to the research into the mitigation of phishing attacks utilizing an adversary-in-the-middle attack, in which attackers lure their victims to a fake website in order to trick them into unknowingly providing their credentials. The specific subscenario addressed was the same as in the study by Sun et al. (2021), namely one in which the user accesses the website on a computer or tablet, and uses their phone for the 2FA. This, on the one hand, makes it possible to take a photo with the phone of the website's URL (namely the browser's address bar displayed on the other device); and, on the other hand, prevents other (and — as Sun et al. (2021) argue — better) technical solutions for the verification, which would be available if only one device was used, such as using session cookies (Sun et al., 2021, p. 12).

In their study, Sun et al. (2021) had already successfully tested the general technical feasibility of the photo-based URL verification. In this thesis, on the other hand, the focus of the research was on factors influencing user decision-making, and therefore on the effect of a *perceived* automation rather than on a working implementation of the verification. For this reason, a prototype was used that simply failed all verification attempts. This also reduced the risk that participants would get stuck at this stage, because the content and quality of the photo they took did not matter.

An overtly fictional scenario was used for this study, meaning that participants were aware the suggested threat was not real and their decisions would not have real-life consequences. This might to some degree have affected participants' responses and behaviour, and reduced the ecological validity of the study. However, designing the study in a way that would attempt to trick victims into believing the scenario was real, would not only have been challenging to implement due to the limited time and lack of budget available for this thesis, it would also have been ethically questionable. In particular, victims in phishing scenarios can

experience stress levels so severe that we considered attempting to expose study participants to them as unethical. Therefore, instead an overtly fictional scenario was combined with a stress inducement based mainly on time and performance pressure which could be expected to result in mild to moderate “everyday” stress during the task (if any) but was not likely to result in severe or lasting distress (see section 3.3.3).

1.4 Outline

In the following, first the theoretical framework and resulting mitigation approach investigated in this thesis are introduced. Then the chosen methodology is explored, i.e. how the study was designed, conducted and its results evaluated. These results are then presented and discussed, and finally conclusions as well as implications for future research are drawn.

2 Theoretical Framework

The ENISA Threat Landscape 2023 report predicts that phishing will continue to pose a significant threat to users (ENISA, 2023), indicating that existing mitigation approaches are not sufficiently effective. In particular, attackers have already succeeded in circumventing existing 2FA implementations (ENISA, 2023, p. 76).

While phishing utilizes technology, it relies on exploiting human error and behaviour in order to be successful (ENISA, 2023). Therefore, mitigation approaches arguably need to take into account both technological and psychological aspects. Norman (2013) argues that when designing products, designers have to understand and consider how users really are, and in particular warns of the risk of assuming them to behave too logically (Norman, 2013, p. 7). In order to address the phishing threat, it is thus arguably necessary to understand both what can make users susceptible as well as what approaches could mitigate this susceptibility. This thesis aimed to contribute to this specifically in the context of 2FA, which to the best of our knowledge had not been researched before.

In this chapter, therefore, first theories and findings from existing research on psychology as well as phishing susceptibility in other contexts are presented and potential factors influencing susceptibility in the 2FA context identified; then, under consideration of these factors, mitigation strategies and the mitigation potential of UX design are explored, based on which the specific UX-based mitigation approach investigated in this thesis was developed.

2.1 Phishing Susceptibility

In this section we explore psychological and individual factors that have been indicated by prior research as relevant to phishing susceptibility, and introduce the resulting theoretical model of phishing susceptibility at the 2FA stage.

2.1.1 Stress

Stress induced in the targets plays an important role in phishing. Attackers commonly use techniques such as “statements invoking fear, threat, excitement, or urgency” (Wang et al. 2009, as cited in Vishwanath et al., 2011, p. 576). “[T]hreats to [...] personal resources such as money” (Swindler, 2021, p. 136) or feelings such as anxiety can trigger acute stress responses (Swindler, 2021, p. 136), which in turn change the way the brain works (Swindler, 2021, p. 17). Higher-order thinking, such as reading (Swindler, 2021, p. 23), analysis and logic, degrade quickly in stress-driven situations (Swindler, 2021, p. 301), reducing people’s ability to think rationally, properly process information or consider potential implications (Johansson and Ask, 2023). The prioritization of “fast, intuitive decision-making over logic and reason” (Swindler, 2021, p. 25), can make people less rational and more error prone (Swindler, 2021, p. 17). Furthermore, in the acute stress phase, focus on addressing the (perceived) threat can intensify to the degree of tunnel vision (Swindler, 2021, p. 39). This is arguably of particular relevance in the phishing context since, for most users, security is already, under normal circumstances, a secondary concern and they are likely to concentrate more “on the real purpose of their online interaction” than on potential cues for deception (Alsharnouby et al., 2015, p. 69) even without stress. Overall, acute stress can thus increase phishing susceptibility.

2.1.2 Information Processing

According to Eveland et al. (2003), information processing is a two-stage process of which attention is the first and elaboration the second stage (Eveland et al. 2003, as cited in Vishwanath et al., 2011, p. 578). In this, attention indicates how much “mental focus [is] given to specific elements of an event or object” (Eveland et al. 2003, as cited in Vishwanath et al., 2011, p. 578), but does not involve any evaluation of those elements (Vishwanath et al., 2011, pp. 578–9). While attention is thus necessary to detect deception, it is not sufficient on its own (Vishwanath et al., 2011, p. 579). Rather, the observed cues need to also be elaborated

on (Vishwanath et al., 2011, p. 579). Elaboration is the second stage of information processing, in which connections are made between the cues and prior knowledge, thus evaluating them (Vishwanath et al., 2011, p. 579). Impaired information processing has been found to increase susceptibility to social engineering attacks (Johansson and Ask, 2023). Furthermore, high stress levels — as are commonly experienced by the targets — negatively affect both attention (tunnel vision) as well as the higher-order thinking necessary for proper elaboration.

2.1.3 Integrated Information Processing Model of Phishing Susceptibility (IIPM)

The integrated information processing model of phishing susceptibility (IIPM) by Vishwanath et al. (2011) was developed to explain individual differences in susceptibility to phishing e-mails, considering aspects of personality based beliefs, cognition and information processing (Vishwanath et al., 2011, p. 578). The model was constructed building on previous research, theories and models in this area (including the Interpersonal Deceptions Theory by Buller et al. 1996, the Theory of Deception by Johnson et al. 1992, and the Elaboration Likelihood Model by Petty et al. 1986 (Vishwanath et al., 2011, p. 578)), and adapting them to the specific context of phishing e-mails.

The IIPM was of interest for this thesis because the context it was developed for is very closely related to the context of our research questions. In the scenario of a phishing attack using a 2FA-capable phishing website, the e-mail context of the IIPM is the first stage, visiting the phishing website the second, and the 2FA context investigated in this thesis the third. Furthermore, these stages take place immediately after one another, i.e. they are part of the same overarching situation. Therefore, it is plausible to assume that external as well as contextual and situational factors remain mainly stable during this time, except for where they are affected by the stages themselves. Consequently, factors that were found to be significant at the e-mail stage can be seen as a good starting point for a model for the 2FA stage.

In their study of 2011, Vishwanath et al. (2011) found "strong support for the model's theoretical structure and causative sequence" (Vishwanath et al., 2011, p. 576) and stated that the IIPM overall "explains close to 50% of the variance in individual phishing susceptibility" (Vishwanath et al., 2011, p. 576). However, in this, they were only able to find support for the significance of some of the factors of the model. According to Vishwanath et al. (2011)'s results, of the model's proposed factors attention, involvement and e-mail load were found to be significant, while knowledge, self-efficacy and elaboration were not (Vishwanath et al., 2011, p. 583). We decided to adopt the former ones based on their likely importance while dropping the rest, aiming not to inflate our model unnecessarily and consequently decrease the feasibility of our study (required sample size). However, we made one exception to this rule by retaining elaboration, which we had already identified as a potentially significant factor based on other research (see section 2.1.2). This also allowed us to validate Vishwanath et al. (2011)'s findings regarding elaboration, which indicated that the "[d]irection [was] as hypothesized" (namely that a higher degree of elaboration reduces susceptibility) "but the relationship was not significant"(Vishwanath et al. (2011, p. 583)), in a slightly different context.

In the following, we outline the IIPM's dependent variable — *Likelihood to Respond* — and the factors found to be its significant predictors, as well as what adaptations we made to apply all of these to the 2FA context.

Likelihood to Respond - *adaptation:* **Likelihood to Proceed**

The IIPM measured susceptibility to phishing as the likelihood to respond to a phishing e-mail, i.e. whether for a given e-mail the user decides to respond or not to respond (Vishwanath et al., 2011, p. 578). This is arguably equivalent to the user's decision to abort or proceed with the login at the 2FA stage. In this, the user is successfully phished in the e-mail case, if they choose to respond, and in the 2FA case, if they choose to proceed. Therefore, the equivalent of this measure in the 2FA context is *Likelihood to Proceed*.

Attention - *adaptation:* **Attention** (without explicit element-specific subfactors)

In the IIPM, Vishwanath et al. (2011) distinguished between four specific elements users may pay attention to, arguing that the direction of their expected influence (negative or positive) on the likelihood to respond differs (Vishwanath et al., 2011, p. 579). Specifically, attention to e-mail title or urgency cues were expected

to increase the likelihood to respond, since they are deliberately designed by the attackers to persuade their targets; attention to source information or spelling and grammar, on the other hand, were expected to decrease the likelihood to respond, because they can compromise the perceived authenticity of the e-mail (Vishwanath et al., 2011, p. 579).

By contrast, all the elements in the 2FA app originate from and are under the control of legitimate sources (provider of the legitimate website, 2FA service). The attackers have no influence on any of them and therefore cannot place any cues working to their advantage. Hence, unlike in the e-mail case, all elements in the 2FA app are designed to support the user and mitigate phishing. Therefore, they should arguably all have an effect (if any) in the same direction, namely to decrease phishing susceptibility. For this reason, we decided not to distinguish between different sub-dimensions of attention, and instead investigate it as a unidimensional construct.

Involvement - *no adaptation*

Involvement indicates how relevant an observer perceives a message or event to be to them (Zaichkowsky 1985, as cited in Vishwanath et al., 2011, p. 580). The higher the perceived relevance, the more likely they are to consciously evaluate it (Vishwanath et al., 2011, p. 580), i.e. pay attention to and elaborate on it. This also means that automatic, subconscious responses, e.g. due to habituation, are less likely to occur (Vishwanath et al., 2011, p. 580). This arguably applies equally to the e-mail and the 2FA stage. Therefore, no adaptations were necessary for this factor.

Email Load - *adaptation:* **2FA Frequency**

According to Vishwanath et al. (2011), a high e-mail load might reduce the amount of time and attention the recipient is able to give to each individual e-mail (Vishwanath et al., 2011, p. 580). Furthermore, a combination of habitual patterns and high e-mail load "tend[s] to trigger automatic responses" (Vishwanath et al., 2011, p. 584) "without consciously paying attention" (Vishwanath et al., 2011, p. 580). Frequent 2FA use can arguably also reduce the time and attention the user is able to give to each authentication process as well as facilitate habituation. Therefore, the equivalent of this measure in the adapted model is *2FA Frequency*.

2.1.4 Other Factors

Two factors which have been frequently investigated for their potential influence on social engineering susceptibility are gender and age. In both cases, however, results have been inconsistent.

According to Sheng et al. (2010), "results suggest that women are more susceptible than men to phishing and participants between the ages of 18 and 25 are more susceptible to phishing than other age groups" (Sheng et al., 2010, p. 373). However, in the detailed discussion of their results, they concede that the observed gender-based difference could partially be explained by the fact that the female participants in their study had less technical training and knowledge than the male ones (Sheng et al., 2010, p. 379). Furthermore, Dhamija et al. (2006) found in their study that "[t]here is no significant difference when comparing the mean scores of males and females" (Dhamija et al., 2006, p. 585), and "[t]here is no significant correlation between participants' scores and ages" (Dhamija et al., 2006, p. 585).

Due to the inconsistent findings from previous research, we only applied descriptive statistics on these demographic data to gain insights about the studied population.

2.1.5 Resulting Model of Phishing Susceptibility at the 2FA stage

Based on all of the above, our resulting model of phishing susceptibility at the 2FA stage comprised the dependent variable *Likelihood to Proceed* and the five independent variables *Stress*, *Attention*, *Elaboration*, *Involvement* and *2FA Frequency*.

2.2 Phishing Mitigation

Vigilance on the part of the user is often recommended as a strategy to mitigate phishing susceptibility (Vishwanath et al., 2011, p. 585). However, vigilance requires a high state of alert that is difficult to maintain at all times (Vishwanath et al., 2011, p. 585). Aside from attention, vigilance arguably also requires elaboration in order to be effective, namely to identify, process and evaluate cues. The high stress level commonly experienced by targets during a phishing attack, however, makes it difficult not only to think rationally but also to recall theoretical knowledge (Swindler, 2021, p. 110), and the intensified focus on the perceived threat (tunnel vision) draws attention away from security-relevant cues and tasks, which for most users are only secondary to begin with (Alsharnouby et al., 2015, p. 69). Research also indicates that awareness campaigns and training are not sufficiently effective in mitigating susceptibility (Williams et al., 2018; Vishwanath et al., 2018). In particular, a series of studies conducted by Ferguson indicates that initial positive effects fade quickly, even within hours, with users falling back into their old patterns (as summarized by Vishwanath et al., 2018, p. 1147). Therefore, it is arguably necessary to find strategies to facilitate alertness and support users in the actual situation when they need to make their decisions.

In principle, there are two possible approaches to support users in the concrete situation: firstly, reducing the burden placed on the user — such as the amount of attention and elaboration required — through automation and predefined workflows, and secondly, providing information and raising awareness right when it is needed. These approaches can arguably complement each other well. In the following we discuss the specific approaches combined in the 2FA app investigated in this thesis.

2.2.1 Automating Detection

In their study of 2021, Sun et al. (2021) introduced the prototype of a 2FA app designed to detect adversary-in-the-middle-based phishing attacks (Sun et al., 2021). As described in more detail in section 1.1, in such a scenario the website the user is interacting with (the phishing website) differs from the website the second factor authentication is requested for (the legitimate website), i.e. their URLs are different. However, the information which website the user is actually visiting is not available to the 2FA service, since the attacker intercepts any communication to and from the user's browser. Sun et al. (2021)'s prototype app, called PhotoAuth, overcomes this problem by prompting the user to take a photograph of their browser's address bar and then analyses said photo using text recognition software (OCR) in order to verify the legitimacy of the URL (Sun et al., 2021). In their study, Sun et al. (2021) successfully conducted a technical proof of concept for this. The PhotoAuth prototype was designed to automatically reject any authentication attempt if the URL verification fails and does not allow users to choose to proceed (Sun et al., 2021, p. 4). Sun et al. (2021) state that false positives are rare (Sun et al., 2021, p. 4), but these can arguably still impair legit use cases, which might be of high importance and urgency to users in real life.

Unlike Sun et al. (2021), we therefore argued that the automation should be used to facilitate, rather than replace user decision-making. This, of course, reintroduced the "human element" social engineering seeks to exploit. What made the PhotoAuth solution still promising in that setting, however, was that it arguably reduces the cognitive burden of the user, including the need for higher-order thinking and information recall. Because the URL verification step is mandatory, the user does not need to remember to do so. Because PhotoAuth prompts the user to take a photo and automates the actual URL verification, the user does not need to remember nor decide what to do or what cues to look out for. Furthermore, the automation arguably also reduces the required extent of attention (taking a photo without needing to pay attention to details such as the URL or security indicators) and elaboration (such as reading and analysing the URL, or interpreting indicators). Overall, the PhotoAuth solution can thus be argued to have a mitigating impact in regards to the factors attention, elaboration, and stress, and is thereby likely to reduce phishing susceptibility.

2.2.2 Supporting User Decision-Making

While a 2FA app is not per se an emergency tool, Swindler argues that a digital product is likely to "be used by someone during or directly following a moment of crisis, even if that is not the product's primary intent" (Swindler, 2021, p. 24). This is arguably particularly true in the case of the phishing scenario this thesis aimed to address, when the user is very likely to be highly stressed by, for example, worrying about

perceived threats suggested by the attackers. Therefore, the app as a whole should be designed in a way that takes into account human stress responses in order to better support users (Swindler, 2021, p. 24). Overall, there are three types of such responses: flight, fight or freeze (Swindler, 2021, p. 137). Swindler states that "[i]t is extremely difficult to predict which [one] will manifest in any given person or situation" (Swindler, 2021, p. 138). It seems reasonable to assume that there is some likelihood for users in our scenario to be experiencing a fight response since they can be seen to actively (not frozen) address (not flee from) the perceived threat. This would make Swindler's de-escalating techniques for UIs (Swindler, 2021, pp. 145–151) particularly interesting for our design. However, even if users are not experiencing a fight response, the basic principles underlying these techniques arguably take into account stress-related effects and impairments, such as the reduced ability to use higher-order thinking, and thereby better support users experiencing high-stress levels in general. We, therefore, included Swindler's advice on de-escalating UIs as well as others of her design tips into the design of the 2FA app investigated in this thesis.

However, in the phishing scenario, the user arguably is stressed because they fell for the phishing narrative, meaning they did not realize they were being tricked. Consequently, they are not aware of the actual threat, namely the phishing attack itself. This means the app also needs to alert them to the real danger. In digital products, warnings are used for this purpose (Bauer et al., 2013, p. 2), and there are many studies on (such as Bravo-Lillo et al., 2013; Felt et al., 2015; Reeder et al., 2018) and established guidelines for their design (such as Bauer et al., 2013; Wogalter et al., 2002; International Organization for Standardization [ISO], 2016).

To create the 2FA app design investigated in this thesis, we built on these findings and guidelines while particularly taking the factors of our theoretical model under consideration. This is outlined in the following.

Interrupt

Prior research showed that blocking warnings, which inhibit the user's previous course of action and force them to actively interact with the warning, are significantly more effective than passive warnings (such as indicators) (Bravo-Lillo et al., 2013, Reeder et al., 2018; Akhawe and Felt, 2013). Interruption can also be seen to be of particular importance in the phishing scenario, when the user is so preoccupied with the perceived threat that they are not likely to notice anything outside their direct line of action (tunnel vision). Only by interrupting that action, therefore, is there arguably a chance of gaining their attention. Furthermore, interruption can also be seen to reduce how much attention is required from the user. Instead of taking an active role and having to remain vigilant the whole time, the user is alerted right in the moment (shorter required duration) using a strong stimulus that is hard to miss (requires little attention to notice). Note that in this we are talking about attention not elaboration, i.e. interruption makes it likelier that the user notices the warning, but does on its own not mean they understand why that happens or what to do about it. Therefore, interruption is arguably only the first step and needs to be combined with other measures to support user decision-making.

Reduce Distractions

In her deescalation techniques, Swindler (2021) advises to minimize distractions and use dedicated interfaces (Swindler, 2021, p. 147). On a similar note, Wogalter et al. (2002) state that removing visual clutter from the vicinity of warning labels has been shown to increase their salience (Wogalter et al., 2002, p. 222). Reducing distractions arguably not only reduces the overall amount of information presented to the user, but also means that the remaining information is relevant to the situation. This in turn can be seen to reduce the amount of attention and elaboration required from the user in order to notice (attention) and identify (elaboration) relevant information as well as potentially increase perceived relevance and thereby involvement.

Reduce Complexity

Being presented with a large number of choices can overwhelm users and lead to so-called *choice paralysis* even under normal circumstances (Swindler, 2021, p. 152). According to Swindler (2021), "[w]hen stress levels rise, having an abundance of choices and no clear direction is a recipe for a freeze response"(Swindler, 2021, p. 152). The resulting indecision can in turn lead to a further increase of the stress level and make it even harder to think rationally (Swindler, 2021, p. 152). In order to prevent this as well as help users

become “unstuck”, Swindler advises giving a clear path forward (Swindler, 2021, p. 152). If the user does need to make a decision (as is the case in our scenario), the number of options should be limited as much as possible (Swindler, 2021, p. 152). To this end, more complex decisions should be split into a series of short decisions, even if this means a longer workflow (Swindler, 2021, p. 153). Similarly, Felt et al. (2015) argue for providing clear instructions and a clear course of action (Felt et al., 2015, p. 2896). Reducing the complexity of decisions to be made can arguably also reduce the amount of elaboration necessary to evaluate the different options.

Opinionated Design

While rational thinking is negatively affected by high stress levels (Swindler, 2021, pp. 22–23), there is little evidence that unconscious choices are also impaired (Swindler, 2021, p. 89). On the contrary, “intuitive decision-making may actually be *more* reliable than rational decision-making [in times of high stress]” (Swindler, 2021, p. 89). It is noteworthy that (according to Swindler (2021)) what Klein refers to as intuition, Kahnemann calls bias (Swindler, 2021, p. 95). Therefore, design can arguably appeal to intuitive thinking by utilizing biases. The “opinionated design” proposed by Felt et al. (2015), which uses “visual design cues to promote a recommended course of action” (Felt et al., 2015, p. 2893), can be seen as such a design, strongly emphasizing and nudging users towards a preferred choice by, for example, utilizing visual attractiveness (Felt et al., 2015, p. 2896). Similarly, Reeder et al. (2018) suggest to de-emphasize the other option, make it more complicated and more effortful to choose (Reeder et al., 2018, p. 3). Furthermore, Felt et al. (2015) state in an opinionated design, users should be able to understand the instruction even without reading (Felt et al., 2015, p. 2896), reducing the need for higher-order thinking (which is required for reading (Swindler, 2021, p. 23)).

Overall, opinionated design arguably reduces the degree of elaboration required. Furthermore, depending on the type and effectiveness of biases utilized, it could also increase involvement.

Visuals

According to the design guidelines compiled by Wogalter et al. (2002), visuals such as colour, high contrast, borders and pictorial symbols can be used to increase the salience of a warning (Wogalter et al., 2002, p. 221). In this, a signal colour is usually chosen to indicate the level of risk. The ISO standard distinguishes three such levels: low, medium and high (ISO, 2016, p. 5). Considering that, on the one hand, the consequences of a phishing attack can be severe, and, on the other hand, the warning in our 2FA app might be triggered by a false positive of the automated URL verification, we argued that the appropriate level for our use case was medium. According to the ISO standard, a combination of orange background and black font should be used for medium risks (ISO, 2016). However, Felt et al. (2015) argued for using the signal colour only as an accent colour, since on digital devices orange does not provide sufficient contrast as a background colour, neither with black (ISO standard) nor white font (Felt et al., 2015). Similarly, Swindler (2021) also emphasizes the importance of contrast between text and background when designing for users experiencing high stress levels (Swindler, 2021, p. 162). In conclusion, it is important to test designs for sufficient contrast and balance the use of signal colours accordingly.

Warnings combining text with pictorial symbols have been found to draw more attention (Wogalter et al., 2002, p. 221) and make the warning more salient (Wogalter et al., 2002, p. 223). However, pictorial symbols are best suited for communicating simple, concrete concepts, and not so effective for abstract ones (Wogalter et al., 2002, p. 223). Phishing as well as adversary-in-the-middle attacks are arguably rather abstract concepts, furthermore ones, which most users can arguably be expected to be only vaguely, if at all, familiar with. For such cases, the ISO standard provides a general purpose symbol: the warning triangle with exclamation mark (ISO, 2016, p. 12), which can be expected to be widely recognized and clearly communicate the risk level.

Comprehensible Message

In order to support users experiencing high stress levels, Swindler (2021) advises to use plain language, simple sentences and (Swindler, 2021, pp. 147–148) “short, simple, common words that can be read intuitively” (Swindler, 2021, p. 162). In this, technical jargon and abbreviations should be avoided (Swindler, 2021,

p. 147). When users are required to make a decision (as is the case when prompted with a warning), they should be provided with all necessary information right in that screen (Swindler, 2021, p. 190). On the other hand, if information is to be understood at a glance, users should only be given as much information as they absolutely need (Swindler, 2021, p. 245). Swindler (2021) also suggests reducing fidelity where applicable, in order to achieve this (Swindler, 2021, p. 245). While the advice regarding understanding information at a glance was given in the context of, for example, medical monitoring systems competing for attention, it is arguably beneficial in any situation in which users can be expected to be inattentive and/or their ability to process information properly is impaired, namely high stress situations in general.

Similar to Swindler's advice, warning design research and guidelines suggest that warning texts should be brief (Felt et al., 2015, p. 2893; Wogalter et al., 2002, p. 224), specific (Felt et al., 2015, p. 2893; Wogalter et al., 2002, p. 221), explicit (Wogalter et al., 2002, p. 221), complete (Wogalter et al., 2002, p. 221), non-technical (Felt et al., 2015, p. 2893) and use simple language (Felt et al., 2015, p. 2893). Overall, the warning message should identify the threat, explain its (potential) consequences and give directives how to avoid it (Wogalter et al., 2002, p. 221). Wogalter et al. (2002) furthermore state the warning should contain a standard signal word in order to draw attention, such as "warning" or "danger" (Wogalter et al., 2002, p. 221). These signal words are both part of the American National Standard Institution (ANSI) standard (according to Wogalter et al., 2002, p. 221) as well as the International Organization for Standardization (ISO) standard (ISO, 2016, p. 5). However, Bauer et al. (2013) argue that the warning icon alone should be used to convey the respective risk level (Bauer et al., 2013, p. 7).

Overall, simplifying the warning text arguably reduces the degree of elaboration required. Furthermore, providing all relevant information on screen arguably reduces the need for higher-order thinking and recalling information from memory.

Familiarity

"Intuition is built through lived experiences" and works based on the recognition of patterns (Swindler, 2021, p. 90). In this, "the more similar [something] is to a known pattern, the faster the subconscious makes the match" (Swindler, 2021, p. 90). According to Swindler (2021), such intuitive comprehension requires no conscious mental effort and involves no uncertainty (Swindler, 2021, p. 90). In order to support users in high stress situations, in which they are likely to rely on intuitive thinking, it is therefore advisable to reuse common patterns and best practices users can be expected to be familiar with. This arguably also reduces the degree of elaboration required, since for example known colour associations or pictorial symbols are recognized without analysing or evaluating them in detail. Furthermore, established standards and guidelines arguably built on experience accumulated from prior research and real-life application, and are likely to comprise methods which have proven to be effective in a wide range of contexts and situations.

2.2.3 Resulting Approach Investigated in This Thesis

The resulting mitigation approach investigated in this thesis was to combine the automated URL verification as proposed by Sun et al. (2021) with verification-based warnings. By taking over the tasks of reading, analysing and evaluating the URL, the automation reduced the need for higher-order thinking on behalf of the user. Furthermore, a warning would be shown to the user, if and only if this verification failed, thereby aiming to increase perceived relevance of the information.

The warnings were designed based on the existing warning design guidelines and findings from prior research discussed in the previous section. In this thesis, it was specifically investigated what impact an opinionated warning design can have compared with an otherwise identical warning design offering a neutral choice, under consideration also of the susceptibility factors included in the theoretical model.

For this purpose, four different warning designs were implemented. Two of these were opinionated, with one promoting the proceed and one promoting the abort option. This aimed to investigate the influence of the "opinionatedness", i.e. the effect of the visual cue, over other aspects of the warning design, such as the message. In addition to the opinionated designs, two neutral choice designs were included in the study. These were identical to the opinionated warning design except for the neutral visual design of the options, thereby not giving visual cues to promote either one of them. The neutral choice designs served to investigate the effect of the absence of "opinionatedness" on users' decisions, compared to that of the

opinionated designs. The only difference between the two neutral choice designs was the order in which the options were presented, aiming to control for potential ordering effects.

3 Method and Implementation

This chapter provides an overview of the methods used in this thesis, a combination of an experiment and a survey. It also includes a description of the artefacts created for this purpose, including a mock-up of a 2FA app. Furthermore, the chapter explores the data collection process for the research, and it discusses the methods employed for data analysis and how the data could help us explore the research questions.

3.1 Method Choice

To address our research questions, we conducted an online study, consisting of two major components: online experiment and survey questionnaires.

We decided on this approach, since we expected to reach a larger group of participants online within the limited time frame, and be able to collect and analyse a larger amount of data using a survey with self-administered questionnaires compared to more time-consuming alternatives, such as interviews. According to Wilson and McLean 1994, a "questionnaire is a widely used and useful instrument for collecting survey information [...], being able to be administered without the presence of the researcher, and often being comparatively straightforward to analyse" (Wilson and McLean 1994, as cited in Cohen et al., 2007, p. 317).

Furthermore, the decision to use both experiment and survey was based on the consideration of how the variables of our model could be measured. In this, we determined that most of the variables (namely stress, attention, elaboration, involvement, and 2FA frequency) could not be observed directly, especially not in an online setting which does not allow for measurement techniques such as tracking eye-movement or monitoring vitals. For these variables, we therefore needed to ask participants questions to measure them, resulting in the use of an online survey with self-administered questionnaires. However, we argued that our dependent variable, the likelihood to proceed, could be observed directly by implementing an interactive mock-up and recording what option (proceed or abort) participants clicked. Because at least some of the processes influencing a user's decision in the investigated scenario are subconscious, as outlined in Chapter 2, we reasoned that we could get a more direct and accurate result by recording actual behaviour than we would have gotten by only asking participants hypothetical questions, which arguably prompt them to use their conscious and rational thinking. While we could have used a pure survey approach for all the variables, we therefore decided to combine it with an experiment. According to Nebeker et al. (n.d.), "an experiment is a study in which a treatment, procedure, or program is intentionally introduced and a result or outcome is observed" (Nebeker et al., n.d., Module 2 Section 2).

As outlined in the previous chapters, stress plays an important role in many real-life phishing cases and consequently many of our decisions in designing our mitigation approach had been informed by considerations of the direct and indirect effects of stress. Aiming to better investigate both the effects of stress as well as the potential of our treatment, we therefore used a 2×4 between-subjects experimental design. In this, participants were randomly assigned firstly into two groups regarding stress (i.e. one being exposed to a stressor and the other not) and secondly into four groups regarding the warning design. The warning designs are shown in the next section, and the implementation of the stressor is discussed below in section 3.3.

Overall, we thus designed a study in which we observed the effect (likelihood to proceed) of our proposed treatment (the design of our interactive mock-up) while also testing for the effects of the factors of our theoretical model, using a combination of experiment and survey.

3.2 Artefact Creation

A fictional scenario was used in this study, for which we created a fictional phishing e-mail that was shown to participants in the survey as well as two interactive mock-ups: one of a login page and one of the 2FA app.

3.2.1 Phishing E-Mail

Taking on the role of attackers when designing the fictional phishing e-mail (Figure 3.1), we decided on a common technique: suggesting a threat and offering a solution. Specifically, the email informed the recipient that there had been a security incident at the online service of their health care provider and their medical data was therefore at risk (threat). We advised them that they must change their password immediately to protect their data and provided a link to do so directly in the e-mail (solution). This e-mail was shown to participants at the beginning of the scenario and served to create a narrative context for the following interactive task. Because susceptibility to phishing e-mails was not the focus of this study and no real e-mails were used anyway, it was decided to allow participants to proceed also via the regular continue button of the survey as an alternative to clicking the emulated link, aiming to mitigate the risk of participants getting stuck at this step.

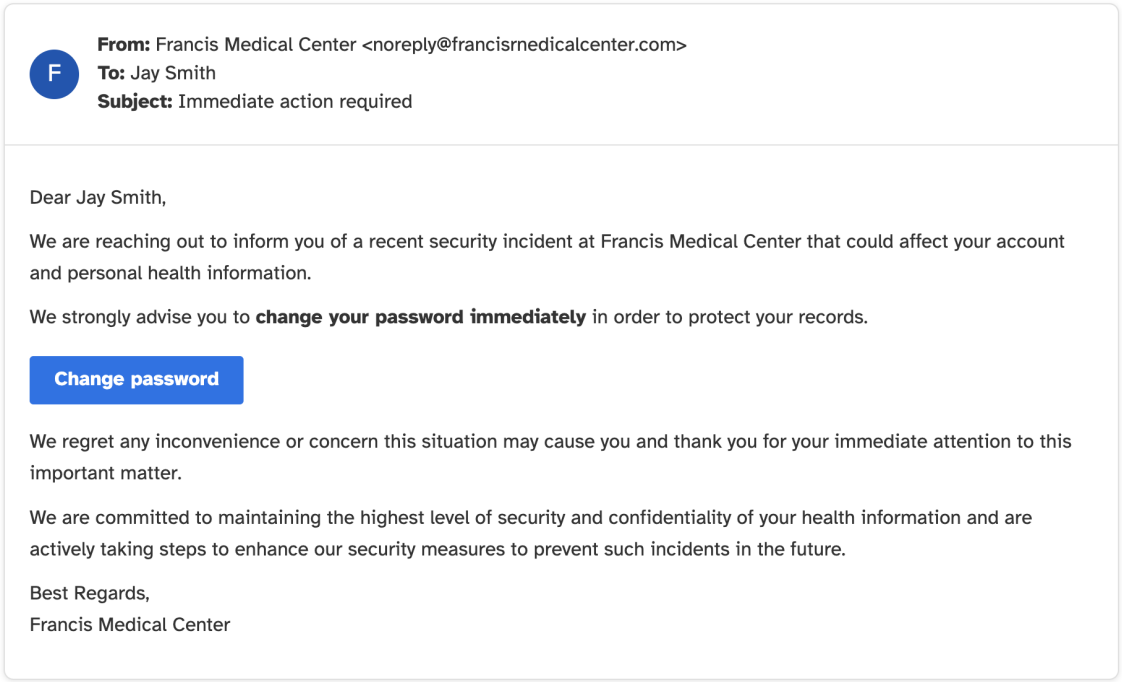

Figure 3.1: Phishing E-Mail

3.2.2 Login Page

After the fake phishing e-mail, participants were shown an interactive mock-up of a login page (Figure 3.2). It was suggested that the link in the e-mail had taken participants there. The username from the fictional scenario was prefilled in the login form. We added subtle phishing indicators by showing an http instead of an https version, and using the fictional domain “francisrnedicalcenter.com” in which the “m” (in “medical”) had deliberately been replaced by “rn” (“r” and “n”). Since attackers cannot register their phishing website under the domain name of the legitimate one, they have to use a different one (Sun et al., 2021, p. 3). In this, using “incorrect and typejacked URLs” is a common technique (Alsharnouby et al., 2015, p. 74). Kumaraguru et al. (2010) found that URLs very similar to the legitimate address confused users in their anti-phishing training software; and only 1 out of 22 participants in the study by Dhamija et al. (2006)) noticed a similar swap of letters with a “w” having been replaced by a “vv”. Furthermore, Swindler (2021) explicitly advises taking measures “to ensure that letter pairs aren’t confused for single letters” (Swindler, 2021, p. 162), suggesting that people in high stress situations are particularly susceptible to such visual deception.

After the participant logged in with the password provided as part of the scenario, they were shown a QR code and asked to scan this with their phone in order to confirm the login.

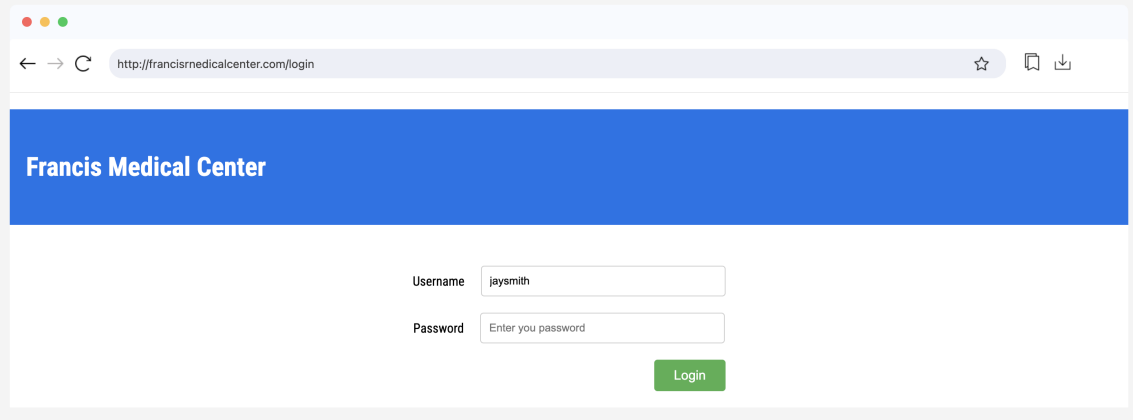

Figure 3.2: Initial Login Page

3.2.3 2FA App

By scanning the QR code, a mock-up of the 2FA app was opened in the browser on the participant's phone. We decided on a web implementation for the mock-up, rather than an actual app, because this way participants did not need to install anything on their phone.

On the first screen of the mock-up app, participants were asked to take a photo of the website they were logging in to. While in a real implementation, the photo- and text-recognition-based URL verification method prototyped by Sun et al. (2021) would have been applied to verify the URL, the mock-up simply always failed all the verification attempts. As a result, a warning was displayed to all participants, alerting them to the risk that the website might not be legitimate. In this, four different warning designs were randomly assigned to participants. Two groups saw a neutral choice design, between which only the order of the buttons (abort vs proceed) was swapped (Figure 3.3 and Figure 3.4). The other two were shown an opinionated design, between which again only the button order (abort vs proceed) was swapped (Figure 3.5 and Figure 3.6). In all cases, participants could then click on abort or proceed/confirm.

3.3 Data Collection

This section covers the recruitment of participants, measurement of the variables, method of the stress inducement, study design (including scenario), and pilot study.

SoSci (www.soscisurvey.de) was used as the sole platform for the entire study, i.e. for both the experiment and survey. "SoSci Survey is a company providing software to create online questionnaires" (Leiner, 2016, p. 374). Its platform is predominantly used by "university students and PhD candidates, who conduct online surveys with convenience samples" (Leiner, 2016, p. 374).

3.3.1 Participants

Considering the limited resources available for this thesis (very limited time and no financial budget), participants were gathered using convenience sampling via Reddit (r/samplesize), which had been used for this purpose in academic research before (Shatz, 2017), as well as via family, friends, colleagues and peers, and their respective networks. Participants had to be 18 or older, fluent in English and did not receive any compensation.

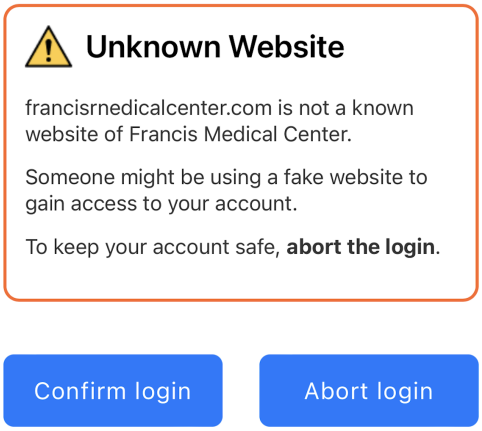

Figure 3.3: Neutral choice warning design with “confirm” on the left and “abort” on the right

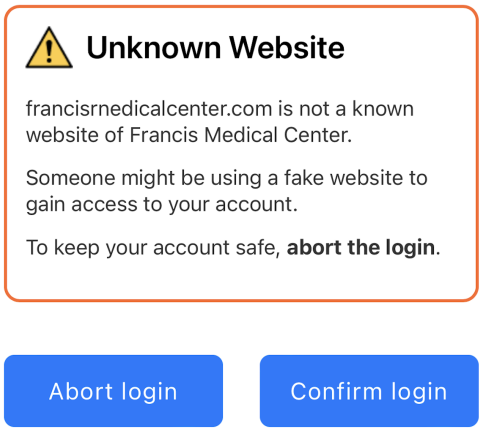

Figure 3.4: Neutral choice warning design with “abort” on the left and “confirm” on the right

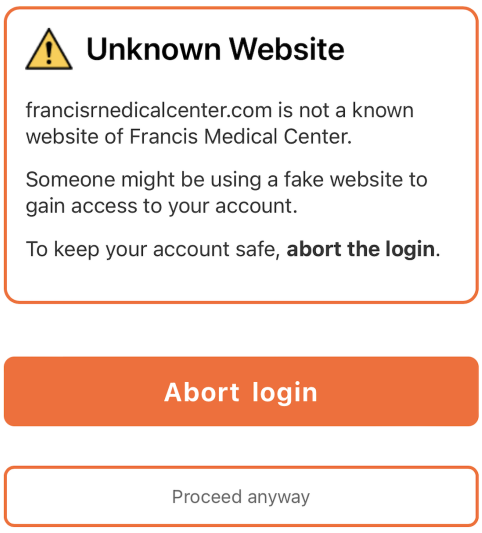

Figure 3.5: Opinionated warning design prompting users to abort

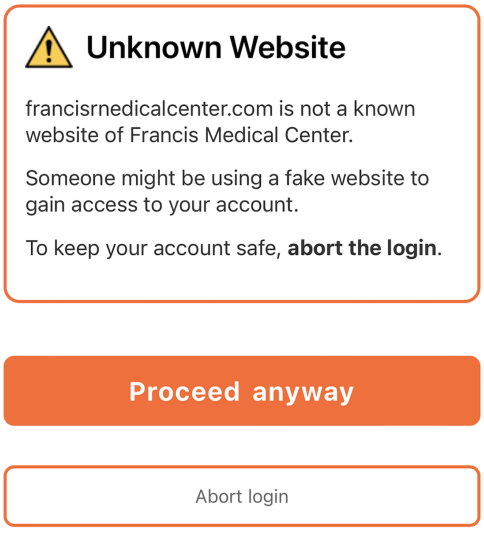

Figure 3.6: Opinionated warning design prompting users to proceed

3.3.2 Measurements

The theoretical model used in this thesis comprised one dependent variable, namely the likelihood to proceed, and five independent variables, namely stress, attention, elaboration, involvement, and 2FA frequency. For descriptive statistical purposes, gender and age group were also collected. In the following, it will be discussed how each variable was measured in our study, as well as how the choices of measurement were informed by and built on prior research. Table 3.1 gives a brief overview of all the variables and their measurements. The full questionnaire can be viewed in the Appendices (Chapter 8).

Table 3.1: Measurement of Variables

Variable	Measurement
Likelihood to Proceed	observed behaviour recorded in experiment 1 variable binary value
Stress	5-item version of state component of State-Trait Anxiety Inventory (STAI) (Zsido et al., 2020) 5 items 5-point items, Likert type
Attention	adapted from 5-item version of the Mindful Awareness Attention Scale (MAAS) (as cited in O. R. Smith et al., 2017) 5 items 5-point items, Likert type
Elaboration	adapted from Message Elaboration Measure (Reynolds, 1997) 12 items 5-point items, Likert type
Involvement	Involvement Scale (Solomon, 2020) 10 items 5-point items, semantic differential
2FA Frequency	direct question 1 item 5-point item, Likert type

Likelihood to Proceed

The dependent variable, likelihood to proceed, measures the decision of a user to abort or proceed with the login. The actual behaviour in the study was observed directly by using an interactive mock-up and recording which option (abort or proceed) participants clicked, resulting in a binary variable.

Stress

According to Zsido et al. (2020), the State-Trait Anxiety Inventory (STAI) by Spielberger et al. (1970) is the most commonly used measure of state and trait anxiety and has been applied to a wide range of contexts (Zsido et al., 2020, pp. 1–2). In previous studies, the questionnaire has been shown to be reliable and valid, and it can be used both in research and clinical practice (Zsido et al., 2020, p. 4). The STAI consists of two components: one for state and one for trait anxiety. In this, the state component measures how the individual feels in the specific moment, i.e. the acute situation, and the trait component how the individual feels generally (Zsido et al., 2020, p. 2). The variable stress in the theoretical model of this thesis refers to the acute stress in the specific situation (see section 2.1.1). Therefore, it was decided to only measure state anxiety. For this, the shorter 5-item version developed by Zsido et al. (2020) was used. Their shortened version of the STAI has been validated in their study (Zsido et al., 2020).

Attention

The Mindfulness Attention Awareness Scale (MAAS) is a well-validated 15-item scale for measuring mindfulness (O. R. Smith et al., 2017, p. 374). According to O. R. Smith et al. (2017), prior research as well as their

own study found the shortened 5-item version of the MAAS to capture the concept equally well (O. R. Smith et al., 2017, pp. 373–374). Mindfulness comprises attention to the immediate experience as well as "an orientation [to one's experience] that is characterised by curiosity, openness and acceptance" (O. R. Smith et al., 2017, p. 373). While for the purposes of this study, the first rather than the second aspect was relevant, the actual questions of the 5-item scale exclusively refer to awareness (of what one is doing), attention and focus, arguably all aspects of the variable attention in the theoretical model of this thesis. Unlike the MAAS, however, the model of this thesis was intended to measure situational attention and not a tendency over time. Therefore, the 5-item MAAS was used to measure attention, but the Likert item points were changed from a frequency range (how often over time) to a degree range (to what extent in the given situation).

Elaboration

The Message Elaboration Measure (Reynolds, 1997) is a scale that was developed to measure the elaboration of messages in general without being limited to a specific message type or format. Hence, it is widely applicable and the questionnaire as well does not contain any references to specific message contents or elements (such as words or icons). Arguably, a warning, such as the one in our study, also constitutes a message. It provides information to the user, regarding both the threat and the possible options available, and requires elaboration in order for the user to decide which action to take. However, the warning message in this study was a lot more concise and specific than messages can be in general, therefore small adaptations in the wording were made to some items of Reynolds's scale in order to adapt them better to this context (such as "not very attentive to the implications" instead of "Not very attentive to the ideas" (Reynolds, 1997, p. 270)).

Involvement

For the IIPM (based on which involvement was included into the model of this thesis), Vishwanath et al. (2011) used a 9-item version of the involvement scale by Zaichkowsky (1985) to measure involvement (Vishwanath et al., 2011, p. 582). However, they did not state which items of the original scale they included. Taking into account the resulting length of our overall questionnaire, we did not want to include the 20-item long original if possible. Therefore, we searched for another instrument to measure involvement. According to Solomon (2020) the involvement scale contained in his book on consumer behaviour is "one of the most widely used scales to assess level of involvement" (Solomon, 2020, p. 181). Furthermore, inspection showed that 9 of its 10 items were identical to items from the scale by Zaichkowsky (1985), and the additional item asked directly about involvement (range "involving - uninvolving"). We therefore decided to use the scale proposed by Solomon (2020) to measure involvement. No adaptions were made.

2FA Frequency

The 2FA Frequency, i.e. how frequently participants use second factor authentication in their daily lives, was measured using a direct question with a 5-point scale ranging from "never" to "every day". In designing this question, we particularly considered two aspects. Firstly, we did not expect every participant to have had previous experience with 2FA. Therefore, we decided to include the answer option "never". Secondly, we did not expect participants, even those who use 2FA apps, to be familiar with this technical term, therefore we avoided it in the question and opted for a descriptive wording instead and including examples which we hoped participants might recognize. In this, we also decided to limit the question to 2FA apps, neglecting other forms of 2FA, since on the one hand, it made the question less complex, and on the other hand, it matched the specific scenario of our thesis. Arguably, user experience is more similar between different 2FA apps in regards to workflow, handling, and habituation, whereas experience with other forms of 2FA (such as SMS-based, card readers, or biometrics) can only partially be applied to 2FA apps.

Additional Questions Regarding Decision-Making

In addition to the questions directly related to the susceptibility factors included in the theoretical model of this thesis, another set of questions was added to the study questionnaires in order to gain further insights into the decision-making of participants regarding their choice to abort or proceed with the login. This

set comprised a list of seven 5-point Likert-scale items, asking specifically about choice difficulty and how participants evaluated their choice made:

1. I was confident in my decision.
2. If I were to come across a similar situation again, I would make the same decision again.
3. I would make the same choice in a real-life situation.
4. I found it difficult to make the decision.
5. I did not think much about the decision.
6. I had an easy time making the decision.
7. I think it was the right decision to make.

Demographics

In addition to the variables of our theoretical model, gender and age group were measured in order to apply descriptive statistics on these demographic data to gain insights about the studied population. Both were measured using a direct question.

3.3.3 Stress Inducement

Almazrouei et al. (2023) proposed and validated "a method to induce stress in human subjects during online participation in research studies without the presence of researchers" (Almazrouei et al., 2023, p. 2575). Prior research had shown that the combination of social–evaluative threats (negative judgement) and uncontrollability "produce[s] the greatest stress response in human subjects" (Almazrouei et al., 2023, p. 2575). In their study, Almazrouei et al. (2023) had participants "answer general knowledge and mathematical questions" people commonly struggle with (Almazrouei et al., 2023, p. 2575). For the stress group, they added time pressure (uncontrollability) as well as emphasized negative feedback and suggested subpar performance for scores below a preset threshold (social–evaluative threats) (Almazrouei et al., 2023, pp. 2576–2577). In their study, they found this implementation to be effective (Almazrouei et al., 2023, p. 2578), which can be seen to be in accordance with prior research.

When deciding on how to induce stress in our study, the approach by Almazrouei et al. (2023) was of particular interest to us because it matched our context (online research without researchers present) and was feasible for us to implement. However, we did not want to risk losing participants by discouraging or offending them with the emphasized negative feedback. While the negativity was meant to create social–evaluative threats as a stress-inducing factor, we expected it to also influence the relationship of participants with the system they were engaging with and its perceived authors, and consequently their willingness to participate, especially since we did not even offer compensation. Unlike in the study by Almazrouei et al. (2023), stress inducement was not our focus. Rather, we needed participants to continue through the main task and survey. Therefore, we decided to "tone down" the feedback to a neutral telling of results, relying more on the time pressure component and possibly on internalized social–evaluative threats, i.e. judging one's own performance. While we thus risked compromising the effectiveness of stress inducement, we argued that mitigating the risk of losing participants took precedence. In order to control for the effect of the stress inducement, we decided on the between-subject design with a control group as well as to measure participants' stress levels both at the beginning of the study and after the interactive task. Furthermore, we argued that the stressor, even if less effective, increased our likelihood of having stressed participants at all, and this was arguably the main purpose of applying a stressor in our study.

The specific stress-inducing tasks and their timing were based on an online concentration test, in which test takers have to count or sum numbers within a deliberately insufficient time span (Psychomeda, n.d.). The test is stated to be a scientific, standardized test by Dr. Satow, and to measure performance, speed, endurance and accuracy (Psychomeda, n.d.). We argued that the test could also be used for our purposes of inducing stress, since Almazrouei et al. (2023) also used mathematical questions in their study, and the concentration test was arguably also designed to be challenging in regards to the measured factors, therefore constituting a test, which we expected people to struggle with to a similar degree as they did with

the questions used by Almazrouei et al. (2023). Furthermore, the concentration test included an arguably established time limit we could reuse.

In summary, our resulting stress inducement implementation combined the approach from the study by Almazrouei et al. (2023) with the concentration test. The tasks and timing were taken from the latter. To this we added immediate feedback for given answers, time out messages and a final score message in accordance with the design by Almazrouei et al. (2023) but with less negative framing.

3.3.4 Study Design

To give participants a mental context for their interaction with our mock-ups, we created a fictional scenario, in which participants were asked to imagine they had registered with a digital service of their healthcare provider and then received an email claiming that the service had been compromised and they needed to change their password immediately in order to keep their account safe. We chose the healthcare context because we assumed that the data at risk, medical data, would be perceived as particularly sensitive, increasing motivation to protect it.

Before taking part in the study, participants were presented with an informed consent form, which they had to acknowledge. Our study then comprised three phases (see Figure 3.7), which are outlined in the following.

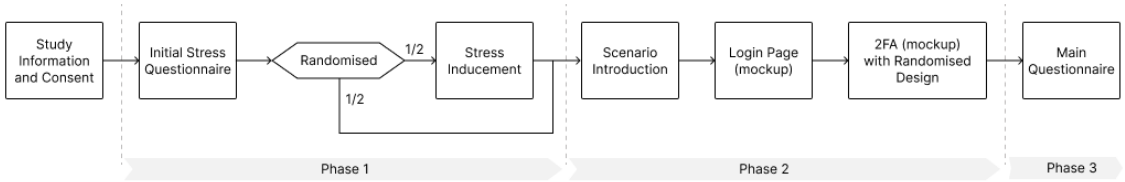

Figure 3.7: Study Design

Phase 1: Entry Questionnaire and Randomized Stress Inducement

In phase 1 of the study, participants were first asked about their stress level, using the stress scale from our questionnaire. Then, participants were randomly assigned into two groups. For one of the groups the stress inducement (as detailed above) was applied, the other could directly progress to the next phase.

Phase 2: Interactive Task with Randomized Design Allocation

At the beginning of the interactive task, participants were introduced to the scenario and shown the fictional phishing email. They then progressed to the interactive mock-up of the login page. After credentials had been entered (provided as part of the scenario), the login page showed a QR code, prompting participants to scan it and confirm the login. By scanning the QR code, they opened the interactive mock-up of the 2FA app on their phone. The 2FA mock in turn asked them to take a photo of the login page's address bar. Once they had provided a photo, a warning about the website was displayed to all participants in the 2FA mock-up. In this, participants were randomly assigned one of the four warning designs. In all designs, the participants could then choose between the option to proceed with or abort the login, which completed the interactive task.

Phase 3: Main Questionnaire

After they had completed the interactive task, participants were directed back to the main survey and asked to fill out the questionnaire. In this, all questions were mandatory except for the demographic ones and one open question in which participants could make a statement about the reason behind their decision.

3.3.5 Pilot Study

Using the pre-test feature of SoSci, a pilot study with six participants was run to identify potential issues with the experimental setup and questionnaire design, and improve them accordingly. As a result, some technical issues with the interactive task were identified and fixed, and some instructional texts were adjusted to make them clearer. Furthermore, strong feedback from two of the test participants led us to reconsider and eventually reduce the length of the stress inducement task. These participants stated the time and effort required for the original stress inducement task upset them to the degree that it made them angry or consider quitting the study. This arguably posed a similar predicament again as discussed in the initial design of that task, namely how to balance the effectiveness of the stress inducement with the effect it might have on participants' willingness to continue with the study. While the concern was that shortening the task might negatively affect its effectiveness, after careful consideration precedence was again given to mitigating the risk of losing participants, in accordance with the reasoning behind our original approach (see section 3.3.3).

3.4 Data Analysis

The data collected in this study was analysed with IBM SPSS Statistics 27.

The regression model comprised all variables of the underlying theoretical model, except for elaboration (due to scale validation issues, see section 4.2): i.e. stress (experienced during the interactive task), attention, involvement and 2FA frequency as independent, and likelihood to proceed as the dependent variable. In addition, the treatment tested in the experiment, i.e. the respective warning design, as well as a new construct derived from the set of questions introduced to gain insights on the decision-making of participants – "confidence in the decision made" (see section 4.2) – were included as independent variables.

Due to the binary nature of the outcome variable, logistic regression was used for the model analysis. In addition, descriptive statistics and correlations between the variables were calculated. The established scales were checked using principal component analyses. The new scale "confidence in the decision made" was examined using exploratory factor analysis. The effect of the stress inducement was tested using an independent sample t-test of stressor (if the stress inducement had been applied) with the change between the respective participant's initial stress level and their stress level experienced during the interactive task.

3.5 Validity and Reliability

After the pilot study, the comments and input of the participants were evaluated for a first preliminary assessment of the validity and reliability of our measurements.

Overall, established scales were used to measure the constructs in our theoretical model as much as possible. Because they were applied to a new context and adaptations had been made to some, principle component analyses were conducted in order to test their respective content validity. Similarly, for the new set of questions regarding participants' decision-making (see section 3.3.2), exploratory factor analysis was used. In both cases, reliability was checked using Cronbach's α.

In order to facilitate the replicability of our study, a full copy of the final questionnaire is provided in the Appendices (Chapter 8).

3.6 Considerations

According to the ENISA Threat Landscape 2023 report, social engineering attacks "persist as a substantial threat to users" (ENISA, 2023, p. 71) and continue to be on the rise (ENISA, 2023, p. 4), with phishing predicted to remain the most prevalent approach over the long term (ENISA, 2023, p. 73). The purpose of this thesis was to contribute to the research concerning the mitigation of these attacks by investigating if and how a combination of automated URL verification and verification-based opinionated warning design can influence user decision-making at the stage of 2FA login. In addition, we investigated how psychological and individual factors can influence this decision-making. To the best of our knowledge, this combined approach

of automation and opinionated design at the stage of 2FA with consideration of psychological and individual factors had not previously been investigated.

The theoretical model underlying this thesis was based on prior research, and existing scales were reused to measure the contained constructs, with adaptions kept as minimal as possible. In addition, statistical methods were applied to test validity and reliability as described in the previous section. Using SoSci for this study not only allowed us to use their infrastructure and secure servers, but arguably also to rely on the accumulated experience and best practices of this research platform, for example, regarding the interface design of different question types or the conduction of the pilot study. Method choice and study design were informed by and based on prior studies and existing methods (as has been described in the previous sections), and the decisions made were guided by considerations of feasibility, reliability, and ethical responsibility.

Because stress plays a major role in phishing susceptibility and was therefore a factor of the theoretical model, we needed to attempt inducing stress in our participants in order to be able to measure resulting effects. However, because of ethical considerations we decided to limit both the type and expected intensity of the stress, aiming to balance the research interest with the well-being of the participants. We, therefore, chose a study design in which participants knew that the scenario was fictional and they were not exposed to any real threat. Furthermore, we expected that stress inducement based mainly on time pressure and to a degree on judgement of performance outside of any real-life context would result in mild to moderate "everyday" stress during the task (if any) but was not likely to result in severe or lasting distress.

Due to ethical and privacy considerations, we also required participants to be 18 years or older and acknowledge an informed consent form before taking part in the study.

While we decided on interactive mock-ups, hoping to collect more accurate data through actual behaviour than we might have gotten relying only on hypothetical questions, we chose a study design in which participants did not receive an actual email, nor did they interact with real systems. Firstly, this arguably increased participants' awareness of engaging with a fictional scenario (see above). Secondly, it ensured that no real data or accounts were at risk. Thirdly, the susceptibility at the e-mail stage was not the focus of this thesis, rather the e-mail was only needed to create the background of the scenario. Therefore we did not want to risk losing participants before the 2FA stage (by them potentially spotting the deception at the e-mail stage and not clicking on the link), and furthermore wanted to avoid unnecessarily collecting personal data (e-mail address).

Instead of implementing an actual app, we decided to use a web application for the 2FA mock-up. The main intention in this was that this way participants did not have to install anything on their phones, which we expected to increase people's willingness and ability to participate, and thus improve feasibility (sample size). In addition, SoSci could thus be used as the sole platform, reducing the effort and infrastructure needed to set up the study, for which we had only very limited time and no budget. While the interactive task can be seen to be less realistic using a web rather than an app mock-up, we argued that it was only slightly so, since the UI and the core interactions were the same, except for being presented in the "frame" of a browser and the user in the end having to close the browser manually rather than the app closing itself. Furthermore, participants were already aware of being in a fictional (i.e. not real) scenario and we expected it not to take much effort to image the mock-up being an app. Overall, we argued the benefit of increased feasibility outweighed this slight compromise in the 2FA mock-up.

During the interactive task, participants were required to take a photo of the (fictional) website. However, for the purpose of the study the content of this photo was not needed. The prototype simply always failed the URL validation regardless of what the photo contained. For this reason, the photo was never processed nor stored. Consequently, even if there were personal or sensitive data contained in such a photo, that data was not accessed. In addition, faking the URL verification also meant that participants did not need to succeed in taking a correct photo (capturing the address bar) of sufficient quality (focus, lighting, etc.), thus reducing the risk participants would get stuck at this stage.

Since 2FA applications are used to confirm logins and transactions in online contexts, this study was arguably investigating online users. Hence, we argued it was adequate to conduct this study and recruit participants online as well. However, because of the limited resources for this thesis (very limited time and no budget) convenience sampling was arguably the only feasible recruitment method. Therefore, the resulting sample was unlikely to be representative and the generalizability of the findings needs to be validated in future research. Hoping to at least mitigate some of the sampling bias resulting from recruiting participants via our direct and extended networks, which we expected in particular to have a higher than usual percentage

of participants with IT backgrounds due to our own such backgrounds, participants were also recruited via Reddit (r/samplesize).

4 Results

In this chapter, the data collected in the present study and the results of the analyses are presented. First, general descriptive statistics are shown, such as overall sample size, frequencies, and average time taken to complete the study. Then the validation and resulting adjustments of the scales are presented as well as the descriptive statistics of their respective distributions. Next, correlations between the continuous variables are explored. Then the results from testing for the effect of the stress inducement are shown, followed by the results of the logistic regression for the overall model. Finally, a descriptive overview of participants' comments on why they made the choice they made is given.

For all assessments of significance, the benchmark of $.05$ was used. Overall, the collected sample ($N = 94$) was too small, in particular for the logistic regression analysis, for which the sample size should at least have been 400 (Hair et al., 2019). Therefore, interpretations of the results must be made with caution.

4.1 Descriptive Statistics

96 valid datasets were collected in the study. This excluded cases, in which respondents did not complete the study, as well as a few incidents, in which technical issues might have corrupted the data. Of the valid datasets, two had to be removed before the data analyses due to group sample size issues regarding 2FA frequency (see below). Consequently, the final overall sample size was 94.

On average, the 94 participants completed the study in 6:24 minutes.

42 of the participants identified as male (44.7%), 34 as female (36.2%), the rest preferred not to disclose their gender. No participants selected "non-binary".

Table 4.1 shows the age group distribution of the study participants. The majority (54.3%) was between 25 and 34 years old, followed by 35- to 44-year-olds at 30.9%.

Table 4.1: Age Group Frequencies

	N	%
18 to 24 years old	6	6.4%
25 to 34 years old	51	54.3%
35 to 44 years old	29	30.9%
45 to 54 years old	8	8.5%

Participants were randomly assigned into two groups, one of which was exposed to the stress inducement, while the other was not. These two groups were almost equal in size, with 45 participants having been exposed to the stress inducement and 49 participants not. The discrepancy from a completely equal distribution was due to the removal of invalid datasets (see above).

Participants were, in addition, also randomly assigned into four groups determining which of the four warning designs a participant was shown. The size of these design groups ranged from 19 to 27 participants (see Table 4.2). Again, the discrepancy from a completely equal distribution was due to the fact that not all initially collected datasets were valid (see above).

Table 4.3 shows the frequencies for the variable measuring 2FA frequency, i.e. how often respondents commonly use 2FA in their daily life. Because the sample size of the "never" group was only two and therefore too small to draw statistical conclusions, the two corresponding datasets were removed from the overall sample, reducing the overall sample size to 94, as mentioned above.

Overall, with 58 of participants the majority (61.7%) chose to abort, while the remaining participants proceeded with the login.

Table 4.2: Design Group Frequencies

	N	%
opinionated design prompting to proceed	26	27.7%
opinionated design prompting to abort	27	28.7%
neutral choice design: “abort” left, “confirm” right	19	20.2%
neutral choice design: “confirm” left, “abort” right	22	23.4%

Table 4.3: 2FA Frequencies

	N	%
Never	2	2.1%
Up to once a month	11	11.5%
Up to once a week	14	14.6%
Up to once a day	41	42.7%
Up to several times a day	28	29.2%

4.2 Scales

In this section the validation and resulting adjustments of the scales are presented as well as the descriptive statistics of their respective distributions. In this, the following benchmarks were used:

- Cronbach's α values $.7$ were considered as acceptable, $.8$ as good, and $.9$ as excellent (Gliem and Gliem, 2003).
- Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) values of $\geq .7$ were considered desirable, and values $< .5$ as unacceptable for factor analysis to be appropriate (Watkins, 2018).
- Communalities $\geq .4$ were considered to signify enough shared variance for factor analysis to be appropriate (Costello and Osborne, 2005).
- For factor loadings, $.3$ was considered the minimum threshold for interpretation, loadings $\geq .5$ as practically significant, and loadings $\geq .7$ as desirable (Hair et al., 2019).

Stress

The stress level of participants was measured using an established 5-item version of the state component of the State-Trait Anxiety Inventory (STAI) (Zsido et al., 2020). This was done twice in the study: once at the very beginning asking participants about their current stress level (Initial Stress) and once directly after the interactive task, asking participants about their stress level during the task (Stress During the Interactive Task). Of these two measures, only the latter was a factor of the theoretical model. The initial stress level, on the other hand, was only recorded in order to be able to test for the effect of the stress inducement (see section 4.4).

Initial Stress

Communalities of the scale items ranged from $.491$ to $.715$, signifying enough shared variance for factor analysis to be appropriate. Furthermore, the Bartlett's test was significant ($p < .001$), and together with the the KMO of $.80$ indicated that principal component analysis (PCA) was acceptable for the sample.

The PCA indicated one component with an eigenvalue of 3.18 and 63.52% of variance explained. All factors loaded very well onto this component with values ranging from $.700$ to $.846$ (see Table 4.4).

The reliability test indicated a good reliability of the scale (α = $.85$).

Table 4.4: Component Matrix for Initial Stress

	Component 1
upset	.700
frightened	.796
nervous	.846
jittery	.811
confused	.825

Possible values for the stress scale ranged from 1 to 5, with 1 representing the lowest and 5 the highest stress level. The overall distribution of scores for initial stress in the collected sample ranged from 1.00 to 3.20 ($M = 1.57, Mdn = 1.40$).

Stress During the Interactive Task

Communalities of the scale items ranged from $.542$ to $.797$, signifying enough shared variance for factor analysis to be appropriate. Furthermore, the Bartlett's test was significant ($p < .001$), and together with the KMO of $.83$ indicated that principal component analysis (PCA) was acceptable for the sample.

The PCA indicated one component with an eigenvalue of 3.34 and 66.85% of variance explained. All factors loaded very well onto this component with values ranging from $.736$ to $.893$ (see Table 4.5).

Table 4.5: Component Matrix for Stress During the Interactive Task

	Component 1
upset	.758
frightened	.834
nervous	.893
jittery	.856
confused	.736

The reliability test indicated a good reliability of the scale ($\alpha = .86$).

Possible values for the stress scale ranged from 1 to 5, with 1 representing the lowest and 5 the highest stress level. The overall distribution of scores for stress during the interactive task in the collected sample ranged from 1.00 to 3.80 ($M = 1.67, Mdn = 1.60$). In this, only 3 participants (3.2%) had stress scores above 3. Furthermore, there was no significant difference between participants with scores below the median vs above ($p = .296$).

Attention

Attention was measured using a scale adapted from an established 5-item version of the Mindful Awareness Attention Scale (MAAS) (as cited in O. R. Smith et al. (2017)).

Communalities of the scale items ranged from $.703$ to $.833$, signifying enough shared variance for factor analysis to be appropriate. Furthermore, the Bartlett's test was significant ($p < .001$), and together with the KMO of $.86$ indicated that principal component analysis (PCA) was acceptable for the sample.

The PCA indicated one component with an eigenvalue of 3.79 and 75.82% of variance explained. All factors loaded very well onto this component with values ranging from $.839$ to $.913$ (see Table 4.6).

The reliability test indicated an excellent reliability of the scale ($\alpha = .92$).

Possible values for the attention scale ranged from 1 to 5, with 1 representing the lowest and 5 the highest attention level. The overall distribution of attention scores in the collected sample ranged from 1.20 to 5.00, ($M = 3.61, Mdn = 3.80$).

Table 4.6: Component Matrix for Attention

	Component 1
not running on automatic	.858
not rushing through unattentively	.913
moment not eclipsed by goal	.892
acting with awareness	.850
paying attention	.839

Involvement

Involvement was measured using the established Involvement Scale (Solomon, 2020).

Communalities of the scale items ranged from $.540$ to $.800$, signifying enough shared variance for factor analysis to be appropriate. Furthermore, the Bartlett's test was significant ($p < .001$), and together with the KMO of $.87$ indicated that principal component analysis (PCA) was acceptable for the sample.

However, the PCA indicated 2 components with eigenvalues of 5.46 and 1.29, respectively, and cumulative variances explained of 54.62% and 67.46%. Table 4.7 shows the factor loadings for these two components.

Table 4.7: Rotated Component Matrix for the Complete Involvement Scale

	Component 1	Component 2
irrelevant / relevant	.868	
not needed / needed	.857	
worthless / valuable	.820	.357
unimportant / important	.806	
uninvolving / involving	.753	.324
means nothing / means a lot	.724	.328
unappealing / appealing	.588	.441
unexciting / exciting		.854
mundane / fascinating		.725
boring / interesting		.711

Running another factor analysis for only the suggested component 1 resulted in a Bartlett's test that was significant ($p < .001$), and a KMO of $.87$, again indicating that principal component analysis (PCA) was acceptable for the sample. Furthermore, the PCA now indicated only one component with an eigenvalue of 4.72 and 67.45% of variance explained. All items loaded very well unto this component with values ranging from $.716$ to $.897$. Compared to the PCA for the complete scale, cumulative variance explained remained almost the same (67.46% vs 67.45%), and the Cronbach's α increased from $.91$ to $.92$, indicating a higher reliability of the reduced scale.

Looking at the items from a conceptual point of view, the items of component 1 arguably all relate well to the context of security concerns, whereas the items of the second component can be seen not to relate to that context well. Participants were asked about their involvement regarding the context of "logging in securely", which arguably is part of a security concern context. Considering this, the first component can be seen as a good match for the measurement intended for the purposes of this study. While according to the principal component analysis the item "unappealing / appealing" could have been included in the first component, it arguably also does not relate well to the security concern context and was hence removed in the next step.

Running a factor analysis for this conceptually motivated reduction resulted in a Bartlett's test that was significant ($p < .001$), and a KMO of $.86$, again indicating that principal component analysis (PCA) was

acceptable for the sample. The PCA again indicated only one component, this time with an eigenvalue of 4.27 and 71.15% of variance explained. All items loaded very well onto this component with values ranging from $.782$ to $.903$ (see Table 4.8). Compared to the purely statistically motivated reduced scale from the previous step the variance explained increased from 67.45% to 71.15%, indicating a better fit, and the Cronbach's α remained almost identical with $.919$ (statistical) vs $.918$ (conceptual), signifying high reliability also of the conceptually motivated reduced scale.

Table 4.8: Component Matrix for Conceptually Reduced Involvement Scale

	Component 1
worthless / valuable	.903
not needed / needed	.892
irrelevant / relevant	.855
uninvolving / involving	.821
means nothing / means a lot	.801
unimportant / important	.782

Because the second component was arguably not of relevance to the purpose of this study, it was removed from the data analysis. In conclusion, the involvement scale was adjusted to include only the items relevant to the context of security concerns (see Table 4.8).

Possible values for the resulting involvement scale ranged from 1 to 5, with 1 representing the lowest and 5 the highest involvement level. The overall distribution of involvement scores in the collected sample ranged from 1.17 to 5.00 ($M = 3.93, Mdn = 4.17$).

Elaboration

Elaboration was measured using a scale adapted from the Message Elaboration Measure (Reynolds, 1997).

Communalities of the scale items ranged from $.437$ to $.708$, signifying enough shared variance for factor analysis to be appropriate. Furthermore, the Bartlett's test was significant ($p < .001$), and together with the KMO of $.76$ indicated that principal component analysis (PCA) was acceptable for the sample.

The PCA indicated three components (see Table 4.9) with eigenvalues of 3.47, 2.32 and 1.43, respectively, and cumulative variances explained of 28.89%, 48.20% and 60.11%. However, the results of subsequent factor analyses and reliability tests for these subcomponents remained weak. In particular, the Cronbach's α for component 2 and 3 fell below the threshold of $.7$ with $.54$ and $.69$, respectively, signifying insufficient reliability; and component 1 and 2 only explained 55.79% and 51.53% of the variance, respectively, falling below the common 60% benchmark for how many components to retain in factor analysis.

Furthermore, splitting the elaboration scale as indicated also did not make sense conceptually. In addition, Vishwanath et al. (2011) — who had also included elaboration in the IIPM — had found no evidence in their study in support of elaboration as a significant predictor. Considering all of the above, the entire elaboration scale was removed from the data analysis in this study.

Confidence in the Decision Made

For the set of questions regarding participants' decision-making, an exploratory factor analysis was conducted.

Communalities of the items ranged from $.473$ to $.773$, signifying enough shared variance for factor analysis to be appropriate. Furthermore, the Bartlett's test was significant ($p < .001$), and together with the KMO of $.77$ indicated that a factor analysis was acceptable for the sample.

The factor analysis indicated 2 components with eigenvalues of 3.37 and 1.63, respectively, and cumulative variances explained of 48.19% and 71.43%. Table 4.10 shows the factor loadings for these two components.

Considering that component 2 only contained two items and "[a] factor with fewer than three items is generally weak and unstable" (Costello and Osborne, 2005, p. 5), these two items were removed and another

Table 4.9: Rotated Component Matrix for Elaboration

	Component 1	Component 2	Component 3
concerned with implications	.789		
not distracted	.769		
attentive to implications	.698		
not resting mind	.671		
pondering message		.811	
extending cognitive effort		.711	.368
not taking it easy		-.705	
exerting mind	.585	.600	
analyzing message		.567	.315
reflecting on implications			.807
pondering implications			.787
pondering what was written		.326	.645

Table 4.10: Pattern Matrix for Complete Scale Confidence in the Decision Made

	Factor 1	Factor 2
would repeat decision	.881	
decision easy	.792	
right decision	.751	
confident in decision	.724	
same decision in real-life	.689	
unreflected decision		.839
decision difficult		.755

factor analysis run for the reduced scale, which equaled component 1. This resulted in a Bartlett's test that was significant ($p < .001$), and a KMO of $.86$, again indicating that factor analysis was acceptable for the sample. The analysis now indicated only one component with an eigenvalue of 3.36 and 67.15% of variance explained. All items loaded very well onto this component with values ranging from $.690$ to $.877$ (see Table 4.11).

Table 4.11: Factor Matrix for Reduced Scale Confidence in the Decision Made

	Factor 1
would repeat decision	.877
decision easy	.795
right decision	.752
confident in decision	.722
same decision in real-life	.690

Compared to the factor analysis for the complete set of items, the cumulative variance explained decreased from 71.43% to 67.15%, but the Cronbach's α increased from $.74$ to $.87$, indicating a higher reliability of the reduced scale.

Looking at the resulting scale from a conceptual point of view, all items can be seen to relate to the confidence of the participant in the decision they made. In this, the item "decision easy" can be seen to relate to confidence in that the more confident one is and thereby the less doubts one has, the easier it can

feel to make a decision. Based on this reasoning, the resulting scale will in the following be referred to as "confidence in the decision made", or "confidence" for short.

Possible values for the confidence scale ranged from 1 to 5, with 1 representing the lowest and 5 the highest confidence level. The overall distribution of scores for confidence in the collected sample ranged from 1.20 to 5.00 ($M = 3.91, Mdn = 4.00$).

4.3 Correlations

In order to assess linear relationships between the continuous variables, a Pearson bivariate correlation analysis was conducted.

As can be seen in Table 4.12, the results showed a significant positive correlation between attention and involvement as well between attention and confidence in the decision made.

Furthermore, there was a significant negative correlation between stress experienced during the interactive task and confidence in the decision made; and a significant positive correlation between stress experienced during the interactive task and initial stress at the beginning of the study.

Table 4.12: Correlations Between Variables

	IST	STR	ATT	INV	CON
initial stress (IST)	1				
stress during interactive task (STR)	.687*	1			
attention (ATT)	-.099	-.054	1		
involvement (INV)	-.057	.067	.497*	1	
confidence in the decision made (CON)	-.145	-.254*	.364*	.197	1

*Correlation is significant at the $.05$ level (2-tailed).

4.4 Effect of the Stress Inducement

In order to test the effect of the stress inducement, an independent sample t-test was conducted of stressor (indicating if the stress inducement had been applied or not) with the change between the respective participant's initial stress level and their stress level experienced during the interactive task. The result of the t-test showed no significance, $t(84) = -.57, p = .57$. Therefore, the null hypothesis could not be rejected, meaning that there was no evidence that the stress inducement had an effect on the change in stress level.

Neither stressor nor the initial stress level of participants were part of the theoretical model and their recording served only to test for the effect of the stress inducement (if any). In addition, in a preliminary logistic regression model with all recorded variables, neither of the two had shown significance (with $p = .960$ and $p = .671$, respectively). Hence, neither stressor nor initial stress level were later included in the regression model.

4.5 Logistic Regression

The regression model comprised all variables of the underlying theoretical model, except for elaboration (see reasoning in section 4.2 above): i.e. stress (experienced during the interactive task), attention, involvement and 2FA frequency as independent, and likelihood to proceed as the dependent variable. In addition, the treatment tested in the experiment, i.e. the respective warning design, as well as the new construct "confidence in the decision made" (see section 4.2 above) were included as independent variables.

Due to the binary nature of the outcome variable (proceed), logistic regression was used for the analysis of the overall model.

Table 4.13: Dummy Variables for the Warning Designs

Design	**Variable**
opinionated design prompting to proceed (see Figure 3.6)	*benchmark category*
opinionated design prompting to abort (see Figure 3.5)	opinionated design prompting to abort
neutral choice design: “abort” left, “confirm” right (see Figure 3.4)	neutral design: “abort” left, “confirm” right
neutral choice design: “confirm” left, “abort” right (see Figure 3.3)	neutral design: “confirm” left, “abort” right

Table 4.14: Dummy Variables for 2FA Frequency

2FA Frequency	**Variable**
up to several times a day	*benchmark category*
up to once a day	2FA frequency up to once a day
up to once a week, *or* up to once a month	2FA frequency up to once a week or less

4.5.1 Introduction of Dummy Variables

Before conducting the logistic regression, the categorical variables were recoded to match the prerequisites of this analysis method. In this, dummy variables were introduced for the designs, with the opinionated design prompting participants to proceed as the benchmark category. This resulted in three design variables as shown in Table 4.13.

For 2FA frequency, group sample sizes were uneven. In particular, the sample sizes for “up to once a month” ($n = 11$) and “up to once a week” ($n = 14$) were far smaller than those of the other two groups, resulting in an overall uneven distribution. Furthermore, in the theoretical framework underlying this thesis, overwhelm due to high 2FA load as well as habituation effects were argued to explain the assumed impact of 2FA frequency on the likelihood to proceed. Arguably, neither up to monthly nor up to weekly 2FA use are likely to result in much overwhelm or habituation. Considering this as well as the small overall sample size collected in this study, the groups “up to once a month” and “up to once a week” were joint into “up to once a week or less”, thereby reducing the number of dummy variables from three to two as well as increasing the minimum group sample size from 11 to 25. The resulting 2FA frequency variables are shown in Table 4.14.

4.5.2 Testing Prerequisites for Logistic Regression

According to Hair et al. (2019), the assumptions of logistic regression comprise independence of observations, and linearity of the logit, i.e. the linearity of the independent variables with the dependent variable. The minimum overall sample size should be 400, and the size of the smaller outcome group should be at least 20. For categorical variables, all cell frequencies in the cross tabulation with the dependent variable “should be greater than 1 and a large majority (75% or more) [...] greater than 5” (Hair et al., 2019, p. 556). This also covers that there should be no complete or semi-complete separation for categorical variables, which would result in some cells having frequencies of 0. For continuous variables, there should be no complete sepa-

ration, i.e. there should not be a pattern where all abort cases are on one side of the variable's range and all proceed cases on the other. Furthermore, there should be no multicollinearity between the independent variables.

Due to the nature of the research design, observations in this study were independent.

The overall sample size of 94 was below the required minimum of 400. The size of the smaller outcome group was 36 (i.e. 36 participants proceeded with the login), which is above the threshold of 20.

For design, all cell frequencies were above the threshold of 5 (minimum cell frequency 6). For 2FA frequency, all cell frequencies were above 1 and 75% were above 5, thereby also meeting the required thresholds. The analysis of the scatter plots of the continuous independent variables with the dependent variable showed no complete separation.

The Box-Tidwell procedure was used to test the linearity of the independent variables with the dependent variable, following the instructions by Hair et al. (2019). The results indicated stress as the only variable exhibiting potential nonlinear effects. Adding its interaction term to the overall model showed a significant, but small improvement in model fit ($\chi^2(1) = 4.25, p = .039$, with $Nagelkerke R^2$ increasing from $.557$ to $.593$). However, the interaction term itself was not significant ($p = .053$). In addition, stress and its interaction term had a high significant correlation ($r(92) = .99, p < .001$), meaning the two variables were nearly identical and adding the interaction term would introduce multicollinearity, violating another assumption of logistic regression. Furthermore, due to the small sample size, the reliability of the significance tests was questionable and therefore results needed to be interpreted with caution. Taking all of this into account, no interaction terms were added to the model.

No multicollinearity was found between the independent variables.

In conclusion, of the requirements for logistic regression, the sample size requirement was not met in this study, meaning that the results of the analysis must be interpreted with caution.

4.5.3 Univariate Significance

"A series of simple logistic regression models were estimated with each independent variable as the only variable in the model to provide the univariate significance tests" (Hair et al., 2019, p. 581).

As can be seen in Table 4.15, the results indicated that attention, involvement, confidence in the decision made and 2FA frequency up to once a day each had a significant negative effect on the likelihood to proceed; while 2FA frequency up to once a week or less had a significant positive effect. None of the remaining variables showed a significant impact.

Table 4.15: Univariate Analysis of Independent Variables

Independent Variable	Logit Coefficient	Significance
stress during the interactive task	.008	.980
attention	-.825	.002
involvement	-1.296	.000
confidence in the decision made	-1.171	.004
2FA frequency up to once a day	-1.094	.016
2FA frequency up to once a week or less	1.721	.000
neutral design: "confirm" left, "abort" right	.631	.200
neutral design: "abort" left, "confirm" right	-.368	.501
opinionated design prompting to abort	.360	.437

4.5.4 Results Logistic Regression

The results of the logistic regression analysis showed that the overall model was significant with $\chi^2(9) = 49.58, p < .001$, and Nagelkerke $R^2 = .557$. 84% of the predictions were correct for a cut value of $.5$ (see classification table in Table 4.16).

Table 4.16: Classification Table for Logistic Regression Model

		Observed	abort	proceed	Percentage Correct
Step1	proceed	abort	53	5	91.4
		proceed	10	26	72.2
		Overall Percentage			84.0

As can be seen in Table 4.17, the independent variables involvement and confidence in the decision made were significant and had a negative impact on the likelihood to proceed. This means, higher involvement and higher confidence each reduced the likelihood to proceed.

Table 4.17: Variables in the Logistic Regression Equation

	B	S.E.	Wald	df	sig	Exp(B)
stress during the interactive task	.075	.567	.018	1	.894	1.078
attention	.481	.439	1.202	1	.273	1.617
involvement	-1.450	.414	12.282	1	.000	.235
confidence in the decision made	-1.300	.534	5.930	1	.015	.273
neutral design: “confirm” left, “abort” right	1.790	.890	4.049	1	.044	5.992
neutral design: “abort” left, “confirm” right	1.061	.965	1.209	1	.272	2.888
opinionated design prompting to abort	1.873	.885	4.478	1	.034	6.511
2FA frequency up to once a day	-1.142	.761	2.254	1	.133	.319
2FA frequency up to once a week or less	1.667	.874	3.640	1	.056	5.295
Constant	7.155	2.729	6.874	1	.009	1281.132

Two design dummy variables, namely the one for the opinionated design prompting users to abort and the neutral one with “confirm” on the left and “abort” on the right, were also significant, but had a positive impact on the likelihood to proceed. This means, participants were more likely to proceed with the login when presented with either of these two designs compared to when presented with the opinionated design prompting users to proceed with the login (benchmark category). This tendency is also reflected in the frequencies of proceed and abort distributed over the designs (see Table 4.18). Participants were by far most likely to abort the login when presented with the opinionated design prompting them to proceed.

Table 4.18: Cross Tabulation of the Warning Designs and the Likelihood to Proceed

	Abort	Proceed	Total
opinionated design prompting to proceed	19	7	26
opinionated design prompting to abort	15	12	27
neutral design: “abort” left, “confirm” right	13	6	19
neutral design: “confirm” left, “abort” right	11	11	22
Total	58	36	94

4.6 Reasons Stated for Decision Made

Following the interactive task, study participants were given the opportunity to comment on how they had come to their decision to abort or proceed with the login, respectively. Answering this open question was

optional.

Seventeen participants did not answer the open question or gave a non-answer (such as “?” or “no reason”), four answers could not be clearly classified.

Seven participants gave as their reason that the scenario was not real (such as “It is a survey”), that there were no consequences (e.g. “it doesn’t matter...”), or that the experiment design was not realistic (e.g. “no real password input”). All of these participants proceeded with the login.

Fifteen participants explicitly stated the warning message had informed them that the website was fake (e.g. “The warning reminds me the website is fake”), or that they had followed the instruction of the warning message (e.g. “warning messages remind me not to log in”). All of these participants aborted the login.

Twenty-six participants stated the website being fake as the reason for their decision, without specifying how they had come to that conclusion (e.g. “not real website”). Of these participants, 23 aborted the login while three proceeded.

Five participants stated warning as their reason, without elaborating on details (e.g. “because of the warning”). Four of these aborted the login, while one proceeded. In this, the participant who proceeded as well as one of those who aborted, had been shown the opinionated design prompting users to proceed.

Ten participants referred to a risk or stated safety concerns, without specifying what their perception was based on (such as “login risk” or “for the safety”). Of these participants, nine aborted the login and one proceeded.

Four participants gave task-related reasons, such as that the authentication processes had been initiated by themselves (e.g. “Because of the prompt to confirm during a process I have initiated myself”) or that they wanted to achieve the task goal (e.g. “want to login”). All of these participants proceeded with the login.

Four participants gave answers suggesting they felt generally unconcerned, without giving a specific reason (such as “Should be fine” or “Didn’t think much”). All of these participants proceeded with the login.

Two participants gave trust as the reason for their decision. In this, one did not specify the source of this trust, while the other stated “becauese the website is send of my friend,I believe she won’t cheat me. [sic]”. Both participants proceeded with the login.

5 Discussion

According to the ENISA Threat Landscape 2023 report, phishing is the most prevalent form of social engineering as well as profitable for attackers (ENISA, 2023), suggesting that often enough it is successful in "exploit[ing] human error or human behaviour" (ENISA, 2023, p. 7). In this, attackers have also already succeeded in circumventing second-factor authentication (ENISA, 2023). In order to address this situation, it is arguably necessary to understand both what can make users susceptible as well as what approaches could mitigate this susceptibility. The present thesis aimed to contribute to this understanding specifically in the context of 2FA (which to the best of our knowledge had not been researched before) by investigating both the potential impact of psychological and individual factors on susceptibility as well as an UX-based mitigation approach, as reflected in the research questions.

In the following, first the results of the present study will be discussed in regards to the research questions, and then the methods used to investigate these questions.

5.1 Results Discussion

In this section, the results presented in the previous chapter are discussed in relation to the research questions of this thesis. These interpretations have to be considered with caution, however, due to the insufficient sample size and use of convenience sampling in this study.

5.1.1 Research Question 1: Psychological Factors

The first research question of this thesis investigated the potential effect of psychological factors on the likelihood to proceed:

RQ1: In the context of automated URL verification and verification-based warnings in a 2FA app, how can psychological factors, such as stress, attention, elaboration and involvement, influence the user's decision to abort or proceed with logging in to a suspicious website?

Of the psychological factors contained in the theoretical model (stress, attention, elaboration, and involvement), involvement was found to have a significant negative effect on the likelihood to proceed based on the collected sample. In addition, confidence in the decision made was identified as a potential predictor, also showing a significant negative effect. Further details as well as considerations regarding the factors that did not show statistical significance in the complete regression model are discussed in the following.

Stress

Neither in the logistic regression model nor the univariate significance test evidence could be found that the stress experienced during the task had an impact on the likelihood to proceed. However, the theoretical framework of this thesis posited that it is the high stress levels that can increase phishing susceptibility, and none of the participants in this study experienced high stress levels . Therefore, the collected sample might not have been sufficient to investigate and test for the assumed effect of stress.

Attention

In the complete logistic regression model, i.e. controlling for the other factors of the model, no evidence was found that attention had a significant impact on the likelihood to proceed. However, in the univariate test, attention had a significant negative effect. The direction of this effect is in accordance with the assumptions of the theoretical model, namely that attention can reduce phishing susceptibility.

In addition, attention significantly correlated with involvement and confidence in the decision made, both of which were indicated as significant predictors in the regression analysis. According to Vishwanath et al.

(2011) involvement has a positive effect on attention. Similarly, in the area of consumer behaviour – where both the definition and measurement of involvement used in this thesis stem from – Celsi and Olson (1988) (in the context of ads) and Behe et al. (2015) (in the context of product choice) found that highly involved subjects paid more attention. Therefore, some of the statistical effect of attention might be subsumed in the observed effect of involvement.

Furthermore, the mitigation approach investigated in this study, i.e. the implementation of the 2FA prototype, was deliberately designed to reduce phishing susceptibility and specifically also to reduce the amount and degree of attention required from users. In particular, due to the automated URL verification no attention to indicators was required at the e-mail or website stage; the use of an interrupting warning that blocked the user's workflow meant that the stimulus was hard to miss, i.e. required little attention to notice, and the duration of attention required was short, since the user was alerted right in the moment; the decluttered interface reducing distractions meant that any attention given was given to relevant information; and the use of visuals, such as signal colours, frames, and use of icon, made the warning salient, requiring less attention to notice (see section 2.2 for an elaboration on these effects). All of these measures on their own – and arguably even more so in combination – can thus contribute to reducing the attention required from users, and thereby can arguably be expected to decrease the significance of attention as a predictor of phishing susceptibility. Therefore, the lack of significance of attention found in this study might indicate that the treatment design was successful in reducing the attention required and contributed towards mitigating some of this aspect of phishing susceptibility.

Elaboration

Because of issues with the scale validity (see section 4.2), the effect of elaboration on the likelihood to proceed could not be investigated in this study. However, according to prior research, involvement can increase the likelihood (Vishwanath et al., 2011) and degree (Behe et al., 2015) of elaboration. Since the effect of elaboration itself could not be controlled for in this study, some of it might therefore have been subsumed in the statistically significant effect found for involvement.

Furthermore, a total of 41 participants either explicitly cited the warning message or stated the website being fake as the basis for their decision, meaning they must have read and evaluated – i.e. elaborated on – the message and/or other indicators (such as elements of the e-mail or website). Of these participants the vast majority aborted the login, which can be seen to be in accordance with the assumption of the theoretical framework that elaboration can reduce the likelihood to proceed.

Involvement

In accordance with the theoretical model, the results of the logistic regression analysis showed a significant negative effect of involvement on the likelihood to proceed. In addition, the univariate tests suggested that involvement was both significant and the strongest individual predictor in the collected sample, with a $Nagelkerke\ R^2$ of $.343$ for the one-variable model, roughly twice as high as the $Nagelkerke\ R^2$ for any of the other significant univariate models and more than half of the $Nagelkerke\ R^2$ for the complete regression model with $.557$. This can be seen to indicate a high predictive strength of involvement. However, as mentioned above, some of this strength might be explained by the fact that the effect of elaboration could not be controlled for in this study due to scale validation issues.

Confidence in the Decision Made

For the collected sample, confidence in the decision made had a significant negative effect on the likelihood to proceed.

Yeung and Summerfield (2012) state that people often "report levels of confidence in their decisions that correlate with objective performance" (Yeung and Summerfield, 2012, p. 1310). However, the respective levels do not always match. Rather, as Fleming and Lau (2014) argue, the degree to which confidence and accuracy correlate can be seen to indicate "how good one is at distinguishing between one's own correct and incorrect judgments" (metacognitive sensitivity) (Fleming and Lau, 2014, p. 1). There are many factors that can lead to over- or underconfidence, such as trait or situational factors, the hard-easy effect (in which difficult tasks tend to lead to overconfidence and easy tasks to "calibrated or even underconfident evaluations"),

or the 'unskilled-and-unaware-of-it' phenomenon (in which overconfidence increases with incompetence) (Boldt et al., 2017, pp. 1521–1522). Furthermore, Zizlsperger et al. (2012) found in their study that attention increased confidence significantly more than it did accuracy (Zizlsperger et al., 2012, p. 1), suggesting that the two can be affected by other factors differently and thus diverge.

Prior research regarding a potential confidence-accuracy association (like the one the results of this study can be seen to indicate) has yielded mixed results, with sometimes contradictory findings (Perfect, 2002; Sauerland and Sporer, 2007; A. M. Smith and Leach, 2019; Volz et al., 2022). In this, results differed not only between contexts but also between studies in the same context, arguably indicating a potential significance of differences in context and study conditions. To the best of our knowledge, no research has previously been conducted in a context related to that of this thesis (such as authentication or warning-related security decisions). If and how any of the findings can or should be applied to the context of this study is therefore arguably difficult to assess. Therefore, instead of explaining the observed association between confidence and accuracy in this study based on prior findings regarding that relationship itself, in the following an explanation based on potential factors that might have influenced both confidence and accuracy in this study will be attempted.

Desender et al. (2018) suggest that low confidence increases the likelihood that subjects will seek information before making their decision (Desender et al., 2018, p. 1). In the context of this study, this arguably meant they were more likely to pay attention to and elaborate on the warning message. Furthermore, information seeking is also a behaviour that is associated with involvement (Behe et al., 2015, p. 11). All three of these (attention, elaboration, and involvement) were factors posited to reduce phishing susceptibility, i.e. increase decision accuracy, with involvement also found significant in the regression analysis. In addition, the message was designed to provide all information necessary to make an informed decision and reduce choice difficulty (regardless of which of the four designs was presented to the user); and the warning as a whole was based on automated URL verification intended to increase perceived relevance (involvement) and credibility. Together, these measures might have increased both confidence and accuracy, thereby contributing towards the respective mitigation of underconfidence (increase confidence to match accuracy) and overconfidence (increase accuracy to match confidence), overall leading to the observed confidence-accuracy association.

If indeed the design of the 2FA implementation (warning design and message, automation) could increase accuracy, this would mean that the UX design investigated in this thesis was to some degree effective in mitigating phishing susceptibility. This, in turn, could then be seen to indicate that a UX-based mitigation approach at the 2FA stage and taking susceptibility factors into consideration in its design (as was the case in this thesis) might be a promising approach towards phishing mitigation.

In addition, a mitigation approach which might increase both confidence and accuracy would arguably be particularly desirable, because post-decision confidence can guide a subject's choice when they are presented with a similar decision again (Folke et al., 2017, p. 1). Specifically, Folke et al. (2017) found in their studies that post-decision confidence "reliably predicted a change of mind when the same choice was presented again" (Folke et al., 2017, p. 5), with those subjects low on confidence less likely to repeat their previous choice (Folke et al., 2017, p. 1). Therefore, a mitigation approach, which might increase both confidence and accuracy would not only be likely to help users make the right decision now, but also again in the future. This is arguably of particular significance in the context of 2FA, since 2FA is by its nature intended to be used repeatedly.

In the present study, confidence in the decision made was not measured only with a direct question, in which participants rated their confidence level as a single 5-point item. Rather, a new extended measure was introduced, which included in addition further questions that reflected some latent assumptions about confidence and its effects (see section 3.3.2). For example, participants were asked to rate how likely they were to repeat their decision, reflecting the assumed relationship discussed in the previous paragraph. The results of the exploratory factor analysis (see section 4.2) can be seen to suggest that the included assumptions were at least to some degree valid for the collected sample. Should this validity not be generalizable, the new measure might be useful to differentiate or ensure a certain type of confidence, namely one that exhibits the assumed relationships. In studies building on these latent assumptions, the scale might then yield more relevant and accurate results, because it would indicate high confidence only where these assumptions are at least to some degree met.

5.1.2 Research Question 2: 2FA Frequency

The second research question of this thesis investigated the potential effect of the frequency of 2FA use on the likelihood to proceed:

RQ2: In the context of automated URL verification and verification-based warnings in a 2FA app, how can 2FA frequency influence the user's decision to abort or proceed with logging in to a suspicious website?

The results of the logistic regression for the complete model showed no evidence that 2FA frequency had an impact on the likelihood to proceed. However, according to the theoretical framework underlying this thesis, the expected relevance and effect of the factor was assumed to be due to overwhelm because of a too high load as well as due to habituation effects (see section 2.1.3). In this study, participants were asked about their usual frequency of 2FA use in their daily lives, meaning that the answer did not necessarily apply to the specific day of their participation. Furthermore, even the highest category of the variable with two or more uses a day might not create an equivalent load on the user as a high e-mail load would, which was the factor of the IIPM 2FA frequency was adapted from. Habituation (if any), on the other hand, can arguably not have been specific to the 2FA implementation of this study, since participants were using it for the first (and only) time. Furthermore, since – to the best of our knowledge – no existing 2FA implementations use verification-based warnings, the workflow and design of the implementation used in this study arguably differed from what participants might have experienced with other apps. Lastly, while the warning was designed based on existing guidelines, it combined several different practices and was not a replica of a pre-existing design. In their study on habituation to security warning design, Anderson et al. (2016) found simply rotating through different variations of designs to be effective in mitigating habituation, which would suggest that habituation effects in this study could be expected to be low. In conclusion, the 2FA load and habituation effects might have been too low across the collected sample to allow to test for the effect on the likelihood to proceed.

Furthermore, while according to Vishwanath et al. (2011) a combination of habituation and high e-mail load increases the likelihood of automatic, subconscious responses, involvement reduces this likelihood. Seeing that the median involvement level among participants in this study was rather high, the results could therefore be seen to be in support of the assumption that involvement can reduce susceptibility resulting from habituation effects also in the 2FA context.

Looking at the results of the univariate tests, on the other hand, both 2FA dummy variables were significant. 2FA usage of up to once a week or less had a significant positive effect on the likelihood to proceed, meaning participants from this group were more likely to continue with the login than those who commonly use 2FA several times a day (benchmark category). On the other hand, 2FA usage between once a week and once a day had a significant negative effect on the likelihood to proceed, meaning participants from this group were more likely to abort the login than those with higher usage frequency (benchmark category). This can be seen to suggest that both too low and too high frequencies might increase susceptibility, and that there might be a point of balance in between, where susceptibility is lowest.

While intuitive comprehension requires no conscious mental effort and involves no uncertainty (Swindler, 2021, p. 90), "[i]ntuition is built through lived experiences" and works based on the recognition of patterns (Swindler, 2021, p. 90). Users with a very low 2FA frequency, however, are arguably likely to have low levels of experience and familiarity. Therefore, their decision-making process is likely to require more elaboration, thereby increasing susceptibility. In addition, they might feel more uncertainty, which could negatively affect their confidence in the decision made, thereby again increasing susceptibility. Furthermore, a very low 2FA frequency could be indicative of an overall low involvement with security concerns (some might even only use 2FA when they are forced to do so, such as in online banking), also increasing susceptibility.

Since participants interacted with the 2FA implementation used in this study only once, this use itself could arguably neither result in an overwhelming load nor in habituation. This would suggest that the potential positive effect of too high 2FA frequency on susceptibility would be due to effects of frequent use that occur and likely accumulate across different 2FA implementations and possibly software in general; i.e., while 2FA implementations differ, users could get desensitized to 2FA use per se; and while verification-based warnings are not commonly used in 2FA implementations, users might get desensitized to such warnings in general. This could be seen to contradict Anderson et al. (2016) finding that a measure as simple as rotating through different variations of design can effectively mitigate habituation, because the exact warning designs arguably differ between different products, constituting different variations.

5.1.3 Research Question 3: Opinionated Design

The third research question of this thesis investigated the potential effect of opinionated design on the likelihood to proceed:

RQ3: In the context of automated URL verification and verification-based warnings in a 2FA app, how can opinionated design influence the user's decision to abort or proceed with logging in to a suspicious website?

The results of this study can be seen to contradict the theoretical assumptions about the effect of opinionated design. In particular, participants who were shown the opinionated design prompting them to proceed were by far more likely to abort than to proceed as well as more likely to abort than participants presented with any of the other designs. This effect was statistically significant not only in the univariate tests but also in the logistic regression for the complete model, meaning even when controlling for the other factors. Furthermore, the opinionated design prompting users to abort resulted in only slightly more abort than proceed choices, and had a lower percentage of aborts than the neutral design with "abort" on the left. This is arguably in stark contrast to the "substantially improved adherence" found by Felt et al. (Felt et al., 2015, p. 2900).

Depending on the purpose they are used for, opinionated designs can arguably be classified as a (combination of) dark pattern(s), because "[t]hey make it possible but asymmetrically difficult for a user to act" (Luguri and Strahilevitz, 2021, p. 52) against the promoted choice, and the distinction between dark patterns and nudging is commonly only based on whether this promoted choice is in the interest of the user or alternatively can be justified by social welfare (nudging), or not (dark pattern) (Luguri and Strahilevitz, 2021, p. 52). The techniques used, on the other hand, are similar in both cases (Luguri and Strahilevitz, 2021, p. 52). In the terms used by Luguri and Strahilevitz (2021), opinionated design can thus be seen to combine the dark patterns "hidden information", which utilizes "smaller print in a less visually prominent location" (Luguri and Strahilevitz, 2021, p. 47), and a type of "aesthetic manipulation", in which one option is less legible than the other (Luguri and Strahilevitz, 2021, p. 51), leading to "[i]mportant information [being] visually obscured" (Luguri and Strahilevitz, 2021, p. 53). Luguri and Strahilevitz (2021) found "hidden information", in particular, to be one of "the most effective dark pattern strategies" (Luguri and Strahilevitz, 2021, p. 47), which "doubled acceptance rates" in their experiment (Luguri and Strahilevitz, 2021, p. 75). Their findings were therefore arguably also in support of the assumed effect and effectiveness of opinionated design posited in the theoretical framework of this thesis, but thereby in contrast to the results of this study.

While both Felt et al. (2015) and Luguri and Strahilevitz (2021) thus argue for the general effectiveness of such designs, Löfgren and Nordblom (2020), on the other hand, posit in their theoretical framework that "[i]mportant choices are rarely nudgeable, as they often are made attentively" (Löfgren and Nordblom, 2020, p. 9). This would suggest that involvement (importance) and attention might mitigate the effect of opinionated designs. While the involvement measured in this study related to security concerns and could explain the cases in which participants aborted despite being prompted to proceed, users can arguably also be involved with other matters (such as task completion), which if they take precedence could then possibly explain cases in which participants proceeded with the login despite being prompted to abort. Such an involvement with matters external to the 2FA prompt could arguably play an important role in phishing scenarios, because targets might be so preoccupied with addressing the perceived threat that their involvement with the phishing narrative could take precedence, rendering the use of opinionated designs – according to Löfgren and Nordblom (2020) – likely ineffective.

In the theoretical framework of this thesis, the assumed effect of opinionated design was explained by its exploitation of biases, and it was argued that high stress levels increase susceptibility to such exploits. However, none of the study participants experienced high stress levels, reducing this susceptibility. Thereby, the low stress levels in the study might have further reduced the statistical significance of the opinionated design.

Furthermore, as stated before, the mitigation approach investigated in this study, i.e. the implementation of the 2FA prototype, was deliberately designed to reduce phishing susceptibility. In particular, the measures aimed to reduce attention and elaboration required, increase involvement and mitigate the effects of stress. Automated URL verification was combined with warning designs comprising a multitude of principles based on established guidelines and findings of prior research (see section 2.2), but only differing in their opinionatedness. Considering the complexity of this treatment, other aspects than the opinionatedness of the

respective design might have had a stronger effect. The statements made by the participants showed that roughly every sixth participant referred to the content of the warning message as the basis of their decision, clearly indicating that they read and evaluated it. If further participants who came to the conclusion the website was fake, were to be assumed to have taken this information from the message also, this proportion could rise to more than 40%. In addition, some participants stated “warning” as the basis for their decision, suggesting that the clear recognition of the warning as a warning might already have had a significant effect on their decision-making. Overall, this could be seen to indicate that the designs – regardless of opinionatedness – were to some degree successful in their intended effects (see also sections on attention, elaboration and confidence in the decision made in the discussion of RQ1 above).

5.2 Method Discussion

Participants for this study were recruited via convenience sampling, meaning that results can likely not be generalized. Attempting to mitigate this to some degree, recruitment was not only conducted via our extended network but also via Reddit (r/samplesize). Since participation was anonymous and the same invitation link was used for all channels, there are no accurate numbers on how many respondents were reached through which channel. However, due to issues with the platform (insufficient “Karma” to post), the recruitment via Reddit started late. Therefore, a rough estimate based on this delayed start can be made, indicating that a maximum of 15% of respondents could have been recruited via Reddit, if any. Not all of these respondents finished the study. This low recruitment rate via Reddit resulted not only in a small overall sample size, but also meant that the aim to mitigate to some degree biases from the network-based sampling was not reached. Due to the very limited time and lack of budget for this thesis, however, the chosen recruitment methods were arguably the only feasible ones for this study.

Although the stress inducement method used in this study was based on prior research (see section 3.3.3), it did not have a significant effect on the stress levels of participants. Since Almazrouei et al. (2023) did not publish the specific questions they used in their study, questions taken from a timed concentration test by Dr. Satow were used instead, since they matched the category of mathematical questions described by Almazrouei et al. (2023). However, while this scientific, standardized test was designed to measure performance, speed, endurance and accuracy (Psychomeda, n.d.), it was not originally intended to induce stress and might not have been a suitable replacement after all. Another reason for the lack of stress induced, might have been that the difficulty and overall length of the test was reduced for this study – particularly after the feedback of the pilot testers – in order to avoid losing respondents due to frustration or lack of patience (see sections 3.3.3 and 3.3.5). As a result, the final version used might have been too easy and/or short to elicit stress. However, the aim of not losing participants was arguably met as shown by the almost even frequencies of participants who had been exposed to the stress inducement and those who had not in the overall sample. Furthermore, according to the SoSci statistics of the survey, only five respondents explicitly quit on one of the stress inducement pages. Lastly, while the stress inducement did not have a significant effect, both the method and result of the inducement met the ethical considerations underlying the method choice. In this, a non-significant effect is arguably still preferable to an unethical inducement.

In an online setting the threshold for quitting can arguably be assumed to be lower than in a lab setting, where respondents have guidance and are observed. Hence, in a lab setting the stress inducement could possibly be made longer and more difficult without risking to lose participants, and thereby might become more effective. Participants’ awareness of being observed might then even add to this effectiveness, since this might contribute to the perceived socio-evaluative threats, which were one of the stress inducing factors posited by Almazrouei et al. (2023).

Furthermore, since the study was conducted online, measurements of the variables relied on self-reporting. A lab setting could have allowed for other and potentially more accurate measurement methods, such as direct observation, monitoring vitals, measuring skin resistance or eye tracking. For example, a person might self-report higher involvement than they actually exhibited, similarly to the so-called privacy paradox, a phenomenon in which people tend to “rarely make an effort to protect [their personal] data actively and often even give it away voluntarily” despite stating that privacy is important to them (Gerber et al., 2018, p. 226).

Lastly, a lab study would have given full control over the setup and devices used, which could have prevented technical issues, which in some cases led to respondents being unable to finish the survey.

While the above arguments outlined potential advantages of a lab study, conducting the study online arguably provided higher ecological validity, because the resulting setting was closer to real-life conditions. Participants used their own devices in their familiar environment and without the potential influence from being observed by researchers. Any potential influence of other people who might have been with the participant (which was not measured in this study), on the other hand, could arguably equally occur in a real scenario and can thereby seen to also contribute to the ecological validity. Furthermore, a lab study would have required both the researchers and the participants to be present in person. This would have limited the population to recruit from to those within traveling distance of Jönköping, increased the effort necessary to participate, and severely limited the amount of sessions that could have been conducted in the very limited time frame of this thesis, overall significantly reducing the resulting sample size and possibly its diversity even further.

Unlike in the study setup of Sun et al. (2021), no prior training had been given to participants regarding the use of the 2FA implementation. Sosci statistics, however, show that none of the respondents aborted the interactive task, meaning all who started it, were able to complete it. While it is unknown, how many of these managed to correctly take a photo including the browser's address bar – because in the prototype implemented for this study the URL verification was faked and the photos never processed – this arguably shows that the self-efficacy of all of these respondents was high enough for them to try and perform the task.

This study was designed so that participants were aware that the scenario was fictional and they were not exposed to any real threat. This might have contributed to the overall rather low stress levels, which in turn might have led to the collected sample not being suitable to test for the effects of high stress (see discussion of stress in section 5.1.1). Furthermore, the overt fictionality might to some degree have affected the decisions made by some participants. In particular, seven participants stated the fictionality (study, no consequences, not real) as the basis for the decision they made. However, many other participants showed high engagement with the scenario, such as the ones elaborating the warning message or other indicators to conclude the website was fake inside the scenario (these alone constituted more than 40% of all participants). In addition, the overtly fictional setup had been deliberately chosen out of ethical considerations. Making respondents believe they were in a real scenario would have been ethically questionable and could have caused them severe distress, as is commonly experienced by targets in real phishing situations.

6 Conclusions and Further Research

6.1 Conclusions

Second-factor authentication is a technology that was introduced to make systems more secure (Siadati et al., 2016, p. 65). However, phishing attackers have already succeeded in circumventing existing implementations, e.g. by using the adversary-in-the-middle technique (AitM) (ENISA, 2023, p. 78) (see section 1.1). In this, Phishing as a Service can enable even attackers with little expertise of their own (ENISA, 2023), arguably making such attacks increasingly likely.

If in such an attack the user is tricked into visiting the fake website set up by the attackers, the 2FA service is the only contact they still have to a legitimate provider, making it the last potential point for intervention (see section 1.1). Thus, mitigating susceptibility at this stage can arguably play an important role in protecting users from substantial harm.

This thesis therefore aimed to contribute to the understanding of what can make users susceptible as well as what approaches could mitigate this susceptibility in the context of 2FA, which – to the best of our knowledge – had not been researched before. In this, the thesis investigated specifically a UX-based mitigation approach for an AitM scenario, combining automated URL verification and verification-based warnings.

While due to the too small sample size and use of convenience sampling in this thesis, the findings need to be validated by further research, they might serve both researchers and practitioners as perhaps an indication, which factors might be of relevance and what mitigation techniques worth considering.

6.1.1 Practical Implications

The purpose of this thesis was to contribute to the research into phishing susceptibility and mitigation, and thereby also inform practical measures taken to protect users. In this, the focus was on the UX design of 2FA apps and its potential towards mitigation.

The results of the present study suggest that confidence in the decision made might be a factor that can reduce phishing susceptibility. 2FA implementations should therefore aim to increase users' confidence in their decisions. A potential explanation for the confidence-accuracy association observed in this study could be that information seeking behaviour was effective. This could suggest that supporting information seeking might be an approach worth considering.

The results of this study further indicate that involvement might reduce phishing susceptibility at the 2FA stage. Therefore, 2FA implementations should aim to facilitate user involvement. In this, situation-aware approaches – like the verification-based warning investigated in this thesis – might increase perceived relevance, since they are suspicion-based and related directly to the situation at hand. While such implementations require more development effort, they are arguably also likely to reduce how often a user is presented with the respective warning or a warning at all. This might mitigate to some degree habituation effects, which can increase susceptibility (Vishwanath et al., 2011).

In addition to a potential direct reduction of susceptibility, involvement can furthermore also increase the likelihood of information seeking behaviour (Behe et al., 2015). Therefore, measures facilitating user involvement are also likely to work well in concert with and contribute towards the aforementioned approach of supporting information seeking aiming to increase confidence in the decision made.

According to the results of this study, relying on opinionated design to influence users' decision-making might not be an effective mitigation strategy in 2FA apps. Furthermore, because it utilizes techniques also found in dark patterns, it might be perceived negatively by some users. Therefore, other approaches should be explored first, and if the use of opinionated design is still considered, tests are advisable to ensure that it does make a significant difference in the specific implementation context.

In regards both to the facilitation of involvement as well as to the use of opinionated design, it is also important to consider that in a phishing scenario the user might be highly involved with the phishing narrative and preoccupied with addressing the perceived threat (or gaining the exciting reward). This might take

precedence over or even eclipse their involvement with security concerns, even if they are usually safety-cautious. Thus, on the one hand, the 2FA app would be competing for the user's involvement, and, on the other hand, the user might be likely to go against the promoted choice in an opinionated design.

The results of this study can, furthermore, be seen to suggest that approaches using automation and design principles that can reduce the attention required from the user might contribute towards the mitigation of phishing susceptibility, especially when combining multiple measures.

According to the results of this study, both too infrequent and too frequent use of 2FA might increase phishing susceptibility, with a point of balance in between, where susceptibility is lowest. When implementing a 2FA solution, it is important to consider that the effects of too high frequency, such as perhaps habituation, might accumulate across different 2FA implementations or even different software products (as the results of the present study can be seen to suggest), thus possibly necessitating a concerted effort across the industry when developing mitigation approaches.

6.1.2 Scientific Implications

To the best of our knowledge, the present study was the first investigating phishing susceptibility and mitigation in the 2FA context, thereby extending the body of knowledge in the broader field of social engineering research. In this, the theoretical framework of this thesis combined findings from prior research from different fields and contexts to develop a new integrated model of susceptibility and a mitigation approach specific to the 2FA context. This study thus investigated both the posited model and approach as well as the generalizability and applicability of prior findings to the 2FA context.

The results of this study can be seen to support the assumption of the theoretical framework that involvement can have a negative effect on phishing susceptibility. This is in accordance with the findings of Vishwanath et al. (2011) in their study on susceptibility at the e-mail stage of phishing attacks. The present study can thus be seen to support their finding and suggest that the mitigating potential of involvement might also apply to the context of 2FA.

Confidence in the decision made was found to have a significant negative effect on the likelihood to proceed in the collected sample. This could be seen to suggest that said confidence might be a significant predictor of phishing susceptibility, which to the best of our knowledge has not previously been posited as a potential factor in prior research related to this topic. Furthermore, the observed effect can be seen to suggest a potential confidence-accuracy association. Further research into the indicated effect, therefore, would contribute to the ongoing scientific debate on this topic, which so far has yielded mixed results, and – to the best of our knowledge – also add a new context to that research area.

The Message Elaboration Measure (Reynolds, 1997) was used to measure elaboration in this study. However, the results of the principle component analysis indicated issues with scale validity for the collected sample (see section 4.2). This could suggest that the scale might not be applicable to the 2FA context.

A significant number of participants in this study went against the promoted choice of the respective opinionated designs. In particular, participants presented with the opinionated design promoting to proceed were by far more likely to abort than to proceed as well as more likely to abort than participants presented with any of the other designs. This would suggest that the effect of opinionated designs on user decision-making might not be as general and as definite as suggested by prior research, such as Felt et al. (2015) or Luguri and Strahilevitz (2021). Rather, the effectiveness of such designs could be influenced by involvement, as posited by Löfgren and Nordblom (2020), or other factors yet to be investigated.

6.2 Further Research

This study investigated one very specific and overtly fictional scenario, utilized convenience sampling and yielded a sample size below the minimum size required for the analysis method applied. Further research is therefore needed to validate as well as further investigate the potential effects suggested by the results of this study, especially in regards to new findings, such as the possible influence of confidence in the decision made on the likelihood to proceed, or the possible point of balance between a too low and too high 2FA frequency.

In the present study a new measure for confidence in the decision made was introduced, which showed good results in the exploratory factor analysis for the collected sample. Further research is required to further validate the scale and assess whether it is correctly measuring the latent construct.

The model introduced in this study was based on prior research but had as a whole not been investigated before. In addition, the study also introduced a complex treatment, which again was based on prior research as well as established guidelines but had as a whole not been investigated before. Therefore, the lack of significance found for some of the susceptibility factors could to a degree also have resulted from interrelations not accounted for in the model or be an indication that the treatment had indeed a mitigating effect in regards to the respective influence of these factors. Similarly, the lack of significance found for the opinionated design might have to some degree also resulted from the combined effectiveness of the remaining aspects of the overall treatment (automation and a multitude of design principles), which might have been strong enough that adding opinionated design did not make a significant difference. However, the respective effects of these measures could not be analysed in this study, since they were applied only in combination. Future research could investigate both the interplay of the model factors and treatment aspects (and different combinations of them), as well as their respective individual contributions, thus further contributing to the understanding of phishing susceptibility and helping to inform mitigation approaches.

7 References

Akhawe, D., & Felt, A. P. (2013). Alice in warningland: A Large-Scale field study of browser security warning effectiveness. *22nd USENIX Security Symposium (USENIX Security 13)*, 257–272.

Almazrouei, M. A., Morgan, R. M., & Dror, I. E. (2023). A method to induce stress in human subjects in online research environments. *Behavior research methods*, *55*(5), 2575–2582. https://doi.org/10.3758/s13428-022-01915-3

Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, *82*, 69–82. https://doi.org/10.1016/j.ijhcs.2015.05.005

Anderson, B. B., Jenkins, J. L., Vance, A., Kirwan, C. B., & Eargle, D. (2016). Your memory is working against you: How eye tracking and memory explain habituation to security warnings. *Decision Support Systems*, *92*, 3–13. https://doi.org/10.1016/j.dss.2016.09.010

Bauer, L., Bravo-Lillo, C., Cranor, L. F., & Fragkaki, E. (2013). Warning design guidelines. https://api.semanticscholar.org/CorpusID:4532891

Behe, B. K., Bae, M., Huddleston, P. T., & Sage, L. (2015). The effect of involvement on visual attention and product choice. *Journal of Retailing and Consumer Services*, *24*, 10–21. https://doi.org/10.1016/j.jretconser.2015.01.002

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology*, *43*(8), 1520–1531. https://doi.org/10.1037/xhp0000404

Bravo-Lillo, C., Komanduri, S., Cranor, L. F., Reeder, R. W., Sleeper, M., Downs, J., & Schechter, S. (2013). Your attention please: Designing security-decision uis to make genuine risks harder to ignore. *Proceedings of the Ninth Symposium on Usable Privacy and Security*. https://doi.org/10.1145/2501604.2501610

Celsi, R. L., & Olson, J. C. (1988). The role of involvement in attention and comprehension processes. *Journal of Consumer Research*, *15*(2), 210–224.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge. https://doi.org/10.4324/9780203029053

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), Article 7. https://doi.org/10.7275/jyj1-4868

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 581–590. https://doi.org/10.1145/1124772.1124861

European Union Agency for Cybersecurity. (2023). *Enisa threat landscape 2023 – july 2022 to june 2023* (I. Lella, E. Tsekmezoglou, M. Theocharidou, E. Magonara, A. Malatras, R. S. Naydenov, & C. Ciobanu, Eds.). https://doi.org/10.2824/782573

Felt, A. P., Ainslie, A., Reeder, R. W., Consolvo, S., Thyagaraja, S., Bettes, A., Harris, H., & Grimes, J. (2015). Improving ssl warnings: Comprehension and adherence. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2893–2902. https://doi.org/10.1145/2702123.2702442

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, Article 443. https://doi.org/10.3389/fnhum.2014.00443

Folke, T., Jacobsen, C., Fleming, S. M., & Martino, B. D. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), Article 0002. https://doi.org/10.1038/s41562-016-0002

Gerber, N., Gerber, P., & Volkamer, M. (2018). Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, *77*, 226–261. https://doi.org/10.1016/j.cose.2018.04.002

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. *2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, 82–88.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (Eighth Edition). Cengage.

International Organization for Standardization. (2016). *ISO 3864-2:2016: Graphical symbols - safety colours and safety signs - part 2: Design principles for product safety labels* (Standard). International Organization for Standardization.

Johansson, L.-O., & Ask, K. (2023). *Utlåtande angående psykologin bakom riskfyllda och irrationella ekonomiska beslut* (statement). Psykologiska Institutionen, Göteborgs Universitet.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching johnny not to fall for phish. *ACM Trans. Internet Technol.*, *10*(2). https://doi.org/10.1145/1754393.1754396

Leiner, D. J. (2016). Our research's breadth lives on convenience samples. a case study of the online respondent pool "sosci panel". *SCM Studies in Communication Media*, *5*(4), 367–396. https://doi.org/10.5771/2192-4007-2016-4-367

Löfgren, Å., & Nordblom, K. (2020). A theoretical framework of decision making explaining the mechanisms of nudging. *Journal of Economic Behavior & Organization*, *174*, 1–12. https://doi.org/10.1016/j.jebo.2020.03.021

Luguri, J., & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. *Journal of Legal Analysis*, *13*(1), 43–109. https://doi.org/10.1093/jla/laaa006

Nebeker, C., Simon, G., Kalichman, M., Talavera, A., Booen, E., & Lopez-Arenas, A. (n.d.). *Basic research concepts (brc)*. https://ori.hhs.gov/basic-research-concepts-brc

Norman, D. (2013). *The design of everyday things*. Basic Books.

Perfect, T. J. (2002). When does eyewitness confidence predict performance? In *Applied metacognition* (pp. 95–120). Cambridge University Press. https://doi.org/10.1017/CBO9780511489976.006

Psychomeda. (n.d.). *Konzentrationstest*. Retrieved February 26, 2024, from https://www.psychomeda.de/online-tests/konzentrationstest.html

Reeder, R. W., Felt, A. P., Consolvo, S., Malkin, N., Thompson, C., & Egelman, S. (2018). An experience sampling study of user reactions to browser warnings in the field. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174086

Reynolds, R. A. (1997). A validation test of a message elaboration measure. *Communication Research Reports*, *14*(3), 269–278. https://doi.org/10.1080/08824099709388670

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law*, *13*(6), 611–625. https://doi.org/10.1080/10683160701264561

Shatz, I. (2017). Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, *35*(4), 537–549. https://doi.org/10.1177/0894439316650163

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. *Pro-*

ceedings of the SIGCHI Conference on Human Factors in Computing Systems, 373–382. https://doi.org/10.1145/1753326.1753383

Siadati, H., Nguyen, T., & Memon, N. (2016). Verification code forwarding attack (short paper). In F. Stajano, S. F. Mjølsnes, G. Jenkinson, & P. Thorsheim (Eds.), *Technology and practice of passwords* (pp. 65–71). Springer International Publishing. https://doi.org/10.1007/978-3-319-29938-9_5

Smith, A. M., & Leach, A.-M. (2019). Confidence can be used to discriminate between accurate and inaccurate lie decisions. *Perspectives on Psychological Science*, *14*(6), 1062–1071. https://doi.org/10.1177/1745691619863431

Smith, O. R., Melkevik, O., Samdal, O., Larsen, T. M., & Haug, E. (2017). Psychometric properties of the five-item version of the mindful awareness attention scale (maas) in norwegian adolescents. *Scandinavian Journal of Public Health*, *45*(4), 373–380. https://doi.org/10.1177/1403494817699321

Solomon, M. R. (2020). *Consumer behavior: Buying, having, and being* (Thirteenth edition, global edition). Pearson.

Sun, Y., Zhu, S., Zhao, Y., & Sun, P. (2021). Let your camera see for you: A novel two-factor authentication method against real-time phishing attacks. *arXiv.Org*. https://doi.org/10.48550/arxiv.2109.00132

Swindler, K. (2021). *Life and death design: What life-saving technology can teach everyday ux designers* [ebook]. Rosenfeld Media.

Tolbert, M. M., Hess, E. M., Nascimento, M. C., Lei, Y., & Shue, C. A. (2022). Exploring phone-based authentication vulnerabilities in single sign-on systems. In C. Alcaraz, L. Chen, S. Li, & P. Samarati (Eds.), *Information and communications security* (pp. 184–200). Springer International Publishing. https://doi.org/10.1007/978-3-031-15777-6_11

Vishwanath, A., Harrison, B., & Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*, *45*(8), 1146–1166. https://doi.org/10.1177/0093650215627483

Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, *51*(3), 576–586. https://doi.org/10.1016/j.dss.2011.03.002

Volz, S., Reinhard, M., & Müller, P. (2022). He confidence-accuracy relation – a comparison of metacognition measures in lie detection. *Applied Cognitive Psychology*, *36*(3), 673–684. https://doi.org/10.1002/acp.3953

Washo, A. H. (2021). An interdisciplinary view of social engineering: A call to action for research. *Computers in Human Behavior Reports*, *4*, 100126–. https://doi.org/10.1016/j.chbr.2021.100126

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Williams, E. J., Hinds, J., & Joinson, A. N. (2018). Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies*, *120*, 1–13. https://doi.org/10.1016/j.ijhcs.2018.06.004

Wogalter, M. S., Conzola, V. C., & Smith-Jackson, T. L. (2002). Research-based guidelines for warning design and evaluation. *Applied Ergonomics*, *33*(3), 219–230. https://doi.org/10.1016/S0003-6870(02)00009-1

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research*, *12*(3), 341–352.

Zizlsperger, L., Sauvigny, T., & Haarmeier, T. (2012). Selective attention increases choice certainty in human decision making. *PloS One*, *7*(7), Article e41136. https://doi.org/10.1371/journal.pone.0041136

Zsido, A. N., Teleki, S. A., Csokasi, K., Rozsa, S., & Bandi, S. A. (2020). Development of the short version of the spielberger state—trait anxiety inventory. *Psychiatry Research*, *291*, 113223. https://doi.org/10.1016/j.psychres.2020.113223

8 Appendices

8.1 Appendix A: Full Questionnaire

8.1.1 Initial Stress

Based on 5-item version of state component of State-Trait Anxiety Inventory (STAI) (Zsido et al., 2020)

5-point Likert items: '1 – Not at all, 2 – Slightly, 3 – Moderately, 4 – Very, 5 – Extremely'

How are you feeling right now at this moment?
1. I feel upset.
2. I feel frightened.
3. I feel nervous.
4. I am jittery.
5. I feel confused.

8.1.2 Open Question After the Interactive Task

For participants who chose to abort the login:
You chose to abort the login. How did you come to this decision?

For participants who chose to proceed the login:
You chose to confirm the login. How did you come to this decision?

8.1.3 Confidence in the Decision Made

5-point Likert items: '1 – Strongly disagree, 2 – Disagree, 3 – Undecided, 4 – Agree, 5 – Strongly agree'

Regarding that decision, how much do you agree with the following statements?
1. I was confident in my decision.
2. If I were to come across a similar situation again, I would make the same decision again.
3. I would make the same choice in a real-life situation.
4. I found it difficult to make the decision.
5. I did not think much about the decision.
6. I had an easy time making the decision.
7. I think it was the right decision to make.

8.1.4 Stress During the Interactive Task

Based on 5-item version of state component of State-Trait Anxiety Inventory (STAI) (Zsido et al., 2020)

5-point Likert items: '1 – Not at all, 2 – Slightly, 3 – Moderately, 4 – Very, 5 – Extremely'

How were you feeling during the interactive task?
1. I felt upset.
2. I felt frightened.

3. I felt nervous.
4. I was jittery.
5. I felt confused.

8.1.5 Attention

Adapted from 5-item version of the Mindful Awareness Attention Scale (MAAS) (as cited in O. R. Smith et al., 2017)

5-point Likert items: '1 – Strongly disagree, 2 – Disagree, 3 – Undecided, 4 – Agree, 5 – Strongly agree'

How much do you agree with the following statements?

During the interactive task,
1. it seemed I was "running on automatic", without much awareness of what I was doing.
2. I rushed through the activities without being really attentive to them.
3. I got so focused on the goal I wanted to achieve that I lost touch with what I was doing right in the moment to get there.
4. I did the task automatically, without being aware of what I was doing.
5. I found myself doing things without paying attention.

8.1.6 Elaboration

Adapted from Message Elaboration Measure (Reynolds, 1997)

5-point Likert items: '1 – Strongly disagree, 2 – Disagree, 3 – Undecided, 4 – Agree, 5 – Strongly agree'

How much do you agree with the following statements?

When presented with the warning on my phone I was
1. attempting to analyze the statements of the message
2. not very attentive to the implications
3. deep in thought about the message
4. unconcerned with the implications
5. extending a good deal of cognitive effort
6. distracted by other thoughts not related to the message
7. not really exerting my mind
8. doing my best to think about what was written
9. reflecting on the implications made
10. resting my mind
11. searching my mind in response to the implications
12. taking it easy

8.1.7 2FA Frequency

How often do you usually use an app (such as Google Authenticator, Microsoft Authenticator, BankID) to confirm logins in your daily life?

1. Never
2. Up to once a month
3. Up to once a week
4. Up to once a day

5. Up to several times a day

8.1.8 Involvement

Based on Involvement Scale (Solomon, 2020)

5-point semantic differential items, with 1 on the left size and 5 on the right side. Some items were reverse scored.

To me, logging in securely is

important	unimportant
boring	interesting
relevant	irrelevant
exciting	unexciting
means nothing	means a lot to me
appealing	unappealing
fascinating	mundane
worthless	valuable
involving	uninvolving
not needed	needed

8.1.9 Gender

What gender do you identify yourself as?

1. Female
2. Male
3. Non-binary
4. Prefer not to say

8.1.10 Age Group

How old are you?

1. 18-24 years old
2. 25 to 34 years old
3. 35 to 44 years old
4. 45 to 54 years old
5. 55 to 64 years old
6. 65 years or older