Information & Computer Security

Sharing passwords with strangers – A laboratory study

Anna Lena Fehlhaber
Department of Empirical Social Research Methods, Institute of Sociology, Leibniz University Hannover, Hanover, Germany

Received 14 October 2025
Revised 3 December 2025
Accepted 21 December 2025

Abstract

Purpose – This study aims to investigate the conditions under which people are inclined to engage with social engineering propositions. Using the contributions of Prospect Theory, the role of individual utility evaluations and risk perception were examined.

Design/methodology/approach – A laboratory experiment was conducted with a sample of 82 people from Germany. Participants were asked to work on tasks when they were approached by a third-party social engineer who offered to help them complete the task in exchange for their social media password. The experiment was conducted in a laboratory setting to control for extraneous factors.

Findings – The results showed that participants were more likely to share their password when the perceived profitability of doing so was high, when they expected that the trust would be rewarded, and when they perceived the stranger to be trustworthy. The findings suggest that this framework could be used to develop more effective strategies to prevent social engineering attacks.

Originality/value – This study uniquely contributes to the literature by applying a Prospect Theory framework to explore the decision-making processes of individuals in the context of social engineering, offering novel insights into the psychosociological mechanisms that influence individuals' susceptibility to phishing tactics.

Keywords Prospect theory, Laboratory study, Social engineering, Password sharing with strangers

Paper type Research paper

1. Introduction

Despite being discouraged due to security risks, password sharing remains common, often within trusted relationships such as families and friends, but occasionally extending to strangers (Singh *et al.*, 2007; Ouytsel, 2021; Ferreira *et al.*, 2013). Malevolent actors may attempt to gain the trust of a person to subsequently obtain unauthorized access to private accounts. This form of deception is known as social engineering, which involves the psychological manipulation of individuals into performing actions or divulging confidential information. Social engineering is regarded as one of the leading threats to information security in the contemporary era (Washo, 2021; Airehrour *et al.*, 2018). Its prevalence has been particularly evident in the context of social media (Chetioui *et al.*, 2022).

The present study addresses a critical gap in the literature on information security by focusing on password-sharing behaviours with strangers, a context that has received limited attention in prior research. Singh *et al.* (2007) examine password-sharing practices in various social and cultural contexts, such as within families, among couples and in remote

Information & Computer Security
Emerald Publishing Limited
2056-4961
DOI 10.1108/ICS-10-2025-0422

Indigenous communities in Australia, where such behaviours are often driven by trust and practical necessity. However, their study does not explore scenarios involving interactions with strangers, leaving a significant gap in understanding how and why individuals might share sensitive credentials in less familiar contexts. Similarly, Ouytsel (2021) highlights that even password sharing between close friends remains poorly understood, indicating a broader research gap in understanding unsafe password-sharing behaviours across varying relationship contexts. Uddin *et al.* (2023) emphasize that password-sharing behaviours significantly increase the risk of cybercrime victimization, particularly in interactions with strangers, yet little is known about the behavioural mechanisms driving such decisions (Uddin *et al.*, 2023).

Building on these findings, the present study bridges the gap in understanding how individuals make decisions about password sharing with strangers in social engineering scenarios. This investigation draws on Prospect Theory (PT) (Kahneman and Tversky, 1979), which provides a framework for understanding decision-making under uncertainty. To examine these processes, a realistic social engineering scenario is simulated in a controlled laboratory environment: Participants are approached by a stranger who offers a monetary incentive in exchange for their social media credentials. This setup mirrors critical aspects of real-world social engineering tactics while ensuring experimental control, enabling a systematic investigation of how individuals balance perceived risks and rewards in these situations.

Unlike previous surveys and role-playing investigations concerning social engineering (Jakobsson *et al.*, 2007; Salahdine and Kaabouch, 2019; Chatchalermpun and Daengsi, 2021; Algarni *et al.*, 2017), this laboratory study enhances ecological validity; By simulating realistic interactions, it captures key behavioural patterns in a way that prior research could not, enabling a deeper understanding of the sociopsychological mechanisms behind password-sharing decisions.

This study contributes to the literature by extending PT to the domain of information security, specifically within the context of social engineering. The research explores how individuals weigh potential rewards against risks when making decisions about password sharing, thus offering novel insights into the application of decision-making theories in security contexts. By focusing on the intersection of psychology, behavioural economics and cybersecurity, this study provides a theoretical foundation for future research aimed at understanding and mitigating the risks associated with social engineering.

While prior studies have explored phishing simulations in corporate email contexts that focus on link-clicking (Egelman *et al.*, 2008; Moody *et al.*, 2017), the broader spectrum of social engineering tactics on social media and private messages remains largely unexplored. This gap is crucial given the growing reliance on these platforms for everyday communication and their increasing exploitation as vectors for cyberattacks. Social engineering attacks, including phishing, often go beyond simple link-clicking. Many involve users being manipulated into logging into fake websites or directly sharing sensitive information, granting attackers extensive access to critical data and increasing the risk of identity theft. Although these tactics frequently rely on indirect methods or building trust over time, attackers sometimes exploit the element of surprise by making immediate requests for confidential information.

The present study captures and analyses these real-world scenarios in which individuals might choose to share their passwords, despite inherent risks. By applying PT, the study explores how participants' evaluations of risks and rewards shape their behaviour in these interactions.

2. Related work

2.1 Authentication and passwords

Usernames and passwords remain the dominant form of authentication, making safe handling essential. Although many alternatives such as biometrics exist (Lyastani *et al.*, 2020; Oesch and Ruoti, 2020; Bonneau *et al.*, 2012; Bachmann, 2014; Mitchell and Shing, 2018; Ozan, 2017), they still depend on passwords for setup, recovery (Cherapau *et al.*, 2015; Bud, 2018; Ryu *et al.*, 2021) or multi-factor authentication (Ibrokhimov *et al.*, 2019).

Common vulnerabilities include weak (Ur *et al.*, 2015; Pearman *et al.*, 2017), reused (Pearman *et al.*, 2017; Wang and Reiter, 2018; Abbott *et al.*, 2018) and even shared passwords, compromising account security. While biometric authentication and multi-factor authentication offer no sharing option or constant reminder for each login attempt, the traditional password – once shared – can be used and abused without the account owner's control (until it is changed). Hence, the associated risk is not to be understated – yet passwords are frequently shared.

2.2 Disclosure of passwords

Passwords are frequently shared within families and close friends to simplify access and demonstrate trust (Singh *et al.*, 2007; Ouytsel, 2021; Ferreira *et al.*, 2013). In some cases, friends logged into an account using shared login information without informing or obtaining consent from the account owner, highlighting risks inherent even in trusted relationships (Ouytsel, 2021).

Passwords are also shared with strangers, particularly in successful phishing attacks. Phishing typically occurs as broad attacks or targeted spear-phishing campaigns aimed at specific individuals. Whether people fall for phishing seems to be dependent on a number of factors, that can be behavioural, educational and demographic in nature, as has been discussed in literature and demonstrated in empirical studies (Blythe *et al.*, 2011; Sheng *et al.*, 2010; Darwish *et al.*, 2012; Moody *et al.*, 2017; Dhamija *et al.*, 2006; Tornblad *et al.*, 2021). Highlighting the importance of understanding user characteristics in predicting susceptibility to phishing attacks, Tornblad *et al.* (2021) identified 32 predictors across various domains such as personality traits, demographics, cybersecurity experience and email behaviours (Tornblad *et al.*, 2021). Atkins and Huang (2013) showed that the two primary triggers in social engineering attacks were signs of verification and the use of persuasion techniques by the attacker. Emotional persuasion and urgency, in particular, led to the people contacted dropping their guard (Atkins and Huang, 2013). Yang *et al.* (2022) developed a machine learning model to predict susceptibility to phishing. A notable correlation between personality and susceptibility to phishing has been demonstrated, with the personality trait most closely associated with vulnerability to phishing identified as extraversion (Yang *et al.*, 2022).

Despite extensive work, few studies empirically examine users' decision-making in phishing scenarios, highlighting a persistent research gap. A systematic literature review sheds light on this research deficit by revealing that a mere 13.9% of the 367 papers analysed focus on users and use empirical research methodologies such as interviews, surveys and in-lab studies. Furthermore, even within this limited subset, there is a striking lack of attention to essential methodological details and participant characteristics, indicating a critical need for further investigation into the human aspect of phishing vulnerability. It is therefore unsurprising that there has been little research conducted on specific phishing scenarios that deviate from the conventional phishing link via email.

Social media platforms, by design, foster information sharing and are thus highly vulnerable to phishing (Lei *et al.*, 2023) while misplaced user trust further heightens the risk of exploitation. In the realm of social network security, investigations have revealed that

common social engineering tactics such as spamming, identity replication and social bot manipulation, largely rely on counterfeit personas (Fire *et al.*, 2014). The potentiality of this occurrence can be attributed to a number of factors, including the propensity of individuals to engage with phishing offers and subsequently grant access to their online accounts.

A theoretically-driven overview of social engineering attack types on social media platforms is provided by Chetioui *et al.* (2022) and Bishnoi *et al.* (2023), who also created an overview of the individual steps involved in social engineering (investigation, hook, play and exit) and reverse social engineering techniques (Bishnoi *et al.*, 2023; Chetioui *et al.*, 2022). Frauenstein and Flowerday (2016) discuss the evolving threat of phishing, particularly through social networks, where users' habits of engaging with content can lead to increased susceptibility to so cial engineering attacks (Frauenstein and Flowerday, 2016). They argue that users' constant interaction with information on social media may lead to information overload, causing them to be less vigilant about security. Another study by Qahri-Saremi and Turel (2023) identified situational variables, including sleep quality, social media ostracism, source likability and fear appeal, as potential predictors of users' susceptibility to phishing messages on social media.

To date, empirical studies about social network platforms and social engineering have been relatively scarce. In a recent study, Ariani and colleagues (2023) examined the utilization of phishing tools on social media, investigating both the prevalence and the methods used (Ariani *et al.*, 2023). In a role-playing experiment on Facebook, Algarni *et al.* (2017) analysed how different persuasion techniques and aspects of source credibility affect the success of social engineering attempts. Nevertheless, the authors themselves have also identified that while role-playing games offer versatility, they may not perfectly mirror real-life behaviours, thus advocating for validation through empirical investigation in authentic settings.

The present study addresses this gap using an alternative theoretical framework without varying specific persuasion techniques, as these lie outside its theoretical scope.

2.3 Rationale for using prospect theory

Phishing scenarios deliberately induce trust and risk-taking by offering potential monetary gains or avoiding losses (Beckers and Pape, 2016; Jakobsson, 2016; Ferreira *et al.*, 2015). The broader field of behavioural science offers several theoretical frameworks for understanding decision-making in these uncertain contexts.

Several frameworks help explain behaviour under uncertainty, including the Theory of Planned behaviour (Ajzen, 1991), Protection Motivation Theory (Rogers, 1975, 1983), Social Exchange Theory (Blau, 1964), Technology Threat Avoidance Theory (Liang and Xue, 2009), Risk Perception Theory (Slovic, 1987) and Dual Process Theories (Kahneman, 2011). These approaches collectively highlight how perceived threat severity, coping appraisals, cognitive effort, and intuitive versus analytical reasoning shape security-related decisions. They offer valuable insights into why individuals form certain intentions or coping responses when confronted with cyber threats. However, these frameworks typically focus on motivational and cognitive antecedents of security behaviour rather than the evaluation of the risky choice itself. Unlike these intention-oriented models, PT directly models how individuals weigh potential gains and losses under uncertainty, making it particularly well suited for analyzing the specific decision structure exploited in phishing situations.

PT (Kahneman and Tversky, 1979) offers a more direct account of such decisions by describing how individuals evaluate potential gains and losses relative to a reference point, rather than through objective probabilities. It introduces mechanisms such as loss aversion, diminishing sensitivity and distorted probability perception (Tversky and Kahneman, 1992;

Kahneman *et al.*, 1982). In contrast to Expected Utility Theory, which serves as a normative baseline, PT captures how people *actually* respond to uncertain outcomes (Camerer, 1995), particularly in situations where subjective perceptions of risk matter more than objective calculations (Bleichrodt *et al.*, 2001).

In phishing contexts, these mechanisms provide a compelling explanation for why individuals may comply with risky requests: trust cues influence the perceived balance of gains and losses, leading to systematically biased judgments under uncertainty (Nguyen *et al.*, 2016). This makes PT particularly well suited for understanding decision-making in high-risk, high-uncertainty environments where individuals weigh immediate potential benefits (e.g. monetary incentives and social rewards) against potential losses (e.g. credential misuse).

Cumulative prospect theory (CPT) extends the original model by applying probability weighting to cumulative distribution functions and is especially useful for complex, multi-outcome lotteries (Tversky and Kahneman, 1992). In the present study, however, participants faced simple, discrete choices – accepting or rejecting a phishing-like offer – rather than continuous probability distributions. For such binary decisions, the qualitative components of the original PT, particularly loss aversion and reference dependence, are sufficient to capture the relevant psychological processes.

Although PT is often associated with formal parameter estimation, its original formulation is fundamentally a *descriptive* model of how individuals evaluate uncertain outcomes. In applied behavioural research, it is widely used without fitting full mathematical value or weighting functions, especially when the goal is to explain underlying cognitive mechanisms rather than derive quantitative utility estimates. Following this established practice, the present study draws on PT conceptually, without estimating its formal parameters. Instead of requiring participants to compute or report precise probabilities – which would not reflect naturalistic decision-making – the experiment elicits their subjective evaluations of risks and rewards and interprets disclosure behaviour through the lens of loss aversion, reference dependence and probability distortion. This approach provides a theoretically grounded yet practically applicable account of real-world decision-making in phishing situations.

2.4 Research model and hypotheses development

PT is applied to explain decision-making in response to phishing offers. Figure 1 depicts the research model used in the present study.

The model assumes that password-sharing decisions depend on perceived benefits, trustworthiness of the stranger and risk evaluation (Simpson, 2007). General attitudes such as risk tolerance and trust influence the decision path, consistent with PT's assumption that individuals evaluate potential gains and losses relative to a reference point rather than in purely objective terms. As illustrated in the model, these processes ultimately culminate in a decision regarding password sharing.

Risk-taking, trust, incentive and initial contact (including reciprocity) are delineated to model the decision-making process regarding password sharing under risk and uncertainty. This experiment investigates how these factors shape subjective evaluations of potential gains and losses in the context of a phishing-like offer.

2.4.1 Risk taking. Sociologically, risk-taking refers to assessing potential negative outcomes within social contexts (Beck, 1992; Lidskog and Sundqvist, 2012). Prior studies link higher risk-taking to greater phishing susceptibility (Abroshan *et al.*, 2021; Ayyagari and Crowell, 2020). People differ in their willingness to take risks (Zuckerman, 2007) because risk-taking depends on personality, experience and social environment (Lam and Ozorio, 2013; Gardner and Steinberg, 2005; Kennison and Chan-Tin, 2020). The general willingness to take risks can be measured using the validated short scale by Beierlein *et al.* (see Appendix

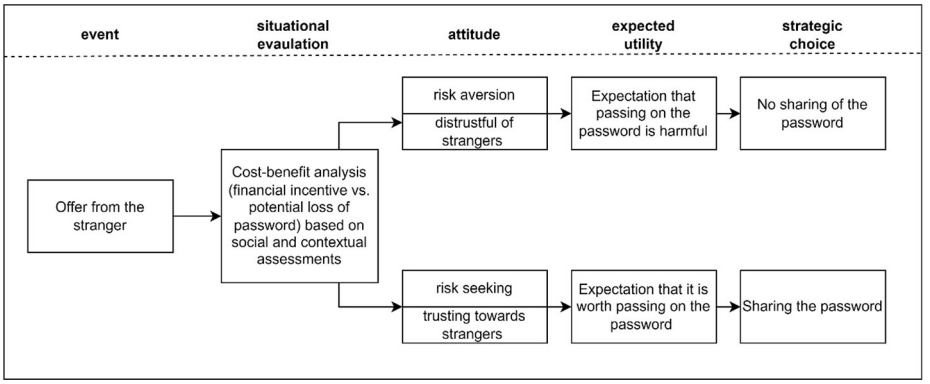

Figure 1. The pathways of the decision-making process in a social engineering situation.
Source(s): Authors' contribution

Table A1 and Beierlein, Kovaleva, Kemper, and Rammstedt, 2015) which incorporated items querying the self-assessment of risk-taking and the assessment of the social environment pertaining to risk-taking.

According to PT (Kahneman and Tversky, 1979: 269), individuals evaluate outcomes based on perceived gains and losses rather than objective probabilities. Due to loss aversion, potential losses weigh more heavily than equivalent gains, shaping risk behaviour. In the context of phishing, this framework suggests that individuals who do not recognize or undervalue the risks associated with sharing passwords (e.g. the potential misuse of personal information) are more likely to engage with the phishing attempt. The lack of awareness diminishes the perceived severity of potential losses, thereby increasing the likelihood of compliance with the phishing request (*H1*). Conversely, when risks are explicitly recognized and evaluated as significant, individuals are more likely to avoid risky behaviours, such as sharing sensitive credentials (*H2*). Risk-averse individuals reject phishing offers emphasizing potential losses, whereas risk-seeking individuals focus on possible rewards (*H3*):

H1. If there is little or no awareness of a potential abuse of trust, the password is more likely to be passed on to the stranger.

H2. The greater the perceived risk of disclosing passwords, the less likely it is that passwords will be disclosed.

H3. Those who are risk-averse are less likely to accept a phishing offer that presents a potential gain than those who are risk-seeking.

2.4.2 Trust. Trust – the expectation that others will act benevolently – is a necessary condition for sharing sensitive data. According to Luhmann (2000), system trust reflects assumptions about a system's stability and reliability and can be applied to the Internet as a social system (Thiedeke, 2004; Baltra, 2011; Falk and Kosfeld, 2006).

On the micro level, trust is shaped by personal preferences, perceptions of the situation and the relationship with the other party (Coleman, 1990) Trust inherently involves a degree of vulnerability, as it requires an advance payment from the trust giver, who risks potential

exploitation by the trusted party. Certain conditions facilitate or impede the formation of trust: For example, the perception of the trustworthiness of the counterpart, the way a situation presents itself, and the incentive that exists when trust has paid off (Deutsch, 1960; Conviser, 1973; Coleman, 1990).

Online trust in strangers depends on general Internet confidence, situational perception and expectations of others' behaviour (Freitag and Traunmüller, 2009; Uslaner, 2002; Bialski and Batorski, 2009; Bauer and Freitag, 2018).

While trust is not explicitly addressed in Tversky and Kahneman's original work on PT (1979), subsequent research has demonstrated its role in shaping subjective risk perception in decision-making under uncertainty (Nguyen *et al.*, 2016; Uslaner, 2007). Trust acts as a contextual factor that influences how individuals evaluate potential outcomes, particularly by altering the perceived likelihood and severity of risks. Trust reduces perceived losses, whereas distrust amplifies them.

In phishing scenarios, trust significantly influences decision-making processes. If an individual perceives the stranger as trustworthy, they are more likely to downplay the associated risks and focus on the potential benefits of compliance (*H4*). Similarly, the perception of a situation as safe and trustworthy reduces the subjective evaluation of risk, encouraging engagement (*H5*). Furthermore, broader assumptions about the likelihood of trust being honoured, such as the belief that the stranger will not exploit the trust given, can override rational assessments of risk, making individuals more susceptible to phishing attempts (*H6*).

By incorporating trust as a contextual factor, PT offers a nuanced explanation of how trust shapes the balance between perceived risks and rewards, influencing individuals' willingness to share sensitive information in uncertain scenarios:

H4. If the stranger is perceived as trustworthy, the password is more likely to be passed on.

H5. If the situation is perceived as trustworthy, the password is more likely to be passed on.

H6. If there is a general expectation that the stranger will not exploit the trust, the password is more likely to be passed on.

2.4.3 Money as an incentive. Financial incentives exploit the "need-and-greed principle" and are common in social engineering, especially phishing (Stajano and Wilson, 2011; Fischer and Evans, 2009) The motivational effect of money varies with personality, situation and perceived value (Jin and Huang, 2014; Lawler, 1981). Although often treated as universal motivators, monetary rewards are socially and individually shaped (Kraemer *et al.*, 2020; Morduch, 2017). The attribution of meaning to money can vary widely, and assuming it as static risks distorting findings (Frey, 1997; Kraemer *et al.*, 2020). Therefore, it is crucial to account for individual differences in the significance ascribed to monetary rewards when assessing their role in decision-making processes.

According to PT, subjective valuation of monetary rewards can distort risk evaluation, leading individuals to prioritize potential gains over losses. In phishing scenarios, the offered monetary reward serves as a key factor shaping the decision to comply with a request. Individuals who assign substantial value to the reward are more likely to engage in risky behaviour, as the perceived benefit outweighs the potential risks (*H7*). Conversely, for those who perceive the reward as insignificant, the same decision might not justify the associated risks. This dynamic highlights how the subjective importance of incentives interacts with individuals' broader evaluations of trust and risk in guiding behaviour.

H7. The greater the participants place the importance of money, the more willing they are to engage in the trade.

2.4.4 Initial contact. Despite limited emotional cues, familiarity and trust can develop online through psychological and social mechanisms (Jones and Moncur, 2018). Empirical studies in e-commerce show that reputation mechanisms and platform trust enable rapid development of initial trust (Liu and Tang, 2018; Li *et al.*, 2012). By contrast, social media platforms lack fiduciary responsibility and make it difficult to assess credibility. The potential trust-giver is initially unacquainted with the trust-requester and may even harbour suspicions of the trust-requester and must first overcome this scepticism so that trust can be established (Bachmann and Zaheer, 2013). In social engineering, several psychological mechanisms are at work to alleviate these doubts and to establish sufficient trust (Aleroud and Zhou, 2017; Fatima *et al.*, 2019; Mouton *et al.*, 2016). For instance, initial reciprocity – even if it is done without risk on the part of the person who subsequently wants to receive trust, can be an icebreaker (Kolm, 2000). Communicative interaction can also help build trust online between strangers (Kim and Kim, 2013). These techniques are referred to in the scientific literature as trust generation.

PT provides a framework for understanding how individuals evaluate potential gains and losses under conditions of uncertainty. Although the theory does not explicitly address mechanisms such as reciprocity, these mechanisms can influence the subjective evaluation of risks and benefits central to the theory. Reciprocity, in particular, creates a psychosociological obligation to reciprocate, which can shift individuals' focus from potential risks to perceived rewards, even when only the promise of a gift is presented (Chao, 2018).

Such gestures of reciprocity could reduce perceived uncertainty and foster an expectation of mutual exchange (*H8*). This dynamic alters the individual's cost-benefit analysis by emphasizing the potential social or relational gains, which may outweigh the consideration of associated risks. In phishing scenarios, this shift can lead individuals to downplay potential risks associated with compliance, increasing their likelihood of sharing sensitive information:

H8. Subjects who had recently interacted with the phisher and received an initial gift were more likely to respond positively to the phisher's subsequent request for their password.

3. Methodology

A combined experiment and post-survey were used to capture decision behaviour under uncertainty, avoiding biases typical of self-assessments, as individuals may provide socially desirable responses or evaluate themselves differently from their actual behaviour in the given situation (Ben-Ner and Halldorsson, 2010; Glaeser *et al.*, 2000).

The experimental design is loosely based on a game theory design, as the expected benefit of PT can be effectively visualized within this framework. Traditionally, these experimental games involve two individuals engaging in competitive decision-making. However, in this simplified game, the decision of the counterpart is not determined by an actual person but is simulated by the experimental administration and all participants acted as trust-givers; the counterpart's responses were simulated. This setup enables measurement of trust and perceived risk (Evans and Krueger, 2011).

After the experiment, participants completed a short questionnaire tailored to their decision path. This mixed-methods approach of lab study involving the experimental game

and subsequent survey aims to provide comprehensive and diverse perspectives regarding the generation of trust in strangers on the Internet.

3.1 Experimental design

The experiment was conducted in a controlled laboratory setting to examine password-sharing behaviour under realistic conditions. To prevent bias, participants were unaware of the phishing element and were initially asked to complete unrelated timed research tasks. The true purpose of the study – assessing whether and under which conditions individuals would grant a stranger access to their Facebook accounts – was revealed only during debriefing, and participants who disclosed their passwords received appropriate support in line with ethical standards.

To ensure ethical integrity while maintaining ecological validity, the study was conducted in accordance with the institutional behavioural research guidelines of a German university. At the time of data collection, deception-based minimal-risk studies in the social and behavioural sciences were not subject to mandatory review by a central ethics board under the institutional policies in place. Although participants entered their real Facebook passwords during the task, passwords were changed together with each participant immediately after debriefing, and no authentication data were stored or accessible to the researchers at any point. All participation was voluntary and based on informed consent, and all data were anonymized upon collection and processed in compliance with GDPR requirements.

After completing the experimental task, participants answered a follow-up survey that captured both quantitative and qualitative predictors of disclosure as well as the perceived influence of the trust-building elements. To ensure consistent experimental conditions, all participants received the same instructions and time limits, and every interaction with the purported stranger followed an identical scripted sequence. The study used a between-subjects design with two conditions that differed solely in the interaction preceding the password request. In the gift condition, participants received unsolicited help from the stranger before the offer was made; in the no-gift condition, the stranger contacted them only once with the same request. This manipulation targeted the initial-contact component of the social engineering scenario, while task structure, timing, incentives and instructions were held constant across conditions. The procedural sequence of both conditions is shown in Figure 2, and the resulting survey branches are outlined in Figure A1.

3.2 Procedure of the experiment

The experiment was conducted at a German university with student and visitor volunteers. All participants provided informed consent, were briefed on purpose and risks, and could withdraw at any time. Eligibility required age ≥18, sufficient language skills and a Facebook account.

Participants joined a private Facebook group to complete timed research tasks for monetary rewards. In the referred group, participants were presented with posts containing formulated questions, necessitating research on the designated topics, with the subsequent task of recording their findings on a digital notepad. They had eight minutes to answer as many questions as possible for payment per correct answer.

Participants were assigned to the two experimental conditions (see Figure 2 and Figure A1). Random assignment via RNG ensured equal group allocation; participants were blind to their condition. Group 1 received no prior contact and was approached later by an anonymous user offering extra payment. Group 2 first received unsolicited help (see Figure A2) from the same user, followed by the same reward offer.

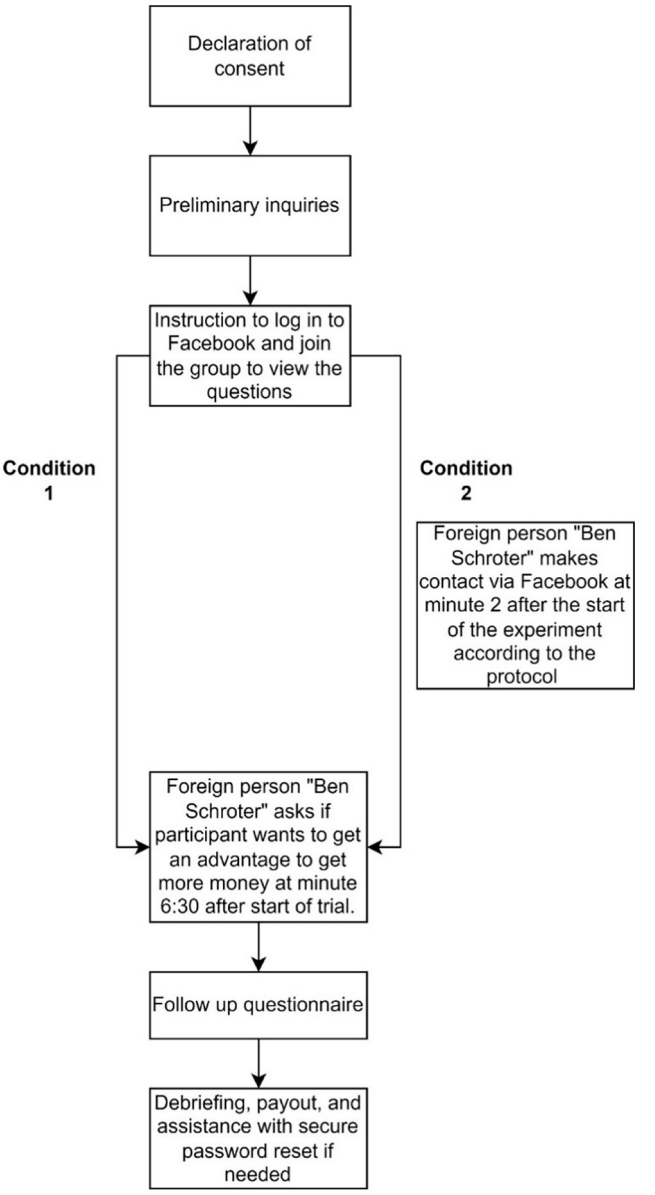

Figure 2. Procedure of the experiment
Source(s): Authors' contribution

Those who accepted were asked to share their Facebook password so the user could allegedly double their payout. Afterward, participants completed a questionnaire adjusted to their experimental branch.

3.3 Statistical procedures

To test the hypotheses (*H1–H8*), various statistical methods were used, chosen based on the nature of the data and the specific hypotheses. For hypotheses involving ordinal data measured on Likert scales (*H2*, *H3*, *H4*, *H5*, *H6* and *H7*), Kendall's Tau correlation coefficient was used. Kendall's Tau as chosen because it has a smaller gross error sensitivity and a smaller asymptotic variance, making it more robust for ordinal data (Croux and Dehon, 2010; Khamis, 2008; Howell, 2010). Given the strongly imbalanced outcome distribution and the high intercorrelations among the trust-related predictors, multivariate logistic regression was not suitable due to issues of separation and unstable coefficient estimates. The hypotheses were therefore tested using bivariate methods that are robust to the distributional properties of the data.

The formula for Kendall's Tau is given by:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c represents the number of concordant pairs and n_d the number of discordant pairs. Confidence intervals for Kendall's Tau were estimated using the normal approximation method based on its asymptotic distribution. While this approach is more accurate with large samples, it is also commonly applied in smaller samples as an approximation when exact or resampling-based methods (e.g. bootstrapping) are not used.

For hypotheses *H1* and *H8*, which involve binary variables, Fisher's Exact Test was used. This test is appropriate when sample sizes are small and expected frequencies in the contingency table are low, as it does not rely on the assumptions of the Chi-Square test. This method allows for accurate inference about the association between the variables of interest, even when conventional tests may fail due to small sample sizes. The test is based on the hypergeometric distribution and is expressed as:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

where a, b, c and d represent the frequencies in the contingency table, and n is the total number of observations. In addition to the exact *p*-value, Fisher's Exact test also provides a confidence interval for the odds ratio, calculated using exact methods based on the noncentral hypergeometric distribution. This interval quantifies the uncertainty around the strength and direction of the observed association, making it especially valuable in small-sample contexts.

In addition to reporting uncorrected p-values for the eight planned hypothesis tests (*H1–H8*), *p*-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure. For completeness, Bonferroni-corrected *p*-values were also computed.

A post-hoc power analysis was conducted for the trust-related hypotheses (*H4–H7*) to evaluate the sensitivity of the correlational tests. The reported Kendall's τ coefficients ($\tau = 0.40$, 0.31, 0.37 and 0.24) were converted into their approximate Pearson *r* values and analysed using standard correlation power calculations with $\alpha = 0.05$ and $n = 82$. The resulting power estimates exceeded 0.99 for $\tau = 0.40$, $\tau = 0.31$ and $\tau = 0.37$ and reached 0.93 for $\tau = 0.24$. These values

indicate that the study design provided sufficient statistical power to detect the medium to large associations observed in the trust-related variables.

4. Data analysis and results

4.1 Descriptive analysis

The sample consisted of 82 participants (31 females, 51 males) with a median age of 21 years (SD: 6.68). They were randomly assigned to one of the experimental conditions. To examine whether randomization resulted in comparable groups, baseline equivalence between conditions was assessed. Age, gender and selected dispositional variables (general risk-taking, internet use and social media use) were compared across conditions using *t*-tests, χ^2-tests, Fisher's exact tests and Wilcoxon rank-sum tests, as appropriate. Regarding internet usage, 97.5% of the sample ($n = 80$) reported regular internet use. In addition, 80.5% ($n = 66$) indicated they regularly used social networks. A total of 46 individuals reported a general sense of safety online, in contrast to the 36 who did not. Among the respondents, 17 individuals indicated that they had already been the victims of internet-facilitated crimes, while 20 respondents expressed concern about becoming victims in the future. A significant relationship was observed between past victimization and the probability of future victimization, as indicated by Fisher's Exact test ($p < 0.001$, 95% CI [2.92, 51.19], Cohen's w = 0.482).

To gain further insights into participants' general risk-taking behaviour, a detailed analysis was conducted. The distribution of risk-taking behaviour revealed two distinct patterns within the sample. A total of 31 participants (37.8%) were classified as "Rather not willing to take risks", indicating a reluctance to engage in risky activities. In contrast, 22 participants (26.8%) were categorized as "Rather willing to take risks", showing a higher propensity for risk-taking behaviour. Together, these two groups make up more than half of the entire sample. The remaining participants exhibited moderate or intermediate levels of risk-taking tendencies, falling between these two extremes (see Figure 3).

In light of these findings, it is also crucial to understand how participants perceived and responded to specific risks within the context those risks were presented (Figure 4). An overwhelming majority of the sample (74%) perceived sharing their password as a risk. Furthermore, a high degree of variability emerges when evaluating the situation's trustworthiness. After all, 26% rated it as neutral, and another 51% experienced the stimulus an (rather) untrustworthy occurrence. The tagged free-text responses show that most participants rated the situation as untrustworthy because the stranger contacted them via Facebook and they had no prior face-to-face contact that could have reduced the perceived stranger status. Moreover, an unencrypted Facebook chat was deemed an unreliable medium for password sharing. Facebook, as a platform, was generally perceived as untrustworthy, with commitments made on it often being disregarded due to the nature of the platform. Other study participants indicated that they found the situation trustworthy because they were in a secure study environment and did not expect anything to happen to them or their account. Finally, data is heterogeneous as to whether sharing passwords is perceived as worthwhile for participants, whereas a clear skewness in the distributions shows that sharing passwords was not perceived as a practice paying off. The majority of the sample (63%) does not perceive password sharing in this specific situation to be beneficial.

To further understand the nuances of these perceptions, it is essential to consider the assessment of the stranger involved in the scenario (Figure 5). Most study participants (59%) rated the person who messaged them on Facebook as untrustworthy, while another 18% said they found the person trustworthy. When looking at the tagged free text responses, it is stated that unknown persons are generally not trusted and certainly not on the anonymous Internet. Those who answered "neutral" specified in the free text question that they had ignored the

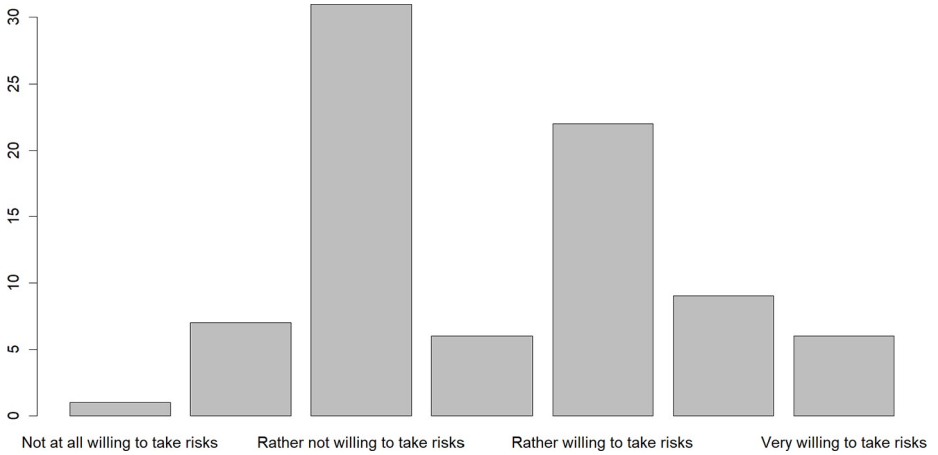

Figure 3. Risk taking propensity of the sample measured with the short scale for willingness to take risks (Beierlein, Kovaleva, Kemper, and Rammstedt, 2015)
Source(s): Authors' contribution

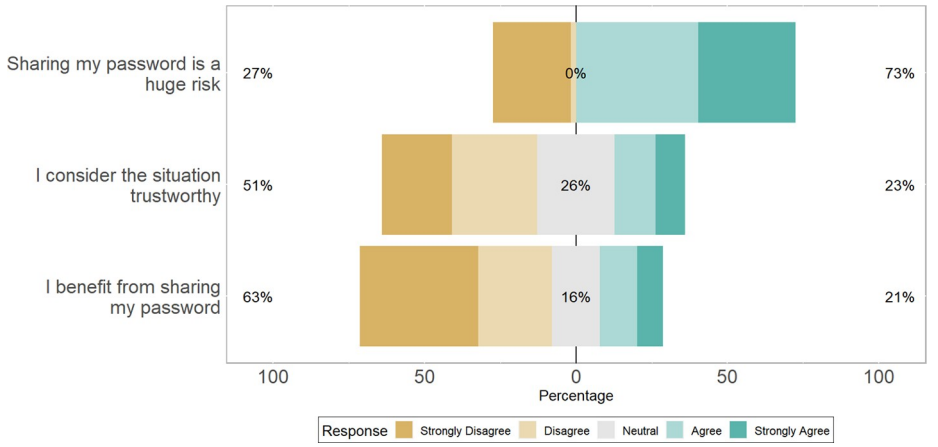

Figure 4. Assessment of the situation
Source(s): Authors' contribution

person either intentionally or unintentionally and did not attach any further importance to him. Those who found him trustworthy indicated that he had seemed likeable in his manner of conversation and honoured that he had given them answers they needed for the study.

The majority (55%) rejects the proposition that the stranger does not gain an advantage from them. Another 27% rate the statement as neutral, and only 18% agree (strongly). Another item asked whether trust in the stranger would pay off, or would have paid off if the offer had been accepted. Just under one-third took a neutral stance, while the majority, with 53% of votes, disagreed that trust in the stranger would pay of and 15% expected a positive outcome. The items "I consider the stranger trustworthy" and "I expect that the stranger will not take advantage

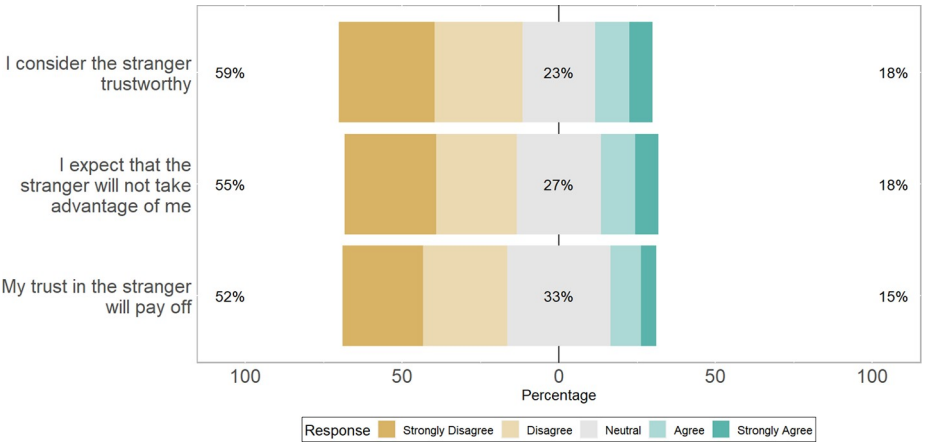

Figure 5. Assessment of the stranger
Source(s): Authors' contribution

of me" are correlated according to relatively conservative Kendall's tau ($p < 0.001$, tau 0.562), and also "My trust in the stranger will pay off" and "I expect that the stranger will not take advantage of me" ($p < 0.001$, tau = 0.487) as well as "My trust in the stranger will pay off" and "I consider the stranger trustworthy" ($p < 0.001$, tau = 0.494).

Throughout the study, a total of 15 participants disclosed their passwords, while the remaining 67 subjects did not. Among those who made their passwords available, 11 were in the first experimental condition. Four of them were in the second experimental condition and received a gift in the form of knowledge from the person subsequently requesting trust. With regard to the demographic factors, a significant correlation of the covariate age at $p = 0.001$ (t (80) = −3.34, 95% CI [−9.59, −2.43]) on the disclosure of the password was found with a small to medium effect of 0.31. For the covariate gender, no significant correlation could be found for the given alpha level of 0.05 (χ^2 (1, n) = 1.16, $p = 0.281$).

4.2 Statistical testing

To examine the psychological and contextual determinants of participants' willingness to disclose their passwords to a stranger (see Appendix Figure A1), eight hypotheses (*H1–H8*) were tested. These covered cognitive risk awareness, individual differences in risk preferences, trust perceptions and situational or motivational factors. The results of the statistical analysis for each hypothesis (*H1–H8*) are summarized in Table 1.

The post-hoc power analysis indicated high statistical power for the trust-related hypotheses. Based on the observed Kendall's τ values ($\tau = 0.40$, 0.31, 0.37 and 0.24), the resulting power estimates exceeded 0.99 for the first three effects and reached 0.93 for the smallest of the four. These values suggest that the study was sufficiently sensitive to detect the medium-to-large associations identified in the trust-related analyses. When controlling the false discovery rate across the eight planned hypothesis tests, all previously significant trust-related associations with password disclosure (*H4–H7*) remained statistically significant at $q < 0.05$ (Benjamini-Hochberg adjustment). Under the more conservative Bonferroni correction (α_Bonferroni = 0.00625), the three strongest trust-related effects (*H4–H6*) remained statistically significant, whereas the smaller effect i*n H7* did not meet the adjusted threshold. The overall pattern of results therefore continues to

Table 1. Hypotheses results with estimates, *p*-values, confidence intervals and effect magnitudes

Variable	Test	Estimate	*p*	CI lower	CI upper	Effect
H1: Awareness of potential abuse of trust and password disclosure	*Fisher's exact test*	*3.28*	0.225	0.5	21.64	Large (OR)
H2: Perceived risk of disclosing passwords	*Kendall's tau*	*0.06*	0.549	−0.09	0.22	Negligible (τ)
H3: Risk Aversion and Phishing Offer Acceptance	*Kendall's tau*	*0.19*	0.055	0.05	0.33	Small (τ)
H4: Perceived trustworthiness of the stranger	*Kendall's tau*	*0.4*	0.0001	0.27	0.52	Moderate–large (τ)
H5: Perceived trustworthiness of the situation	*Kendall's tau*	*0.31*	0.0022	0.17	0.43	Moderate (τ)
H6: Expectation that trust will not be exploited	*Kendall's tau*	*0.37*	0.0002	0.24	0.49	Moderate (τ)
H7: Importance of money	*Kendall's tau*	*0.24*	0.017	0.1	0.38	Small (τ)
H8: Prior interaction with the phisher	*Fisher's exact test*	*0.29*	0.084	0.09	1.02	Large (protective OR)

Note(s): Estimates represent Kendall's τ for *H2–H7* and odds ratios for *H1* and *H8*. Effect magnitude thresholds follow conventional guidelines for τ (< 0.10 negligible, 0.10–0.29 small, 0.30–0.49 moderate, ≥ 0.50 large) and for odds ratios (OR ≈ 1.22, 1.86, and 3.00 indicating small, medium and large effects, respectively; Chen, Cohen and Chen, 2010; protective effects interpreted symmetrically for OR < 1)

support the central role of trust-related evaluations in participants' willingness to disclose their passwords.

4.2.1 Trust-related perceptions. Trust-related variables emerged as particularly strong predictors of disclosure behaviour.

H4 hypothesized that participants who perceived the stranger as trustworthy would be more likely to share their password. The analysis confirmed this assumption, revealing a strong positive association, $\tau = 0.40$, 95% CI [0.27, 0.52], $p < 0.001$, representing a moderate–large effect.

Similarly, *H5* proposed that perceived trustworthiness of the situation would increase the likelihood of disclosure. This hypothesis was also supported, $\tau = 0.31$, 95% CI [0.17, 0.43], $p = 0.002$, indicating a moderate effect.

H6 focused on participants' expectations about whether the stranger would exploit their trust. Participants who assumed that their trust would not be taken advantage of were significantly more likely to disclose their password, $\tau = 0.37$, 95% CI [0.24, 0.49], $p < 0.001$, again reflecting a moderate effect.

Taken together, these findings highlight the central role of interpersonal and contextual trust in participants' decision-making.

4.2.2 Risk-related factors. The role of risk perception and individual risk preference was assessed in *H2* and *H3*.

H2 posited that higher perceived risk would reduce the likelihood of disclosure. However, the data showed no significant association, $\tau = 0.06$, 95% CI [−0.09, 0.22], $p = 0.549$, indicating that perceived risk alone did not deter participants from engaging in potentially compromising behaviour.

In contrast, *H3* focused on dispositional risk aversion, predicting that risk-averse individuals would be less likely to disclose. This hypothesis was not supported, although the data showed a small positive association that narrowly missed conventional significance thresholds, $\tau = 0.19$, 95% CI [0.05, 0.33], $p = 0.055$. This pattern suggests that trait-level risk sensitivity may play a role in disclosure behaviour, though the effect should be interpreted with caution.

4.2.3 Cognitive awareness and motivational influences. H1 examined the effect of participants' awareness of potential trust abuse on disclosure behaviour. While the odds ratio suggested a trend towards increased disclosure when awareness was low, OR = 3.28, 95% CI [0.50, 21.64], $p = 0.225$, the result did not reach statistical significance, and thus the hypothesis was not supported. Instrumental motivation was addressed in *H7*, which proposed that participants who placed greater importance on monetary gain would be more likely to share their passwords. The analysis revealed that a higher valuation of money was a significant predictor of password disclosure, $\tau = 0.24$, 95% CI [0.10, 0.38], $p = 0.017$, representing a small effect. Finally, *H8* examined whether prior interaction with the phisher, including receiving a gift beforehand, increased the likelihood of password disclosure. While there was a trend in the expected direction, the association was not statistically significant, $p = 0.084$ and the odds ratio, OR = 0.29, 95% CI [0.09, 1.02], did not provide sufficient evidence to support the hypothesis. In summary, the findings support hypothesis *H4*, *H5*, *H6* and *H7* (Table 1).

5. Discussion and implications

5.1 Key findings

The findings of this experiment focus on several key areas related to the likelihood of password disclosure within a social media context, guided by the principles of PT. The discussion is structured to highlight the unique contributions of this research, comparing and contrasting with existing studies, and addressing the specific hypotheses in detail. Before discussing the hypotheses, the demographic effects – being the most easily comparable across studies – are evaluated.

5.1.1 Age. The analysis revealed an age-related pattern: older participants were more willing to disclose their passwords than younger ones. Prior research has produced mixed findings, with some studies reporting greater caution among older adults (Sarno *et al.*, 2020) and others showing reduced discrimination of phishing attempts (Grilli *et al.*, 2021; Oliveira *et al.*, 2017). Younger adults have, in turn, been found to share passwords more frequently (Whitty *et al.*, 2015). Although the age range in this sample is relatively narrow, the present results suggest that susceptibility may vary gradually across age rather than only between extreme groups.

5.1.2 Gender. No gender differences were observed. This aligns with the largely inconsistent evidence on gender and cybersecurity behaviour (Moody *et al.*, 2017; Leukfeldt, 2014; Abroshan *et al.*, 2021; Broadhurst *et al.*, 2019). Studies reporting gender effects (Verkijika, 2019; Anwar *et al.*, 2017) typically involve highly heterogeneous samples; such patterns are less likely to appear in a relatively homogeneous academic context.

5.1.3 Awareness of potential trust abuse (H1). Awareness of potential trust abuse did not reduce disclosure. This contrasts with work suggesting that awareness can improve phishing resistance (Sarno *et al.*, 2020; Alnajim and Munro, 2009), but it is consistent with evidence that knowledge often fails to translate into behaviour (Downs *et al.*, 2007). Within the PT framework, this can be explained by the limited personal relevance of the threat in a controlled setting, reducing the salience of potential losses.

5.1.4 Perceived risk of disclosing passwords (H2). Perceived risk showed no association with disclosure. Previous studies also indicate that perceived severity alone is a poor predictor of protective behaviour (Downs *et al.*, 2007). The laboratory setting may have reduced the subjective relevance of potential negative outcomes, diminishing the expected effect of loss aversion.

5.1.5 Risk aversion (H3). Dispositional risk aversion showed only a small, nonsignificant association with password disclosure. Earlier studies linked risk attitudes to susceptibility (Sheng *et al.*, 2010; Abroshan *et al.*, 2021), but the present findings indicate that such traits can be overshadowed by situational cues. PT accounts for this by emphasizing that framing effects may outweigh stable risk preferences during decisions under uncertainty.

5.1.6 Perceived trustworthiness of the stranger (H4). Perceived trustworthiness of the stranger was a strong predictor of password disclosure. Participants who viewed the counterpart as sincere and reliable were substantially more willing to share their credentials, which aligns with prior research demonstrating that perceived credibility is central to social engineering success (Algarni *et al.*, 2017). In terms of PT, trust acts as a framing cue: when the interaction partner appears trustworthy, individuals mentally downweigh the probability of negative outcomes and attend more to the potential benefits of cooperation (Kahneman and Tversky, 1979). This shift in subjective evaluation offers a coherent explanation for the significant impact of interpersonal trust in the present experiment.

5.1.7 Perceived trustworthiness of the situation (H5). Trust in the situation also significantly influenced disclosure. Participants who rated the overall context as credible were more likely to comply, indicating that situational cues can be as influential as interpersonal trust. Social media phishing benefits from precisely this effect: familiar interfaces and everyday communication patterns create a misleading sense of safety, whereas email phishing often contains detectable irregularities (McAlaney and Hills, 2020; Kleitman *et al.*, 2018). From a PT perspective, a credible setting reduces the perceived likelihood of losses (Tversky and Kahneman, 1992), making individuals less vigilant and more susceptible to manipulation. The present findings support this mechanism.

5.1.8 Expectation regarding trust exploitation (H6). Expectations regarding whether trust would be exploited had a similarly strong effect. Participants who believed the stranger would not misuse their password were markedly more willing to disclose it. This pattern suggests that decisions were shaped less by objective risk evaluation and more by anticipated interpersonal outcomes. PT offers a clear explanation: individuals assess potential gains and losses relative to their expectations rather than objective probabilities. When a positive outcome is expected, the subjective weight of potential losses decreases, making disclosure appear less risky (Tversky and Kahneman, 1992). This aligns with the strong association observed in the present data.

5.1.9 Importance of money (H7). Participants who attached greater importance to the monetary incentive were more likely to disclose, reflecting the effectiveness of financial lures in social engineering (Atkins and Huang, 2013). PT explains this through reference-dependent valuation, whereby potential gains draw attention away from uncertain losses.

5.1.10 Prior interaction with the phisher (H8). Prior interaction and gifting did not increase disclosure. This suggests that reciprocity alone is insufficient to elicit trust in the

absence of stronger credibility cues, which is consistent with research on persuasion in phishing contexts (Algarni, 2019) and PT's certainty effect (Tversky and Kahneman, 1992; Chao, 2018).

5.2 Limitations

Before examining the theoretical and practical implications of this study, it is essential to consider its limitations. Due to the location of recruitment, the sample consisted primarily of students and members of the university and is therefore limited in its representativeness. Moreover, because all participants were recruited at a single German university, the findings reflect one specific cultural and institutional context and may not readily generalize to other countries or cultural settings. In addition, the sample shows limited diversity with respect to academic background and digital literacy. Because recruitment took place exclusively on campus, participants shared similar educational trajectories and digital habits, which precludes meaningful analysis of whether these factors systematically influence disclosure behaviour. Future studies would need to draw on more heterogeneous, non-academic populations to examine whether the dynamics observed here differ in groups with broader demographic profiles. Furthermore, the relatively complex execution of the experiment and the subsequent follow-up survey limit the total number of subjects with whom the experiment can be conducted due to time constraints. Although the sample size was inherently limited by the structure of the experimental procedure, the post-hoc analyses indicate that the study was adequately powered to detect medium to large effects in the trust-related hypotheses. Nonetheless, smaller effects may not have been reliably identified under these conditions, and non-significant findings should therefore be interpreted with caution. It is also plausible that physical proximity to the university environment fostered a social desirability effect, leading individuals to exhibit a heightened sense of being well-informed and cautious in handling their data.

In addition, there are some difficulties inherent to laboratory experiments, especially in modelling complex phenomena and the control of all possibly influencing variables. Besides, it is noteworthy that under time-constrained circumstances, there is a discernible elevation in the overall propensity for cooperation, as evidenced by numerous laboratory experiments, including those featuring competitive conditions (Cone and Rand, 2014; Rand *et al.*, 2012). Because time stress is inherent in many real-world phishing situations, this variable was not experimentally altered, but it could be worthwhile to investigate in further research whether omitting time stress influences the willingness to respond to the phishing offer and, in this case, to make the password accessible. This consideration is particularly relevant because, remarkably, in the evaluation of the free text responses, 19 respondents subsequently stated that they had intentionally or unintentionally ignored or not read the message and eight other respondents felt under too much pressure to react due to time pressure.

5.3 Theoretical implications

The findings of this study contribute to theoretical work on social engineering by clarifying how trust cues influence subjective evaluations of risk and reward within the framework of PT. Rather than relying on general attitudes towards risk, participants responded primarily to interpersonal and situational trust signals. This supports the view that decision-making in phishing contexts is shaped by reference-dependent evaluations and framing effects rather than objective probability assessments. Trust reduced the perceived likelihood of loss, allowing potential gains to take precedence – an effect consistent with PT's account of how individuals weigh outcomes under uncertainty.

The results also complement cognitive and behavioural cybersecurity models by demonstrating where PT extends beyond intention-focused frameworks such as Protection Motivation Theory or Technology Threat Avoidance Theory. While these models explain how users form threat perceptions and coping appraisals, PT explains the final choice behaviour: how users trade off potential losses and benefits when confronted with a concrete disclosure request. The strong influence of trust-related cues in this study illustrates the importance of incorporating psychological framing mechanisms into existing models of phishing susceptibility.

Finally, the absence of significant effects for several dispositional variables, including risk aversion, suggests that situational framing may override stable traits when individuals make rapid decisions in social engineering encounters. This highlights the need for future theoretical work to integrate PT with contextual and interpersonal factors, and to examine how trust and framing shape decision dynamics across different digital environments.

5.4 Managerial implications

Security awareness initiatives should account for the trust-based nature of social media attacks. In this study, participants disclosed credentials not because they lacked technical knowledge, but because their trust was deliberately manipulated. organizations should therefore complement conventional, technically oriented training with psychosocial awareness programmes that teach users to critically evaluate contextual cues such as tone, familiarity and timing, using realistic simulations of seemingly benign social interactions involving rewards or peer-like communication.

Technical and monitoring measures on social media platforms also need to reflect these dynamics. Passwords were revealed despite two-factor authentication, indicating that interface protections alone are insufficient when social trust is exploited. Platforms can respond by introducing adaptive trust warnings, real-time detection of manipulative behaviour, and brief confirmation prompts when users attempt to share credentials in chats. Building on Aun *et al.* (2023), integrating deep-learning-based phishing detection into social-media messaging could support timely, context-sensitive intervention.

Finally, platform and interface design strongly shape perceived safety. Many participants felt secure simply because the interaction took place in a controlled environment, illustrating how environmental cues can override technical literacy. UX and UI designers should counter this illusion of safety by incorporating trust indicators and carefully placed friction prompts when credentials are shared in unverified conversations. User-centred design that reduces cognitive load and encourages brief reflection can help curb impulsive disclosure.

Beyond immediate organizational practices, the results also highlight broader implications for cybersecurity policy and digital literacy. Because the decision to disclose a password emerged not from lack of technical knowledge but from trust-driven distortions in evaluating gains and losses, interventions must extend beyond teaching factual indicators of phishing. Integrating these psychological mechanisms into digital literacy curricula – whether in schools, vocational training or community programmes – can help users recognize how interpersonal cues shape subjective risk assessments. At the policy level, emphasizing trust awareness and decision-making under uncertainty in national cybersecurity guidelines could strengthen preventive efforts and reduce the societal impact of social engineering attacks. In this sense, the findings not only inform organizational training but also contribute to a wider cultural shift towards recognizing social manipulation as a central cybersecurity threat. In essence, organizations and developers should move from reactive cybersecurity to anticipatory, behaviourally informed strategies based on how trust and interface dynamics shape user decisions.

6. Conclusion and future directions

This study has made several original contributions to understanding how individuals assess the risks and benefits of disclosing passwords to strangers on social media, using the framework of PT. The innovative methodological approach combines a controlled laboratory experiment featuring actual simulated phishing attempts with subsequent detailed surveys – a significant advancement over previous studies that relied primarily on surveys or role-playing scenarios (Algarni, *et al.*, 2017; Algarni, *et al.*, 2016; Das, *et al.*, 2019; Ouytsel, 2021; Whitty, *et al.*, 2015). This dual-method approach provides more reliable insights into real-world behaviour while maintaining experimental control, offering a new standard for research in this field.

The findings underscore the complexity of decision-making in social engineering scenarios, particularly how trust and expected benefits significantly influence cybersecurity behaviours. The application of PT to social media phishing represents a novel theoretical contribution, demonstrating how trust-related considerations influence the evaluation of gains and losses. In social engineering and phishing contexts, individuals do not merely rely on a generalized inclination to assume or eschew risks; rather, they engage in nuanced decision-making processes guided by their subjective assessments of potential gains and losses in these scenarios.

The study revealed that trust can significantly influence these evaluations, prompting a re-assessment of potential risks and benefits that may lead to a higher likelihood of password disclosure. This re-evaluation process is crucial, as it underscores the importance of situational factors – like the perceived trustworthiness of the phisher and the specific context of the interaction – over broad dispositional traits.

Beyond explaining the underlying psychological mechanisms, the findings also offer several implications for practice, policy and digital literacy. In theoretical terms, the results demonstrate that the mechanisms described by PT; particularly framing effects, reference-dependent evaluations and the down weighting of negative outcomes under trust, provide a robust explanation for why individuals disclose sensitive information in socially engineered situations. Practically, the findings indicate that interventions must address not only technical knowledge but also the interpersonal and situational cues that shape subjective evaluations of gains and losses. This includes designing phishing awareness programmes that simulate social interaction patterns, updating organizational policies to create clearer rules for verifying interpersonal requests, and developing digital literacy training that teaches users how trust cues distort risk perception. More broadly, these insights can support efforts to strengthen cybersecurity culture by shifting users from purely technical vigilance towards socially informed decision-making.

Building on these insights, future work and practice should focus on how trust-induced distortions in risk evaluation can be mitigated at both user and system level. At the user level, training programmes can emphasize the recognition of interpersonal manipulation and encourage verification and slow-down strategies when sensitive information is requested. At the system level, platforms and organizations can experiment with lightweight friction and context-sensitive warnings that prompt users to reconsider disclosing credentials in private conversations. Evaluating such measures in longitudinal and field studies would deepen our understanding of how trust, risk perception, and interface design jointly shape disclosure decisions in everyday digital environments.

References

Abbott, J., Calarco, D. and Camp, L.J. (2018), "Factors influencing password reuse: a case study", *TPRC 46: the 46th Research Conference on Communication, Information and Internet Policy 2018*, doi: 10.2139/ssrn.3142270.

Abroshan, H., Devos, J., Poels, G. and Laermans, E. (2021), "Phishing happens beyond technology: the effects of human behaviors and demographics on each step of a phishing process", *IEEE Access*, Vol. 9, pp. 44928-44949, doi: 10.1109/ACCESS.2021.3066383.

Airehrour, D., Nair, N. and Madanian, S. (2018), "Social engineering attacks and countermeasures in the New Zealand banking system: advancing a user-reflective mitigation model", *Information*, Vol. 9 p. 110, doi: 10.3390/info9050110.

Ajzen, I. (1991), "The theory of planned behavior", *Organizational Behavior and Human Decision Processes*, Vol. 50 No. 2, pp. 179-211, doi: 10.1016/0749-5978(91)90020-T.

Aleroud, A. and Zhou, L. (2017), "Phishing environments, techniques, and countermeasures: a survey", *Computers and Security*, Vol. 68, pp. 160-196, doi: 10.1016/j.cose.2017.04.006.

Algarni, A. (2019), "What message characteristics make social engineering successful on Facebook: the role of Central route, peripheral route, and perceived risk", *Information*, Vol. 10 No. 6, doi: 10.3390/info10060211.

Algarni, A., Xu, Y., and Cha, T. (2016), "'Measuring source credibility of social engineering attackers on Facebook", *49th Hawaii International Conference on System Sciences (HICSS)*, pp. 3686-3695, doi: 10.1109/HICSS.2016.460.

Algarni, A., Xu, Y. and Chan, T. (2017), "An empirical study on the susceptibility to social engineering in social networking sites: the case of Facebook", *European Journal of Information Systems*, Vol. 26 No. 6, pp. 661-687, doi: 10.1057/s41303-017-0057-y.

Alnajim, A., and Munro, M. (2009), "'Effects of technical abilities and phishing knowledge on phishing websites detection", *Proceedings of the IASTED International Conference on Software Engineering*, pp. 120-125.

Anwar, M., He, W., Ash, I., Yuan, X., Li, L. and Xu, L. (2017), "Gender difference and employees' cybersecurity behaviors", *Computers in Human Behavior*, Vol. 69, pp. 437-443, doi: 10.1016/j.chb.2016.12.040.

Ariani, P., Jayanti, K.S., Atmaja, I.G.B., Dewi, I.G.A.A., Saskara, G.A.J. and Listartha, I.M.E. (2023), "Comparative analysis of phishing tools on social media sites", *Ultimatics: jurnal Teknik Informatika*, Vol. 15 No. 1, pp. 22-27, doi: 10.31937/ti.v15i1.2920.

Atkins, B. and Huang, W. (2013), "A study of social engineering in online frauds", *Open Journal of Social Sciences*, Vol. 1 No. 3, pp. 23-32, doi: 10.4236/JSS.2013.13004.

Aun, Y., Gan, M.-L., Wahab, N.H.B.A. and Guan, G.H. (2023), "Social engineering attack classifications on social media using deep learning", *Computers, Materials and Continua*, Vol. 47 No. 3, pp. 4917-4931, doi: 10.32604/cmc.2023.032373.

Ayyagari, R. and Crowell, A. (2020), "Risk and demographics' influence on security behavior intentions", *J. Southern Assoc. Inf. Syst*, Vol. 7 No. 1, pp. 1-13, doi: 10.17705/3JSIS.00013.

Bachmann, M. (2014), "Passwords are dead: alternative authentication methods", 2014 IEEE Joint Intelligence and Security Informatics Conference, p. 322, doi: 10.1109/JISIC.2014.67.

Bachmann, R., and Zaheer, A. (2013), *Handbook of Advances in Trust Research*, Edward Elgar Publishing, Cheltenham, UK, doi: 10.4337/9780857931382.

Baltra, A.I. (2011), "Cheshire: a design framework for alternate reality games", In Anacleto, J.C., Fels, S., Graham, N., Kapralos, B., Saif El-Nasr, M. and Stanley, K. (Eds), *Entertainment Computing – ICEC 2011. ICEC 2011. Lecture Notes in Computer Science*, Vol. 6972, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-24500-8_38.

Bauer, P.C., and Freitag, M. (2018), "Measuring trust", In *Uslaner, E.M. The Oxford Handbook of Social and Political Trust*, Oxford University Press, Oxford, doi: 10.1093/oxfordhb/9780190274801.001.0001.

Beck, U. (1992), *Risk Society: towards a New Modernity*, SAGE Publications, New York, NY, doi: 10.2307/3341155.

Beckers, K. and Pape, S. (2016), "A serious game for eliciting social engineering security requirements", *2016 IEEE 24th International Requirements Engineering Conference (RE)*, pp. 16-25, doi: 10.1109/RE.2016.39.

Beierlein, C., Kovaleva, A., Kemper, C.J. and Rammstedt, B. (2015), "Kurzskala zur erfassung der risikobereitschaft/short scale for assessing the willingness to take risks", doi:10.6102/zis236, available at: https://zis.gesis.org/skala/Beierlein-Kovaleva-Kemper-Rammstedt-Kurzskala-zur-Erfassung-der-Risikobereitschaft-(R-1)

Ben-Ner, A. and Halldorsson, F. (2010), "Trusting and trustworthiness: What are they, how to measure them, and what affects them", *Journal of Economic Psychology*, Vol. 31 No. 1, pp. 64-79, doi: 10.1016/j.joep.2009.10.001.

Bialski, P., and Batorski, D. (2009), "From online familiarity to offline trust. How a virtual community creates familiarity and trust between strangers", In Zaphiris, P. and Ang, C.S. (Eds), *Social Computing and Virtual Communities*, Chapman and Hall/CRC, New York, NY, doi: 10.1201/9781420090437-c8.

Bishnoi, A., GarvBishnoi, S. and Gupta, N. (2023), "Comprehensive assessment of reverse social engineering to understand social engineering attacks", *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, doi: 10.1109/ICSSIT55814.2023.10061054.

Blau, P.M. (1964), "Justice in social exchange", *Sociological Inquiry*, Vol. 34, pp. 193-206, doi: 10.1111/J.1475-682X.1964.TB00583.X.

Bleichrodt, H., Pinto, J.L. and Wakker, P.P. (2001), "Making descriptive use of prospect theory to improve the prescriptive use of expected utility", *Management Science*, Vol. 47 No. 11, pp. 1498-1514.

Blythe, M., Petrie, H.L., and Clark, J.A. (2011), "F for fake: four studies on how we fall for phish", *Proceedings of the International Conference on Human Factors in Computing Systems (CHI 2011)*, doi: 10.1145/1978942.1979459.

Bonneau, J., Herley, C., Oorschot, P.C.V. and Stajano, F. (2012), "The quest to replace passwords: a framework for comparative evaluation of web authentication schemes", *2012 IEEE Symposium on Security and Privacy. IEEE*. pp. 553-567, doi: 10.1109/SP.2012.44.

Broadhurst, R., Skinner, K., Sifniotis, N., Matamoros-Macias, B. and Ipsen, Y. (2019), "Phishing and cybercrime risks in a university student community", *SSRN Electronic Journal*, Vol. 2 No. 1, pp. 4-23, doi: 10.2139/ssrn.3176319.

Bud, A. (2018), "Facing the future: the impact of apple FaceID", *Biometric Technology Today*, Vol. 1, pp. 5-7, doi: 10.1016/S0969-4765(18)30010-9.

Camerer, C. (1995), "Individual decision making", In Kagel, J.H. and Roth, A.*E. (Eds), The Handbook of Experimental Economics*, Princeton University Press, Princeton, doi: 10.1515/9780691213255-010.

Chao, M. (2018), "Intentions-based reciprocity to monetary and non-monetary gifts", *Games*, Vol. 9 No. 4, p. 74, doi: 10.3390/g9040074.

Chatchalermpun, S. and Daengsi, T. (2021), "Improving cybersecurity awareness using phishing attack simulation", In IOP Conference Series: materials Science and Engineering, *Pupose-Led Publishing*, doi: 10.1088/1757-899X/1088/1/012015.

Cherapau, I., Muslukhov, I., Asanka, N., and Beznosov, K. (2015), "On the impact of touch ID on iPhone passcodes", *Symposium on Usable Privacy and Security (SOUPS) 2015*, pp. 257-276, doi: 10.5555/3235866.3235888.

Chetioui, K., Bah, B., Alami, A.O. and Bahnasse, A. (2022), "Overview of social engineering attacks on social networks", *Procedia Computer Science*, Vol. 198, pp. 656-661, doi: 10.1016/j.procs.2021.12.302.

Coleman, J.S. (1990), *Foundations of Social Theory*, The Belknap Press, Cambridge, Mass./London.

Cone, J. and Rand, D.G. (2014), "Time pressure increases cooperation in competitively framed social dilemmas", *PLoS ONE*, Vol. 9 No. 12, p. e115756, doi: 10.1371/journal.pone.0115756.

Conviser, R.H. (1973), "Toward a theory of interpersonal trust", *The Pacific Sociological Review*, Vol. 16 No. 3, pp. 377-399, doi: 10.2307/1388493.

Croux, C. and Dehon, C. (2010), "Influence functions of the spearman and Kendall correlation measures", *Statistical Methods and Applications*, Vol. 19, pp. 497-515, doi: 10.1007/s10260-010-0142-z.

Darwish, A., Zarka, A.E. and Aloul, F. (2012), "Towards understanding phishing victims' profile", *2012 International Conference on Computer Systems and Industrial Informatics*, pp. 1-5, doi: 10.1109/ICCSII.2012.6454454.

Das, S., Kim, A., Tingle, Z., and Nippert-Eng, C. (2019), "All about phishing: exploring user research through a systematic literature review", *Proceedings of the Thirteenth International Symposium on Human Aspects of Information Security and Assurance*, doi: 10.48550/arXiv.1908.05897.

Dhamija, R., Tygar, J.D. and Hearst, M. (2006), "Why phishing works", *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 581-590, doi: 10.1145/1124772.1124861.

Downs, J.S., Holbrook, M. and Cranor, L.F. (2007), "Behavioral response to phishing risk", eCrime '07: *proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pp. 37-44, doi: 10.1145/1299015.1299019.

Egelman, S. Cranor, L.F., and Hong, J. (2008), "You've been warned: an empirical study of the effectiveness of web browser phishing warnings", CHI '08: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, doi: 10.1145/1357054.1357219.

Evans, A.M. and Krueger, J.I. (2011), "Outcomes and expectations in dilemmas of trust", *Judgment and Decision Making*, Vol. 9 No. 2, pp. 90-103.

Falk, A. and Kosfeld, M. (2006), "The hidden cost of control", *American Economic Review*, Vol. 96 No. 5, pp. 1611-1630, doi: 10.1257/aer.96.5.1611.

Fatima, R., Yasin, A., Liu, L. and Wang, J. (2019), "How persuasive is a phishing email? A phishing game for phishing awareness", *Journal of Computer Security*, Vol. 27 No. 3, pp. 1-32, doi: 10.3233/JCS-181253.

Ferreira, A., Correia, R., Chadwick, D.W., Santos, H., Gomes, R., Reis, D. and Antunes, L. (2013), "Password sharing and how to reduce it", In *IT Policy and Ethics: Concepts, Methodologies, Tools, and Applications*, IGI Global, Hershey, PA, doi: 10.4018/978-1-4666-2919-6.ch002.

Ferreira, A., Coventry, L. and Lenzini, G. (2015), "Principles of persuasion in social engineering and their use in phishing", International Conference on Human Aspects of Information Security, Privacy, and Trust, pp. 36-47, doi: 10.1007/978-3-319-20376-8_4.

Fire, M., Goldschmidt, R. and Elovici, Y. (2014), "Online social networks: threats and solutions", *Communications Surveys and Tutorials, IEEE*, Vol. 16 No. 4, pp. 2019-2036, doi: 10.1109/COMST.2014.2321628.

Fischer, S.L.P. and Evans, K.M. (2009), "The psychology of scams: provoking and committing errors of judgement", United Kingdom: University of Exeter School of Psychology.

Frauenstein, E.D. and Flowerday, S.V. (2016), "Social network phishing: becoming habituated to clicks and ignorant to threats?", Information Security for South Africa (ISSA), pp. 98-105, doi: 10.1109/ISSA.2016.7802935.

Freitag, M. and Traunmüller, R. (2009), "Spheres of trust: an empirical analysis of the foundations of particularised and generalised trust", *Eur. J. Polit. Res*, Vol. 48, pp. 782-803, doi: 10.1111/J.1475-6765.2009.00849.X.

Frey, B. (1997), *Not Just for the Money: An Economic Theory of Personal Motivation*, Edward Elgar Publishing, Cheltenham.

Gardner, M. and Steinberg, L. (2005), "Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study", *Developmental Psychology*, Vol. 41 No. 4, pp. 625-635, doi: 10.1037/0012-1649.41.4.625.

Glaeser, E.L., Laibson, D.I., Scheinkman, J.A. and Soutter, C.L. (2000), "Measuring trust", *The Quarterly Journal of Economics*, Vol. 115 No. 3, pp. 811-846, doi: 10.1162/003355300554926.

Grilli, M.D., McVeigh, K.S., Hakim, Z.M., Wank, A.A., Getz, S.J., Levin, B.E., Ebner, N.C. and Wilson, R.C. (2021), "Is this phishing? Older age is associated with greater difficulty discriminating between safe and malicious emails", *The Journals of Gerontology: Series B*, Vol. 76 No. 9, pp. 1711-1715, doi: 10.1093/geronb/gbaa228.

Howell, D.C. (2010), *Statistical Methods for Psychology*, Wadsworth, Cengage Learning, Belmont, CA.

Ibrokhimov, S., Hui, K.L., Al-Absi, A.A., Lee, H.J. and Sain, M. (2019), "Multi-factor authentication in cyber physical system: a state of art survey", *21st International Conference on Advanced Communication Technology (ICACT)*, pp. 279-284, doi: 10.23919/ICACT.2019.8701960.

Jakobsson, M. (2016), *Understanding Social Engineering Based Scams*, Springer Science+Business, New York, NY, doi: 10.1007/978-1-4939-6457-4.

Jakobsson, M., Tsow, A., Shah, A., Blevis, E., and Lim, Y.-K. (2007), "What instills trust? A qualitative study of phishing", *Financial Cryptography and Data Security*, pp. 356-361, doi: 10.1007/978-3-540-77366-5_32.

Jin, L. and Huang, Y. (2014), "When giving money does not work: the differential effects of monetary versus in-kind rewards in referral reward programs", *International Journal of Research in Marketing*, Vol. 31 No. 1, pp. 107-116, doi: 10.1016/j.ijresmar.2013.08.005.

Jones, H.S. and Moncur, W. (2018), "The role of psychology in understanding online trust', In McAlaney, J., Frumkin, L.A. and Benson,V. (Eds). *Psychological and Behavioral Examinations in Cyber Security*, IGI Global, Hershey, PA, doi: 10.4018/978-1-5225-4053-3.ch007.

Kahneman, D. (2011), *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York, NY.

Kahneman, D. and Tversky, A. (1979), "Prospect theory: an analysis of decision under risk", *Econometrica*, Vol. 47 No. 2, pp. 263-292.

Kahneman, D., Slovic, S.P., Slovic, P. and Tversky, A. (1982), *Judgment under Uncertainty: Heuristics and Biases*, Erlbaum, Hillsdale, NJ, doi: 10.1017/cbo9780511809477.012.

Kennison, S.M. and Chan-Tin, E. (2020), "Taking risks with cybersecurity: using knowledge and personal characteristics to predict Self-Reported cybersecurity behaviors", *Frontiers in Psychology*, Vol. 11, doi: 10.3389/fpsyg.2020.546546.

Khamis, H. (2008), "Measures of association how to choose?", *JDMS*, Vol. 24, pp. 155-162, doi: 10.1177/8756479308317006.

Kim, D. and Kim, J.H. (2013), "Understanding persuasive elements in phishing e-mails: a categorical content and semantic network analysis", *Online Information Review*, Vol. 37 No. 6, pp. 835-850, doi: 10.1108/OIR-03-2012-0037.

Kleitman, S., Law, M.K.H. and Kay, J. (2018), "It's the deceiver and the receiver: individual differences in phishing susceptibility and false positives with item profiling", *PLoS ONE*, Vol. 13 No. 10, doi: 10.1371/journal.pone.0205089.

Kolm, S.-C. (2000), "The theory of reciprocity', In Gérard-Varet, L.-A., Kolm, S.-C. and Ythier, J.M. *The Economics of Reciprocity, Giving and Altruism*, Palgrave Macmillan, London, doi: 10.1007/978-1-349-62745-5_5.

Kraemer, K., Jakelja, L., Brugger, F. and Nessel, S. (2020), "The social ambiguity of money: empirical evidence on the multiple usability of money in social life", *Review of Social Economy*, doi: 10.1080/00346764.2022.2076150.

Lam, D. and Ozorio, B. (2013), "The effect of prior outcomes on gender risk-taking differences", *Journal of Risk Research*, Vol. 16 No. 7, pp. 791-802, doi: 10.1080/13669877.2012.737824.

Lawler, E.E. (1981), *Pay and Organization Development*, Addison-Wesley Publishing Company, Boston.

Lei, W., Hu, S. and Hsu, C. (2023), "Uncovering the role of optimism bias in social media phishing: an empirical study on TikTok", *Behaviour and Information Technology*, pp. 1827-1841, doi: 10.1080/0144929X.2023.2230305.

Leukfeldt, E.R. (2014), "Phishing for suitable targets in the Netherlands: routine activity theory and phishing victimization", *Cyberpsychology, Behavior, and Social Networking*, Vol. 17 No. 8, pp. 551-555.

Li, F., Pieńkowski, D., Moorsel, A. and Smith, C. (2012), "A holistic framework for trust in online transactions", *International Journal of Management Reviews*, Vol. 14, pp. 85-103, doi: 10.1111/j.1468-2370.2011.00311.x.

Liang, H. and Xue, Y. (2009), "Avoidance of information technology threats: a theoretical perspective", *MIS Quarterly*, Vol. 33 No. 1, pp. 71-90.

Lidskog, R., and Sundqvist, G. (2012), "Sociology of risk", In Roeser, S., Hillerbrand, R., Sandin, P. and Peterson, M. (Eds), *Handbook of Risk Theory*, Springer Science+Business Media, Luxemburg, doi: 10.1007/978-94-007-1433-5_40.

Liu, Y. and Tang, X. (2018), "The effects of online trust-building mechanisms on trust and repurchase intentions: an empirical study on eBay", *Information Technology and People*, Vol. 31 No. 3, pp. 666-687, doi: 10.1108/ITP-10-2016-0242.

Luhmann, N. (2000), *Vertrauen: Ein Mechanismus Der Reduktion Sozialer Komplexität*, UVK, Gänsheide, München.

Lyastani, S.G., Schilling, M., Neumayr, M., Backes, M. and Bugiel, S. (2020), "Is FIDO2 the Kingslayer of user authentication? A comparative usability study of FIDO2 passwordless authentication", *2020 IEEE Symposium on Security and Privacy (SP). IEEE*. pp. 268-285, doi: 10.60882/cispa.24613269.

McAlaney, J. and Hills, P.J. (2020), "Understanding phishing email processing and perceived trustworthiness through eye tracking", *Frontiers in Psychology*, Vol. 11, doi: 10.3389/fpsyg.2020.01756.

Mitchell, C. and Shing, C.-C. (2018), "Discussing alternative login methods and their advantages and disadvantages", *14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1353-1356, doi: 10.1109/FSKD.2018.8687163.

Moody, G.D., Galletta, D.F. and Dunn, B.K. (2017), "Which phish get caught? An exploratory study of individuals´ susceptibility to phishing", *European Journal of Information Systems*, Vol. 26, pp. 564-584, doi: 10.1057/s41303-017-0058-x.

Morduch, J. (2017), "'Economics and the social meaning of money'", In Bandelj, N, Wherry, F.F. and Zelizer, V.A. (Eds), *Money Talks. Explaining How Money Really Works*, Princeton University Press, Princeton, doi: 10.1515/9781400885268-004.

Mouton, F., Leenen, L. and Venter, H.S. (2016), "Social engineering attack examples, templates and scenarios", *Computers and Security*, Vol. 59, pp. 186-209, doi: 10.1016/j.cose.2016.03.004.

Nguyen, Q., Villeval, M.C. and Xu, H. (2016), "Trust and trustworthiness under the prospect theory: a field experiment in Vietnam", *Economic Development and Cultural Change*, Vol. 64 No. 3, pp. 545-572, doi: 10.1016/j.cose.2016.03.004.

Oesch, S., and Ruoti, S. (2020), "That was then, this is now: a security evaluation of password generation, storage, and autofill in browser-based password managers", *Proceedings of the 29th USENIX Security Symposium*. pp. 2165-2182.

Oliveira, D., Rocha, H., Yang, H., Ellis, D., Dommaraju, S., Muradoglu, M., Weir, D., Soliman, A., Lin, T. and Ebner, N. (2017), "Dissecting spear phishing emails for older vs young adults: on the interplay of weapons of influence and life domains in predicting susceptibility to phishing", Proceedings of the 2017 chi conference on human factors in computing systems, pp. 6412-6424, doi: 10.1145/3025453.3025831.

Ouytsel, J.V. (2021), "The prevalence and motivations for password sharing practices and intrusive behaviors among early adolescents' best friendships – a mixed-methods study", *Telematics and Informatics*, Vol. 63, doi: 10.1016/j.tele.2021.101668.

Ozan, E. (2017), "Password-free authentication for social networks", *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1-5, doi: 10.1109/CCWC.2017.7868484.

Pearman, S., Thomas, J., Naeini, P.E., Habib, H., Bauer, L., Christin, N., Cranor, L.F., Egelman, S. and Forget, A. (2017), Let's go "in for a closer look: observing passwords in their natural habitat. Proceedings of the 2017 ACM", SIGSAC Conference on Computer and Communications Security (CCS '17), doi: 10.1145/3133956.3133973.

Qahri-Saremi, H. and Turel, O. (2023), "Situational contingencies in susceptibility of social media to phishing: a temptation and restraint model", *Journal of Management Information Systems*, Vol. 40 No. 2, pp. 503-540, doi: 10.1080/07421222.2023.2196779?

Rand, D.G., Greene, J.D. and Nowak, M.A. (2012), "Spontaneous giving and calculated greed", *Nature*, Vol. 489, pp. 427-430, doi: 10.1038/nature11467.

Rogers, R.W. (1975), "A protection motivation theory of fear appeals and attitude Change1", *The Journal of Psychology*, Vol. 91 No. 1, pp. 93-114, doi: 10.1080/00223980.1975.9915803.

Rogers, R.W. (1983), "Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. 153-177", In Cacioppo, J. and Petty R. (Eds), *Social Psychophysiology*, Guilford Press, New York, NY.

Ryu, R., Yeom, S., Kim, S.-H. and Herbert, D. (2021), "Continuous multimodal biometric authentication schemes: a systematic review", *IEEE Access*, Vol. 9, pp. 34541-34557, doi: 10.1109/ACCESS.2021.3061589.

Salahdine, F. and Kaabouch, N. (2019), "Social engineering attacks: a survey", *Future Internet*, Vol. 11 No. 4, doi: 10.3390/FI11040089.

Sarno, D.M., Lewis, J.E., Bohil, C.J. and Neider, M.B. (2020), "Which phish is on the hook? Phishing vulnerability for older versus younger adults", *Human Factors*, Vol. 62 No. 5, pp. 704-717, doi: 10.1177/0018720819855570.

Sheng, S., Lanyon, M.B., Kumaraguru, P., Cranor, L. and Downs, J. (2010), "Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions", *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, doi: 10.1145/1753326.1753383.

Simpson, J.A. (2007), "Psychological foundations of trust", *Current Directions in Psychological Science*, Vol. 16 No. 5, pp. 264-268, doi: 10.1111/j.1467-8721.2007.00517.x.

Singh, S., Cabraal, A., Demosthenous, C., Astbrink, G. and Furlong, M. (2007), "Password sharing: implications for security design based on social practice", *CHI '07: proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 895-904, doi: 10.1145/1240624.1240759.

Slovic, P. (1987), "Perception of risk", *Science*, Vol. 236 No. 4799, pp. 280-285, doi: 10.1126/science.3563507.

Stajano, F. and Wilson, P. (2011), "Understanding scam victims: seven principles for systems security", *Communications of the ACM*, Vol. 54 No. 3, pp. 70-75, doi: 10.1145/1897852.1897872.

Thiedeke, U. (2004), *Soziologie Des Cyberspace. Medien, Strukturen Und Semantiken*, VS Verlag für Sozialwissenschaften, Wiesbaden, doi: 10.1007/978-3-322-80482-2.

Tornblad, M.K., Jones, K.S., Namin, A.S. and Choi, J. (2021), "Characteristics that predict phishing susceptibility: a review", *Proceedings of the 2021 HFES 65th International Annual Meeting*, pp. 938-942.

Tversky, A. and Kahneman, D. (1992), "Advances in prospect theory: cumulative representation of uncertainty", *Journal of Risk and Uncertainty*, Vol. 5, pp. 297-323, doi: 10.1007/BF00122574.

Uddin, M.A., Supti, A.Z., Asgar, R. and Mridha, S. (2023), "Cyber security awareness (CSA) and cybercrime in Bangladesh: a statistical modeling approach", *Australian Journal of Engineering and Innovative Technology*, Vol. 5 No. 1, pp. 15-25, doi: 10.34104/ajeit.023.015025.

Ur, B., Noma, F., Bees, J., Segreti, S.M., Shay, R., Bauer, L., Christin, N. and Cranor, L.F. (2015), "I added '!' at the end to make it secure": observing password creation in the lab", *Symposium on Usable Privacy and Security (SOUPS) 2015*.

Uslaner, E.M. (2002), *The Moral Foundations of Trust. Cambridge*, Cambridge University Press, doi: 10.2139/ssrn.824504.

Uslaner, E.M. (2007), "Trust and risk",. In Siegrist, M. and Gutscher, H. (Eds), *Trust in Risk Management - Uncertainty and Scepticism in the Public Mind*, Routledge, London, doi: 10.4324/9781849776592.

Verkijika, S.F. (2019), "If you know what to do, will you take action to avoid mobile phishing attacks" self-efficacy, anticipated regret, and gender", *Computers in Human Behavior*, Vol. 101, pp. 286-296, doi: 10.1016/j.chb.2019.07.034.

Wang, K.C., and Reiter, M.K. (2018), "How to end password reuse on the web", *Proceedings of the 26th ISOC Network and Distributed System Security Symposium (NDSS 2019)*, pp. 1-17, doi: 10.48550/arXiv.1805.00566.

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

Whitty, M., Doodson, J., Creese, S. and Hodges, D. (2015), "Individual differences in cyber security behaviors: an examination of who is sharing passwords", *Cyberpsychol Behav Soc Netw*, Vol. 18 No. 1, pp. 3-7, doi: 10.1089/cyber.2014.0179.

Yang, R., Zheng, K., Wu, B., Li, D., Wang, Z. and Wang, X. (2022), "Predicting user susceptibility to phishing based on multidimensional features", *Computational Intelligence and Neuroscience (Special Issue: Artificial Intelligence and Machine Learning-Driven Decision-Making)*, pp. 1-11, doi: 10.1155/2022/7058972.

Zuckerman, M. (2007), *Sensation Seeking and Risky Behavior*, American Psychological Association, Washington, DC, D.C, doi: 10.1037/11555-000.

Further references

Austin, C.R., Bobek, D.D. and Jackson, S. (2021), "Does prospect theory explain ethical decision making? Evidence from tax compliance", *Accounting, Organizations and Society*, Vol. 94, p. 101251, doi: 10.1016/j.aos.2021.101251.

Barberis, N.C. (2013), "Thirty years of prospect theory in economics: a review and assessment", *Journal of Economic Perspectives*, Vol. 27 No. 1, pp. 173-196, doi: 10.1257/jep.27.1.173.

Chadee, A.A., Chadee, X.T., Chadee, C. and Otuloge, F. (2022), "Violations at the reference point of discontinuity: limitations of prospect theory and an alternative model of risk choices", *Emerging Science Journal*, Vol. 6 No. 1, pp. 37-52, doi: 10.28991/ESJ-2022-06-01-03.

Clark, W.A.V. and Lisowski, W. (2017), "Prospect theory and the decision to move or stay", *Proceedings of the National Academy of Sciences*, Vol. 114 No. 36, pp. 7432-7440, doi: 10.1073/pnas.1708505114.

Khaw, M.W., Stevens, L. and Woodford, M. (2021), "Individual differences in the perception of probability", *PLoS Computational Biology*, Vol. 17 No. 4, p. e1008871, doi: 10.1371/journal.pcbi.1008871.

Rossiter, J.R. (2019), "A critique of prospect theory and framing with particular reference to consumer decisions", *Journal of Consumer Behaviour*, Vol. 18 No. 5, pp. 399-405, doi: 10.1002/cb.1779.

Ruggeri, K., Alí, S., Berge, M.L., Bertoldo, G., Bjørndal, L.D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., Gibson, S.P., Jarke, H., Karakasheva, R., Khorrami, P.R. and Kveder, J. (2020), "Replicating patterns of prospect theory for decision under risk", *Nature Human Behaviour*, Vol. 4, pp. 622-633, doi: 10.1038/s41562-020-0886-x.

Appendix

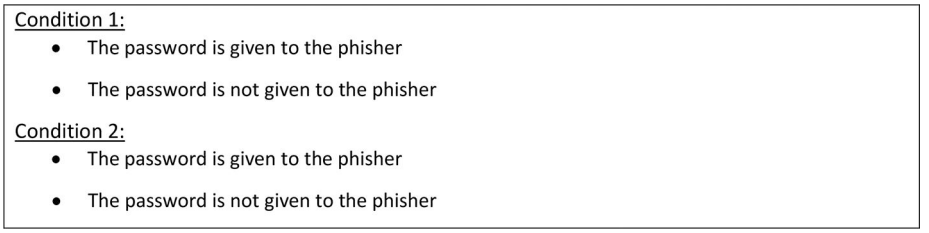

Figure A1. The four possible branches that determine the wording of the final survey

Table A1. Closing questionnaire (not all items were used in this paper)

Question	Evaluation	Theoretical concept/empirical measurement tool	Branches and wording differences
Q1: How high did you find the incentive to take up the phisher's offer (the person who wrote to you on Facebook)?	5-point-Likert scale, 1 = no incentive; 5 = very high incentive	Situational awareness and personal preferences (risk, social) according to Coleman (1990)	none
Q2: How high did you perceive the risk of permanently losing your password by sharing it with the phisher?	5-point-Likert scale, 1= no risk; 5 = very high risk	Situational awareness and personal preferences (risk, social) according to Coleman (1990)	none
Q3: I expect that the stranger will not/would not have take/took advantage of me	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Expectations of the other person's behaviour	Wording depends on the decision for or against password sharing
Q4: I believe I can/could have benefited from this exchange	5-point-Likert scale, 1= strongly disagree; 5 = strongly agree	Situational awareness and personal preferences (risk, social) according to Coleman (1990)	Wording depends on the decision for or against password sharing
Q5: The person who made me the offer seemed trustworthy to me	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Conditions under which trust can be established according to Coleman (1990)	none
Q6: Why did the person (not) seem trustworthy to you?	Free text box	Conditions under which trust can be established according to Coleman (1990)	Wording according to the answer from Q5
Q7: The situation in which the offer was made to me seemed trustworthy	5-point-Likert scale, 1= strongly disagree; 5 = strongly agree	Conditions under which trust can be established according to Coleman (1990)	none
Q8: Why did the situation (not) seem trustworthy to you?	Free text box	Conditions under which trust can be established according to Coleman (1990)	Wording according to the answer from Q7
Q9: Did the conversation with the person on Facebook have any influence on you (not) sharing your password?	Free text box	Conditions under which trust can be established according to Coleman (1990)	Only for condition 2 (prior contacting of the person taking the trust)

(*continued*)

Table A1. Continued

Question	Evaluation	Theoretical concept/empirical measurement tool	Branches and wording differences
Q10: I took/would take a risk by giving away my password	5-point-Likert scale, 1= strongly disagree; 5 = strongly agree	Awareness of taking a risk	Wording depends on the decision for or against password sharing
Q11: I expect to receive/have received the promised additional money	Binary (yes/no)	Social expectations and willingness to trust/distrust according to Falk and Kosfeld (2006)	none
Q12: I expect my trust to pay off/have paid off	5-point-Likert scale, 1= strongly disagree; 5 = strongly agree	Social expectations and willingness to trust/distrust according to Falk and Kosfeld (2006)	Wording depends on the decision for or against password sharing
Q13: How do you personally rate yourself: Are you generally a risk-taker, or do you try to avoid risk?	7-point-Likert	Single-item scale for assessing risk tolerance by Beierlein *et al.* (2015).	none
Q14: How often do you use the Internet?	often/rarely/never	Risk exposure/frequency of use	none
Q15: How often do you use social networks (e.g. Twitter, Facebook and Snapchat)?	often/rarely/never	Risk exposure/frequency of use	none
Q16: I am familiar with the Internet and the social network I use	5-point-Likert scale, 1= strongly disagree; 5 = strongly agree	Perceived empowerment - how familiar subjects are with digital environment	none
Q17: I am confident that the Internet is safe	Yes/No/I don't' know	System trust according to Luhmann, adapted by Thiedeke (2004) for cyberspace	none
Q18: I am confident that the social network I use is secure	Yes/No/ I don't' know	System trust according to Luhmann, adapted by Thiedeke (2004) for cyberspace	none

(*continued*)

Table A1. Continued

Question	Evaluation	Theoretical concept/empirical measurement tool	Branches and wording differences
Q19: Internet security is important to me in general	Binary (yes/no)	Personal security relevance	none
Q20: It is important to me to protect myself on the Internet	Binary (yes/no)	Personal security relevance	none
Q21: I find that the protective measures to guard yourself on the Internet are easy to apply	Binary (yes/no)	Perceived empowerment cybersecurity	none
Q22: I believe that there is no point in taking protective measures on the Internet	Binary (yes/no)	System trust according to Luhmann, adapted by Thiedeke (2004) for cyberspace	none
Q23: Why (not)?	Free text box	Specify the reasons	Wording according to the answer from Q22
Q24: I do not feel safe on the Internet	Binary (yes/no)	System trust according to Luhmann, adapted by Thiedeke (2004) for cyberspace	none
Q25: I am worried about becoming a victim of a digital crime (e.g. identity theft, hacking, social engineering)	Binary (yes/no)	Fear of becoming a victim	none
Q26: I have been a victim of a digital crime in the past	Binary (yes/no)	Fear of becoming a victim	none
Q27: I think it is likely that I will be a victim of a digital crime (again)	Binary (yes/no)	Fear of becoming a victim	Wording according to the answer from Q26
Q28: I limit my use of the Internet because of the dangers involved	Binary (yes/no)	Rejection or acceptance of the threat according to the technology acceptance model	Asked only when either Q24, Q25, Q26 or Q27 indicate that there are cybersecurity concerns
Q29: Please elaborate why	Free text box	Rejection or acceptance of the threat according to the technology acceptance model	Asked only if answer at Q28 is "Yes"

(*continued*)

Table A1. Continued

Question	Evaluation	Theoretical concept/empirical measurement tool	Branches and wording differences
Q30: I limit my use of social media because of the dangers involved	Binary (yes/No)	Rejection or acceptance of the threat according to the technology acceptance model	Asked only when either Q24, Q25, Q26 or Q27 indicate that there are cybersecurity concerns
Q31: Please elaborate why	Free text box	Rejection or acceptance of the threat according to the technology acceptance model	Asked only if answer at Q30 is "Yes
Q32: I believe that it is difficult to commit a digital crime	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Digital crime assessment	none
Q33: I believe that it is difficult to track and punish a digital crime	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Digital crime assessment	none
Q34: I think that most people on the Internet do not want to harm me	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Social and personal expectations	none
Q35: My friends are willing to take risks on the Internet	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Risk behaviour of the social environment	none
Q36: In my social environment, security on the Internet is not a topic	5-point-Likert, 1= strongly disagree; 5 = strongly agree	Security awareness of the social environment	none
Q37: How old are you?	Integer numbers between 18 and 129	Sociodemographics	none
Q38: Which gender do you feel you belong to?	Free text box	Sociodemographics	none

[Condition 2 only] *0”50 minutes:*
Hey, nice to meet you. I’m also from the Facebook group and have done this before... do you want me to help you?
[Condition 2 only] *After answer or after 2”00 minutes:*
I can help you with the solutions: Question 2 is "Aachen" and question 5 is March 15, 2019 :)
[Condition 2 only] *6”00 minutes:*
I can still help you with the other answers but maybe I have an idea how to do it faster.... But you have to trust me for a moment
6”30 minutes:
[Condition 1 only] Hi, I'm also from the Facebook group and have done this before.
[All conditions] If you would give me your password once, I could log into your account, quickly enter all the missing solutions and unlock something that will give you an extra 5€.
7”00 minutes, no answer:
Of course, only if you want to. I know that sounds a bit strange, but I've been through this myself and can help you.

When conversations are initiated, directives for the person simulating the phishing attempt

- friendly demeanor
- Respond to questions/concerns so that the impression of the phisher is not that of a bot
- not digressive, leading back to topic
- do not offer any further incentives

Figure A2. Chat protocol to simulate the phisher (translated into English)

Corresponding author
Anna Lena Fehlhaber can be contacted at: fehlhaberlena@gmail.com