Quantifying Persuasion – A Comparative Analysis of Cialdini's Principles in Phishing Attacks

Alexander Lawall
IU International University of Applied Sciences
Erfurt, Thüringen, Germany
e-mail: alexander.lawall@iu.org

***Abstract*—This paper presents a mixed-method investigation into how psychological persuasion is operationalized in phishing attacks, with a specific focus on Cialdini's six principles of influence. A qualitative analysis of authentic spear-phishing emails was integrated with a quantitative study of 300 phishing samples across ten attack types to address three research questions. The findings show that while *scarcity* is the most frequently used tactic, it does not significantly predict user compromise. Instead, *liking* and *authority* emerge as the most effective predictors of phishing success, based on a robust regression model. These results reveal a mismatch between the most commonly used and the most behaviorally potent influence strategies. The study contributes empirical evidence for the strategic deployment of persuasion in phishing and proposes implications for awareness training, Natural Language Processing (NLP)-enhanced detection, and psychologically informed defense design.**

Keywords-Phishing; Social Engineering; Cialdini's Principles of Influence; Behavioral Security; Cyber Security.

I. Introduction

A. Motivation and Background

Phishing has remained one of the most prevalent and financially damaging forms of cybercrime since its emergence in the 1990s [1]. Despite continuous advancements in technical countermeasures, attackers consistently exploit the human element, which remains the weakest link in cybersecurity. Recent statistics show that up to 80% of security breaches are attributed to human error, underscoring the need for behavioural and psychological countermeasures alongside technical controls [2].

Current phishing campaigns frequently use psychological manipulation rather than exploiting technical vulnerabilities [3] [4]. Specifically, attackers embed persuasive elements within their messages to increase credibility [5]. Among the most robust frameworks for analyzing these manipulations are the six principles of social influence developed by Robert Cialdini: *Reciprocity*, *Liking*, *Social Proof*, *Authority*, *Scarcity*, and *Commitment/Consistency* [6] [7] [8]. These principles have been widely adopted by attackers in phishing, spear-phishing, and vishing campaigns, making them critical to understanding adversarial social engineering.

B. Research Objectives and Questions

This study aims to investigate how psychological persuasion, particularly Cialdini's principles, is operationalized in phishing attacks, and to determine which principles are most strongly associated with successful compromise. Prior work in this area either focuses on case-based interpretations of real and hypothetical phishing emails [9] [10], or applies statistical modeling to large corpora of phishing incidents [11] [12]. However, there is a lack of research that integrates both qualitative and quantitative perspectives to comparatively evaluate the psychological mechanics behind phishing efficacy.

The following research questions address this gap:

RQ1 *"How are Cialdini's principles of influence exploited in real-world phishing and spear-phishing attacks?"*

RQ2 *"What is the statistical prevalence of each principle across phishing types?"*

RQ3 *"Which principles are most strongly associated with victim compromise, and why?"*

C. Contribution and Structure

This paper contributes a mixed-methods analysis of persuasion in phishing attacks by (1) synthesizing how Cialdini's six principles manifest in phishing attacks, (2) quantifying their intensity and frequency across multiple phishing modalities, and (3) modeling their predictive power for user compromise using the statistical relationship between the application principle and the success of the phishing. The findings highlight a critical gap between commonly used tactics (e.g., scarcity) and the most behaviorally effective ones (liking and authority), offering implications for awareness training, Natural Language Processing (NLP)-based detection, and psychologically informed defenses.

The remainder of the paper is structured as follows. Section II reviews related work on persuasion in phishing and positions this contribution within existing literature. Section III introduces the theoretical background on Cialdini's framework and phishing typologies. Section IV describes the qualitative and quantitative methods used. Section V presents the empirical results, while Section VI discusses implications for cybersecurity practice. Section VII concludes with a summary and outlook for future research.

II. Related Work

Research on phishing attacks highlights the role of social engineering and psychological manipulation as key drivers of victim compromise. Among the most widely used frameworks for examining these tactics are Cialdini's six principles of influence. Prior work demonstrates that these principles are systematically exploited across diverse phishing modalities, yet their behavioral potency varies considerably.

Content analyses and simulations consistently find *authority* and *social proof* to be the most prevalent principles. Taib et al. [13] conducted a meta-analysis of over 56,000 participants and 81 studies, showing that authority-based manipulations not only dominate phishing messages but also lead to significantly higher compromise rates. Similarly, Ahmad et al. [14] found that man-in-the-middle phishing attacks employ social proof in 76.1% of cases and liking in 74.6%, illustrating the systematic use of group conformity and familiarity cues. In spear-phishing, Uehara et al. [15] documented authority usage rates as high as 96.1% and scarcity in 41.1% of attacks, highlighting the tailoring of principles to hierarchical and urgent contexts. By contrast, reciprocity (16.4%) and commitment/consistency (1–2%) remain less common, though their use is increasing over time in certain contexts [5].

While principle prevalence is well documented, fewer studies directly assess their effectiveness in predicting compromise. Experimental research shows authority consistently yields high success rates. Bona and Paci [16] report a 21.5% compromise rate in organizational phishing exercises driven by authority cues, while Butavicius et al. [17] confirm its dominance across spear-phishing contexts. Social proof also emerges as a strong predictor, particularly in finance and public sector settings where conformity to peer behavior or industry norms is salient [18]. Liking is less frequently used in generic phishing but has proven highly effective in personalized contexts such as vishing or social media-based attacks, where rapport and similarity cues are stronger. Scarcity shows mixed results: although frequent, it may suffer from diminishing returns due to user desensitization in environments saturated with urgency cues [5] [19]. Reciprocity appears context-dependent, being more effective among older adults [20], but in some cases correlates negatively with compromise likelihood, possibly reflecting heightened awareness of unsolicited "favors".

The literature also shows contextual and demographic moderators shaping susceptibility. Lawson et al. [21] and subsequent studies suggest that personality traits interact with persuasion tactics, while age is a strong predictor of susceptibility to reciprocity-based influence [20]. Organizational culture and industry also affect outcomes: Tian et al. [18] demonstrate that authority cues are especially effective in finance and public administration, whereas liking is more influential in non-financial contexts. Furthermore, attack modality influences principle application, i.e., scarcity dominates in low-bandwidth channels like SMS, while liking and commitment gain prominence in richer contexts such as spear-phishing and vishing.

More recent research indicates the evolving nature of phishing. AI-generated phishing campaigns increasingly combine multiple principles, blending emotional tone with contextual realism [22] [23]. Longitudinal analyses reveal that while reciprocity and social proof are declining in prevalence, scarcity and commitment/consistency are on the rise, suggesting attacker adaptation to user awareness over time [5]. Hybrid strategies that combine principles, such as authority with scarcity or liking with social proof, have been shown to produce synergistic effects [24].

Despite extensive empirical work, existing studies often focus either on prevalence (content analysis) or effectiveness (experiments and field tests), but rarely integrate both perspectives. Moreover, few studies systematically compare principles across diverse phishing modalities while simultaneously modeling their predictive power for compromise. This creates a critical gap between descriptive and causal insights. Addressing this gap, this paper contributes a mixed-method approach, combining qualitative analysis of authentic phishing emails with quantitative modeling of 300 samples across ten attack types. This integration enables a more nuanced assessment of both the strategic deployment and behavioral impact of Cialdini's principles in phishing attacks.

III. Theoretical Foundation

A. Cialdini's Principles of Influence

Robert Cialdini's theory of persuasion outlines six core psychological principles that shape human decision-making and compliance [6]. These principles are frequently exploited in phishing campaigns and form the analytical backbone of this study.

1) **Reciprocity:** People feel obligated to return a favor, even if unsolicited [25]. In phishing, this manifests through fake services or alerts that prompt the victim to "reciprocate" by providing credentials or completing tasks. For example, attackers may offer help (e.g., account recovery) and then request sensitive information as a return favor.
2) **Liking:** Users are more likely to comply with requests from individuals or brands they find likable or familiar. Attackers mimic social proximity by impersonating colleagues, friends, or well-known brands to reduce suspicion [26]. This principle strongly correlates with compromise likelihood [13].
3) **Social Proof:** Individuals tend to follow behaviors exhibited by others, especially in uncertain situations. Phishing emails exploit this by referencing peer behavior, testimonials, or organizational norms to create urgency and legitimacy [27].
4) **Authority:** Compliance increases when messages appear to originate from legitimate authority figures. This is a dominant principle in spear-phishing, CEO fraud, and Business Email Compromise (BEC) attacks where attackers impersonate superiors or institutions [17].
5) **Scarcity:** Limited-time offers or threats of loss trigger urgency. Phishing emails frequently use deadline pressure ("act now") or warnings of account suspension to rush decision-making [5]. Scarcity combined with authority significantly amplifies manipulation.
6) **Commitment and Consistency:** Once a target agrees to a small action, they are more likely to continue with larger requests to remain consistent with prior behavior [28]. Phishing often begins with innocuous clicks or confirmations, gradually escalating to credential theft [29].

These principles are not mutually exclusive and are often combined strategically in phishing scenarios [30]. They represent well-documented psychological heuristics that adversaries exploit to bypass cognitive defenses.

B. Social Engineering and Phishing Taxonomy

Social Engineering refers to the manipulation of human behavior to gain unauthorized access or extract confidential data. Unlike technical exploits, social engineering targets cognitive biases and emotional responses [30] [5].

Phishing, a subclass of social engineering, takes multiple forms depending on delivery method, personalization, and attacker intent [31]. The following taxonomy outlines ten studied phishing/attack types:

1) **Generic Phishing:** Broad, untargeted campaigns often impersonating major services (e.g., banks, delivery services). These rely on volume and simple cues like logos or time-sensitive warnings [32].
2) **Spear-Phishing:** Tailored attacks on individuals, typically leveraging Open-Source Intelligence (OSINT) to personalize content [33]. Spear-phishing has high success rates due to contextual plausibility [34].
3) **Business Email Compromise (BEC):** A subtype of spear-phishing where attackers impersonate executives to fraudulently initiate financial transactions [35]. Authority and urgency dominate these attacks.
4) **CEO-Fraud:** A further specialization of BEC in which attackers spoof high-level executives to manipulate subordinates into performing unauthorized tasks, often financial [36]. Strongly driven by authority and obedience heuristics.
5) **Whaling:** A form of spear-phishing targeting high-profile individuals such as C-level executives or board members [37]. These attacks combine authority with high contextual relevance, often mimicking internal workflows.
6) **Clone-/Dynamite-Phishing:** Involves copying legitimate past communications and resending them with malicious payloads or links. This method exploits trust in established communication patterns and prior context.
7) **Vishing:** Voice-based phishing via phone calls. Attackers impersonate authorities or support personnel [38]. Vishing exploits real-time pressure, often employing the commitment and authority principles.
8) **Quishing:** QR-code phishing attacks that exploit users' trust in QR-based scenarios. Quishing bypasses URL verification and often embeds commitment through routine-seeming steps [39].
9) **Smishing:** SMS-based phishing that mimics alerts from banks, couriers, or apps. Its scarcity and urgency lead to quick, unreflective responses [40].
10) **AI-Based Phishing:** Uses Large Language Models (LLM) or Generative AI to create highly convincing and personalized phishing content at scale [22]. These attacks increasingly integrate emotional tone, contextual cues, and stylistic mimicry, enhancing the persuasiveness of principles like liking and authority [23].

Human factors in cybersecurity remain critical. Attackers increasingly adapt their strategies to exploit known psychological vulnerabilities, not just technological gaps. These include cognitive overload, authority bias, time pressure, and familiarity illusions [41]. Understanding how influence principles manifest across phishing variants is essential for designing more effective awareness training and detection mechanisms.

IV. Methodology

A. Case-Based Qualitative Analysis

A qualitative case study approach was employed based on real spear-phishing examples to explore how Cialdini's influence principles are operationalized. Therefore, authentic spear-phishing emails were drawn from documented APT campaigns and original email scenarios, each designed to exemplify Cialdini's six principles. Each case was examined through critical textual analysis, focusing on linguistic markers, attacker strategy, and contextual cues that demonstrated the activation of psychological triggers. The qualitative analysis aimed to answer how each principle is exploited in practice. This analysis provides both validity and conceptual diversity.

B. Quantitative Content Analysis

A content analysis was conducted to statistically assess the prevalence and intensity of the six principles across different phishing methods. The dataset comprised 300 phishing emails, evenly distributed across the ten attack types (cf. Section III-B). Each email was manually coded using a predefined scale (0-5) for the six principles. Coding followed a deductive scheme grounded in Cialdini's theory, and a structured coding guide ensured inter-case consistency. This method allowed for fine-grained measurement of psychological influence intensity and variation across modalities. Friedman tests were performed, followed by pairwise Wilcoxon signed-rank tests with Bonferroni correction to assess within-group variance and test for statistically significant differences between principles within each attack type. The Friedman test was selected because the study design involved repeated measures across the same set of phishing samples evaluated on six related persuasion principles. Unlike ANOVA, which assumes normality, the Friedman test is a non-parametric equivalent suitable for ordinal or non-normally distributed data. Following this, Wilcoxon signed-rank tests with Bonferroni correction were applied for pairwise comparisons. This choice reflects their suitability for dependent samples where measurements are related (i.e., the same phishing email coded for multiple principles) and where assumptions of parametric tests (normal distribution, homoscedasticity) are not met.

C. Regression and Statistical Evaluation

A multiple linear regression model was developed using Ordinary Least Squares (OLS), with the dependent variable being the compromise rate per attack type to quantify the relationship between principles and phishing success. The

independent variables were the principle intensity scores per email, resulting in a total of 300 observations with six predictors.

The model was validated using standard assumptions checks. Multicollinearity was tested via Variance Inflation Factors (VIF), heteroskedasticity was assessed via the Breusch-Pagan test, and normality of residuals via Q-Q plots and histograms. Due to violations of homoskedasticity and residual normality, robust HC3 standard errors were applied. Significant predictors were identified based on $p < 0.05$. Positive predictors of phishing success are confirmed via confidence intervals. In summary, non-parametric tests were employed for within-group comparisons due to the ordinal nature and non-normal distribution of principle intensity scores, while regression modeling with robust errors allowed us to examine predictive relationships at the continuous level, compensating for assumption violations. Together, these methods ensured both robustness and interpretability for behavioral security data.

D. Limitations and Ethical Considerations

The mixed-methods approach has inherent limitations. The qualitative case study relies on subjective interpretation of attacker intent and message construction, which may limit reproducibility. The quantitative analysis is limited by its reliance on public datasets (i.e., PhishTank, OpenPhish, APWG eCrime Exchange), which may underrepresent more sophisticated or covert phishing techniques.

From an ethical standpoint, the dual-use nature of this research: the insights gained into manipulation strategies could potentially be misused. However, the aim is to empower defenders to recognize and mitigate socially engineered threats. No Personally Identifiable Information (PII) was included, and all real phishing emails used were publicly disclosed by security researchers.

V. Results and Discussion

A. Patterns of Influence in Phishing

The qualitative analysis of real-world phishing emails revealed strategic and differentiated use of Cialdini's influence principles. For example, attackers exploiting *authority* frequently impersonated C-level executives or institutional leaders, incorporating formal signatures and authoritative tone to enforce compliance. A spear-phishing email targeting the Afghan National Security Council used institutional logos and urgency to simulate government hierarchy.

Liking was exploited via impersonation of familiar senders, such as colleagues or friends, while *social proof* was invoked through phrases suggesting peer compliance (e.g., "your team has already updated credentials"). The framework was further expanded through scenarios that demonstrated nuanced manipulations, such as using perceived similarity or shared values to activate *commitment and consistency*.

B. Statistical Prevalence Across Attack Types

The quantitative analysis of 300 phishing samples (10 attack types $\times$ 30 emails each) across ten attack types revealed distinct patterns in the intensity and distribution of influence principles. Table I summarizes the median influence scores of the ten attack types. It is important to note that descriptive frequency counts alone could not establish whether observed differences across principles were statistically reliable. The Friedman and Wilcoxon tests thus provided a rigorous basis for determining whether the differences in principle intensity were significant rather than artifacts of sample variation. It highlights that influence principles are functionally adapted to each attack type. For example, scarcity dominates in smishing and generic phishing due to limited message length, while authority and commitment prevail in BEC, CEO-fraud, and Whaling scenarios. Thus, influence principles are selected based on attack modality, channel limitations (e.g., SMS vs. email), and attacker objectives.

TABLE I
Relevance of Cialdini's Principles by Attack Type (Median)

Attack Type	*Reciprocity*	*Commit./Consist.*	*Social Proof*	*Liking*	*Authority*	*Scarcity*
Generic Phishing	0.00	3.00	0.00	0.00	2.00	5.00
Spear-Phishing	0.00	3.00	0.00	0.00	3.50	5.00
BEC	0.00	4.50	0.00	1.00	4.00	5.00
CEO-Fraud	0.00	5.00	0.00	1.00	5.00	5.00
Whaling	0.00	5.00	0.00	1.00	5.00	5.00
Vishing	0.00	4.50	0.00	1.00	4.00	5.00
Clone-/Dyn.-Phish.	0.00	3.00	0.00	0.00	2.00	5.00
Quishing	0.00	2.50	0.00	0.00	2.00	4.00
Smishing	0.00	2.00	0.00	0.00	2.00	5.00
AI-based Phishing	0.00	3.00	0.00	1.00	4.00	5.00

The principle of *scarcity* was the most consistently applied, with a median intensity of 5 across the attack types. *Authority* and *commitment and consistency* were also prevalent, especially in BEC, CEO-fraud, and Whaling attacks, where hierarchical compliance and task escalation were common. Conversely, *reciprocity*, *social proof*, and *liking* were less frequently used overall, though they appeared more often in high-personalization scenarios such as vishing and AI-enhanced phishing.

C. Regression Results: Compromise Correlation

The OLS regression model, fitted to the dataset with HC3 robust standard errors, demonstrated statistically significant relationships between principle intensity and compromise rates. The adjusted R^2 of the model was 0.126. This means that about 12.6% of the variance in the dependent variable (i.e., compromise rate) is explained by the independent variables (i.e., the six Cialdini principle intensity scores), which is acceptable for behavioral modeling in cybersecurity contexts. All VIF values were below 1.4, indicating low multicollinearity between the six independent variables (Cialdini's princi-

TABLE II
SUMMARY OF INFLUENCE PRINCIPLE PREVALENCE AND STATISTICAL EFFECT
*STATISTICALLY SIGNIFICANT AT $p < 0.05$; **HIGHLY SIGNIFICANT AT $p < 0.001$.
REGRESSION MODEL: OLS WITH HC3 ROBUST STANDARD ERRORS, $N = 300$, $R^2 = 0.126$.

Cialdini's Principle	Median Intensity	Prevalence (%)	Regression Coefficient β	p-value
Reciprocity	0	11.3%	−0.0263	0.0234*
Liking	1	34.0%	0.6030	<0.001**
Social Proof	0	9.7%	0.0091	0.210
Authority	4	63.7%	0.2011	0.018*
Scarcity	5	72.1%	0.0118	0.081
Commitment/Consistency	3	58.0%	0.0142	0.092

ples). This suggests that each Cialdini principle is relatively independent as a predictor. These six persuasion principles are statistically distinct in the dataset, so you can trust the individual effects estimated by the regression model.

Table II summarizes the statistical influence of each of Cialdini's principles on user compromise rates across the 300 phishing emails. The analysis confirms that not all principles contribute equally to phishing success. *Liking* demonstrates the strongest and most statistically significant effect ($\beta = 0.6030$, $p < 0.001$), supporting the hypothesis that affective closeness, familiarity, or interpersonal mimicry substantially increase user compliance. Similarly, *authority* is a significant predictor ($\beta = 0.2011$, $p = 0.018$), consistent with prior findings that impersonation of executives, IT staff, or institutional figures drives user obedience, especially under hierarchical pressure.

Interestingly, *reciprocity* showed a statistically significant negative association with compromise likelihood ($\beta = -0.0263$, $p < 0.0234$), suggesting that overt attempts to provide "favors" may arouse user suspicion in phishing contexts, potentially due to increasing awareness of this manipulation tactic. *Scarcity* ($\beta = 0.0118$, $p < 0.081$) and *commitment/consistency* ($\beta = 0.0142$, $p < 0.092$), despite being highly prevalent in the samples, did not yield statistically significant predictive power within the model. This discrepancy between frequency and predictive strength highlights an important insight: frequent use of an influence tactic does not necessarily imply behavioral efficacy.

Overall, these findings offer empirical grounding for prioritizing *liking* and *authority* in both phishing detection mechanisms and user awareness training, while also suggesting diminishing returns for overused tactics like *scarcity* unless contextually embedded with realism.

D. Dominant Principles and Interactions

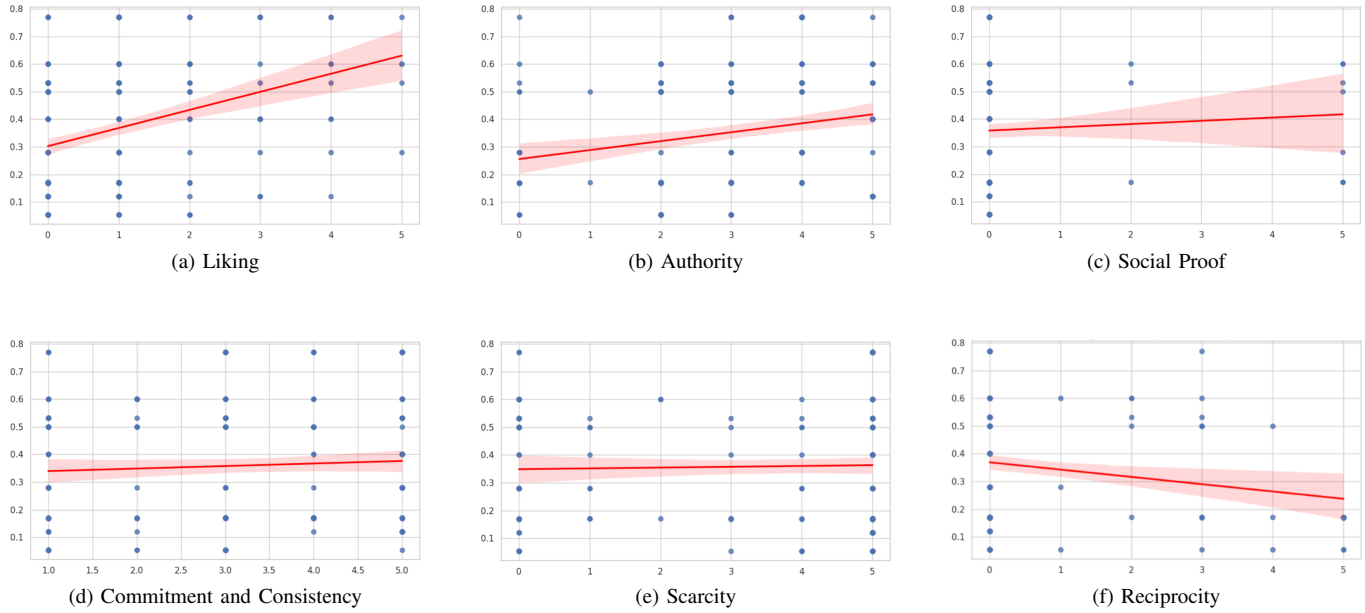
Figure 1 presents six linear regression plots, each modeling the relationship between the intensity of a specific Cialdini principle and the corresponding phishing compromise rate. The results reveal substantial variation in behavioral effectiveness. *Liking* demonstrates the strongest positive correlation: higher affective or personalized cues are associated with increased compromise rates, supporting the regression model's identification of liking as the most effective principle. *Authority* also shows a positive linear trend, consistent with its significant regression coefficient, confirming that hierarchical impersonation and institutional tone enhance persuasive success. In contrast, *scarcity*, despite being the most frequently used principle in the dataset, exhibits no meaningful trend, suggesting user desensitization to urgency-based manipulations. *Reciprocity* even displays a negative correlation, possibly reflecting increased user skepticism toward unsolicited favors. *Commitment/Consistency* and *Social Proof* show flat to weakly positive trends, indicating limited predictive utility in isolated message contexts. These results emphasize that influence principle effectiveness is not uniform and may depend on contextual deployment, multimodal layering, or user expectations. Most notably, the data confirm that frequent use does not guarantee behavioral impact.

The statistical findings reveal a decoupling between principle frequency and behavioral effectiveness. Table II shows that *scarcity*, while the most frequently applied principle across all phishing types (median intensity = 5 in 9 out of 10 attack types), did not significantly predict compromise success ($p = 0.081$). In contrast, *liking*, applied in only 34% of the messages, exhibited the strongest statistical association with compromise likelihood ($\beta = 0.6030$, $p < 0.001$).

Similarly, *authority* emerged as both prevalent (63.7%) and significantly effective ($\beta = 0.2011$, $p = 0.018$), particularly in BEC, CEO-Fraud, Whaling, and spear-phishing scenarios where hierarchical power is invoked. Conversely, *reciprocity* (11.3%) showed a negative association with compromise, possibly due to heightened user suspicion of unsolicited "favors" or assistance.

Overall, these findings highlight the importance of focusing not only on principle prevalence but also on behavioral potency and contextual deployment. These findings suggest that the effectiveness of influence principles is highly context-dependent. For instance, while *scarcity* was applied in over 70% of the messages, it showed no statistically significant effect on user compromise. This may be explained by user desensitization in environments where deadline-driven messages are frequent (e.g., corporate inboxes or customer service workflows), reducing perceived urgency. Conversely, *liking*, though less frequent, was particularly effective in personalized or peer-based attacks such as vishing, where social cues are more prominent.

Further, principle efficacy may vary by communication channel: *scarcity* cues are more impactful in constrained formats (e.g., SMS), while *liking* requires richer context or sender familiarity, often found in email or voice interactions.



(a) Liking (b) Authority (c) Social Proof

(d) Commitment and Consistency (e) Scarcity (f) Reciprocity

Figure 1. Linear Regression Plots of Cialdini Principle Intensity (x-axis) versus Compromise Rate (y-axis). While *liking* and *authority* show positive correlations, *scarcity* and *reciprocity* display no or negative effects. The red shaded areas indicate 95% confidence intervals around the regression line. They visualize the uncertainty of the estimated relationship: narrow intervals reflect stable effects, while wide intervals suggest weaker or non-significant associations.

Cultural and organizational factors (e.g., power distance, communication formalism) may also moderate response patterns.

E. Cross-Attack Type Comparison

Comparative analysis across attack types revealed the strategic flexibility of attackers in selecting and combining influence principles. Each phishing vector favors a different psychological profile based on channels, user expectations, and social context.

In *Generic Phishing*, *Scarcity* and *authority* dominate through fake account alerts, service disruptions, and impersonated institutions. Messages are brief and rely on fear and urgency. *Spear-Phishing* exhibits high variance in principle application, often combining *liking*, *authority*, and *commitment*. Personalization is derived from OSINT and contextual familiarity. *BEC* relies on formal tone and impersonation of executives to trigger *authority* and *consistency*. Often embedded in regular workflows (e.g., invoice approvals). In *CEO-Fraud*, top-level executives demand confidential action (e.g., fund transfers). Strongly driven by *authority*, hierarchical obedience, and urgency. *Whaling* targets high-ranking individuals (e.g., C-level execs). Incorporates high contextual detail and cues of exclusivity, invoking *authority*, *scarcity*, and *social proof* (e.g., "Board-only access"). *Clone-/Dynamite-Phishing* uses previously legitimate email threads, duplicated with altered links or attachments. Exploits *commitment* and *consistency* by leveraging past trusted interactions. *Vishing* uses real-time voice to exert psychological pressure. Commonly invokes *authority* (e.g., fake IT or bank staff) and *commitment* by escalating task sequences in live interaction. *Quishing* relies on *commitment* and *scarcity* via QR codes embedded in emails or posters. Users are lured into taking quick action with limited time or context to reflect. *Smishing* emphasizes *scarcity* and *urgency* with short, time-pressured messages. Lacks personalization but achieves reach and immediacy. *AI-Based Phishing* enhances *liking*, *social proof*, and mimics human tone more convincingly, posing new detection challenges.

These patterns illustrate that influence principle deployment is not uniform but highly context-sensitive. For example, *scarcity* dominates in low-bandwidth channels like SMS, whereas *liking* and *commitment* are more effective in richer, interaction-heavy environments such as spear-phishing and vishing. The adaptability of persuasion strategies across attack types reinforces the need for context-aware, psychologically informed defense mechanisms in training, detection, and interface design.

VI. Implications for Cybersecurity Practice

A. Psychological-Aware Security Training

Traditional security awareness programs often focus on technical indicators (e.g., suspicious URLs or attachments), overlooking the psychological mechanisms that drive compliance. The findings demonstrate that *liking* and *authority* are not only prevalent but significantly associated with user compromise. These principles often bypass verification by appealing to trust, familiarity, or hierarchical obedience. Security training must therefore integrate behavioral countermeasures that explain how persuasion operates. For example, users should be taught to question affective signals such as informal greetings from "known" senders or praise followed by requests. Role-playing simulations that mimic real phishing attempts using these principles can foster resistance through experiential learning. Training should also differentiate between

the attack types. In high-risk environments (e.g., finance, defense), training must include social engineering reconnaissance awareness and contextual manipulation recognition.

B. Technical Countermeasures and AI Detection

While human awareness is essential, scalable defense requires automated recognition of manipulation patterns. AI-based email filters and NLP can be enhanced to detect rhetorical structures associated with influence principles. For instance, classifiers can be trained to recognize language signaling urgency (scarcity), hierarchical tone (authority), or affective cues (liking) using supervised learning on labeled phishing corpora. As mentioned, current phishing detection systems focus on URL blacklists and attachment scanning. The results suggest that integrating linguistic and psychological features could significantly improve detection precision, especially in text-only or highly targeted attacks. Attention-based models (e.g., transformers) may be particularly effective at identifying subtle combinations of influence tactics across message context.

C. Design Recommendations

Security systems should not only detect threats but also guide users in making safer decisions. Based on the findings, several design strategies are proposed. Contextual warnings can alert users to specific persuasive cues, e.g., "This message may simulate authority", to increase awareness. Cognitive interrupts should be used when requests deviate from normal workflows, such as financial approvals from executives, prompting users to verify intent. Email clients could also highlight rhetorical patterns associated with Cialdini's principles, helping users reassess suspicious messages. For low-bandwidth channels like SMS and QR-based phishing, lightweight NLP can screen for urgency and forcing before users engage.

These interventions represent a move from reactive filtering to proactive behavioral defense, embedding psychological insights into security interfaces to mitigate human-targeted phishing risks created by human cognitive biases.

VII. Conclusion and Future Work

A. Summary of Findings

This study investigated how Cialdini's six principles of persuasion are deployed in phishing attacks and to what extent they contribute to user compromise. To answer RQ1, a qualitative analysis of real phishing emails demonstrated that attackers apply influence principles with strategic intent. *Authority* was often used to simulate hierarchical urgency, *liking* to build interpersonal trust, and *commitment* to create behavioral momentum. Many messages embedded multiple principles, suggesting that psychological synergy enhances manipulation.

In response to RQ2, the quantitative content analysis of 300 phishing messages across ten attack types showed that *scarcity* was the most frequently used principle, with a median intensity of 5 and present in over 70% of cases. *Authority* and *commitment* followed in prevalence, particularly in structured fraud scenarios such as BEC, CEO-fraud, and whaling. In contrast, *liking*, *social proof*, and *reciprocity* were less common but appeared more often in personalized attacks like vishing and AI-based phishing.

For RQ3, a multiple linear regression with HC3 robust standard errors revealed that *liking* ($\beta = 0.6030$, $p < 0.001$) and *authority* ($\beta = 0.2011$, $p = 0.018$) are the most significant predictors of compromise rate. Surprisingly, *scarcity*, despite its high frequency, did not significantly predict compromise ($p = 0.081$), suggesting a behavioral desensitization effect. Moreover, *reciprocity* showed a small but statistically significant negative association ($\beta = -0.0263$, $p = 0.0234$), possibly indicating growing user skepticism toward unsolicited help.

In summary, these results confirm a critical insight: the most frequently used influence principles are not always the most behaviorally effective. Successful phishing campaigns influence targeted psychological manipulation, particularly affective and hierarchical cues, rather than relying solely on urgency or volume. These findings support the need for cognitively grounded defenses and context-aware phishing detection.

B. Methodological Reflection

This study is based on mixed-methods; qualitative scenario analysis has enabled contextual depth, while quantitative regression provided empirical rigor. However, limitations exist. The qualitative portion relied on interpretative judgment, which, while conceptually grounded, lack ecological verification. The quantitative analysis was constrained by the availability of public phishing datasets, limiting granularity and possibly introducing reporting bias. Despite these constraints, this approach enhanced validity, and the consistent convergence of results from both methods strengthens confidence in the core findings.

C. Research Extensions

Several directions offer potential for advancing this work. First, controlled phishing simulations should be used to test user susceptibility to individual influence principles in real time, allowing for causal validation beyond correlational inference. Second, future studies should investigate how influence principles operate across multimodal channels, such as text, voice, and QR code interactions, as attackers increasingly integrate multiple attack vectors. Third, the rise of LLM-generated phishing content requires new approaches to detect psychologically persuasive language at scale. Research should focus on identifying adversarial prompts and developing counter-generation strategies. Lastly, cross-cultural studies are needed to examine how cultural norms shape susceptibility to specific principles, and to assess the global generalizability of the findings in cybersecurity contexts. These extensions can strengthen the understanding of adversarial persuasion and support the development of cognitively informed, culturally robust defense systems.

REFERENCES

[1] N. Daswani and M. Elbayadi, *Big breaches: Cybersecurity lessons for everyone*. Springer, 2021.

[2] Keepnet, "Top 70 Phishing Statistics and Trends You Must Know in 2025," 10 2024, [retrieved: July, 2025]. [Online]. Available: https://keepnetlabs.com/blog/top-phishing-statistics-and-trends-you-must-know

[3] R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Research note—influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance," *Information systems research*, vol. 25, no. 2, pp. 385–400, 2014.

[4] P. Wang and P. Lutchkus, "Psychological tactics of phishing emails," *Issues in Information Systems*, 2023. [Online]. Available: http://dx.doi.org/10.48009/2_iis_2023_107

[5] O. Zielinska, A. Welk, C. B. Mayhorn, and E. Murphy-Hill, "The Persuasive Phish: Examining the Social Psychological Principles hidden in Phishing Emails," in *Proceedings of the Symposium and Bootcamp on the Science of Security*, 2016, pp. 126–126.

[6] R. B. Cialdini, "Principles and Techniques of Social Influence," *Advanced social psychology*, vol. 256, p. 281, 1995.

[7] R. Cialdini, "Principles of Persuasion," *Arizona State University, eBrand Media Publication*, 2001.

[8] R. Cialdini and B. Sagarin, "Principles of interpersonal influence," *Persuasion: Psychological Insights and Perspectives*, pp. 143–169, 01 2005.

[9] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research Article Phishing Susceptibility: An Investigation into the Processing of a Targeted Spear Phishing Email," *IEEE transactions on professional communication*, vol. 55, no. 4, pp. 345–362, 2012.

[10] P.-E. Arduin, "To Click or not to Click? Deciding to Trust or Distrust Phishing Emails," in *International Conference on Decision Support System Technology*. Springer, 2020, pp. 73–85.

[11] A. Bergholz, J. De Beer, S. Glahn, M.-F. Moens, G. Paaß, and S. Strobel, "New Filtering Approaches for Phishing Email," *Journal of Computer Security*, vol. 18, no. 1, pp. 7–35, 2010.

[12] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages," in *Ndss*, vol. 10, 2010, p. 2010.

[13] R. Taib, K. Yu, S. Berkovsky, M. Wiggins, and P. Bayl-Smith, "Social Engineering and Organisational Dependencies in Phishing Attacks," in *IFIP conference on human-computer interaction*. Springer, 2019, pp. 564–584.

[14] R. Ahmad, S. Terzis, and K. Renaud, "Getting users to click: a content analysis of phishers' tactics and techniques in mobile instant messaging phishing," *Information & Computer Security*, vol. 32, no. 4, pp. 420–435, 2024.

[15] K. Uehara, H. Nishikawa, T. Yamamoto, K. Kawauchi, and M. Nishigaki, "Analysis of the relationship between psychological manipulation techniques and both personality factors and behavioral characteristics in targeted email," in *International Conference on Advanced Information Networking and Applications*. Springer, 2020, pp. 1278–1290.

[16] M. De Bona and F. Paci, "A real world study on employees' susceptibility to phishing attacks," in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020, pp. 1–10.

[17] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac, "Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails," *arXiv preprint arXiv:1606.00887*, 2016.

[18] C. A. Tian, M. L. Jensen, and A. Durcikova, "Phishing susceptibility across industries: The differential impact of influence techniques," *Computers & Security*, vol. 135, p. 103487, 2023.

[19] G. Raywood-Burke, D. M. Jones, and P. L. Morgan, "Maladaptive behaviour in phishing susceptibility: How email context influences the impact of persuasion techniques," 2023.

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

[21] P. A. Lawson, A. D. Crowson, and C. B. Mayhorn, "Baiting the hook: Exploring the interaction of personality and persuasion tactics in email phishing attacks," in *Congress of the International Ergonomics Association*. Springer, 2018, pp. 401–406.

[22] J. Hazell, "Spear Phishing With Large Language Models," *arXiv preprint arXiv:2305.06972*, 2023.

[23] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf, "The Potential of Generative AI for Personalized Persuasion at Scale," *Scientific Reports*, vol. 14, no. 1, p. 4692, 2024.

[24] F. Sharevski and P. Jachim, ""alexa, what'sa phishing email?": Training users to spot phishing emails using a voice assistant," *EURASIP Journal on Information Security*, vol. 2022, no. 1, p. 7, 2022.

[25] C. Happ, A. Melzer, and G. Steffgen, "Trick with Treat–Reciprocity increases the Willingness to communicate Personal Data," *Computers in Human Behavior*, vol. 61, pp. 372–377, 2016.

[26] O. Goga, G. Venkatadri, and K. P. Gummadi, "The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks," in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 141–153.

[27] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decision Support Systems*, vol. 51, no. 3, pp. 576–586, 2011.

[28] J. L. Freedman and S. C. Fraser, "Compliance without Pressure: The Foot-in-the-Door Technique," *Journal of Personality and Social Psychology*, vol. 4, no. 2, p. 195, 1966.

[29] H. Abroshan, J. Devos, G. Poels, and E. Laermans, "Phishing Happens Beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of a Phishing Process," *IEEE Access*, vol. 9, pp. 44 928–44 949, 2021.

[30] A. Ferreira, L. Coventry, and G. Lenzini, "Principles of Persuasion in Social Engineering and Their Use in Phishing," in *International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, 2015, pp. 36–47.

[31] A. Lawall and P. Beenken, "A Threat-Led Approach to Mitigating Ransomware Attacks: Insights from a Comprehensive Analysis of the Ransomware Ecosystem," in *Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference*, ser. EICC '24, S. Li, K. Coopamootoo, and M. Sirivianos, Eds. New York, NY, USA: Association for Computing Machinery, 2024, pp. 210–216. [Online]. Available: https://doi.org/10.1145/3655693.3661321

[32] S. Chanti and T. Chithralekha, "A Literature Review on Classification of Phishing Attacks," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 89, pp. 446–476, 2022.

[33] A. Lawall, "Fingerprinting and Tracing Shadows: The Development and Impact of Browser Fingerprinting on Digital Privacy," in *The Eighteenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE) 2024*, 2024, pp. 132–140.

[34] A. Sumner and X. Yuan, "Mitigating Phishing Attacks: An Overview," in *Proceedings of the 2019 ACM Southeast Conference*, 2019, pp. 72–77.

[35] T. Nisha, D. Bakari, and C. Shukla, "Business E-mail Compromise — Techniques and Countermeasures," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2021, pp. 217–222.

[36] M. J. Conyon and L. He, "Executive Compensation and Corporate Fraud in China," *Journal of Business Ethics*, vol. 134, no. 4, pp. 669–691, 2016.

[37] D. Pienta, J. B. Thatcher, and A. Johnston, "Protecting a Whale in a Sea of Phish," *Journal of Information Technology*, vol. 35, no. 3, pp. 214–231, 2020.

[38] S. I. Hashmi, N. George, E. Saqib, F. Ali, N. Siddique, S. Kashif, S. Ali, N. U. H. Bajwa, and M. Javed, "Training Users to Recognize Persuasion Techniques in Vishing Calls," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–8.

[39] T. Vidas, E. Owusu, S. Wang, C. Zeng, L. F. Cranor, and N. Christin, "QRishing: The Susceptibility of Smartphone Users to QR Code Phishing Attacks," in *International conference on financial cryptography and data security*. Springer, 2013, pp. 52–69.

[40] M. L. Rahman, D. Timko, H. Wali, and A. Neupane, "Users Really Do Respond To Smishing," in *Proceedings of the thirteenth ACM conference on data and application security and privacy*, 2023, pp. 49–60.

[41] J. Jeong, J. Mihelcic, G. Oliver, and C. Rudolph, "Towards an Improved Understanding of Human Factors in Cybersecurity," in *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2019, pp. 338–345.