"I can say I'm John Travolta...but I'm not John Travolta": Investigating the Impact of Changes to Social Media Verification Policies on User Perceptions of Verified Accounts

Carson Powers, Nickolas Gravel, and Christopher Pellegrini, *Tufts University;*
Micah Sherr, *Georgetown University;* Michelle L. Mazurek, *University of Maryland;*
Daniel Votipka, *Tufts University*

https://www.usenix.org/conference/soups2024/presentation/powers

This paper is included in the Proceedings of the
Twentieth Symposium on Usable Privacy and Security.

August 12–13, 2024 • Philadelphia, PA, USA

978-1-939133-42-7

“I can say I’m John Travolta . . . but I’m not John Travolta.”[*] Investigating the Impact of Changes to Social Media Verification Policies on User Perceptions of Verified Accounts

Carson Powers*, Nickolas Gravel*, Christopher Pellegrini*, Micah Sherr**
Michelle L. Mazurek†, and Daniel Votipka*
*Tufts University; †University of Maryland; **Georgetown University
{carson.powers,nickolas.gravel,christopher.pellegrini,daniel.votipka}@tufts.edu
mmazurek@cs.umd.edu; msherr@cs.georgetown.edu

Abstract

Until recently, almost all social media platforms verified the identities behind notable accounts. Prior work showed users understood this process. However, Twitter/𝕏’s switch to an open, less rigorous verification process represented a significant policy shift. We conduct a U.S. Census-representative survey to investigate how this and subsequent verification changes across social media impact users’ verification perceptions. We find most users generally recognize the changes to Twitter/𝕏’s policy, though many still believe Twitter/𝕏 verifies account holders’ true identities. However, users are less aware of subsequent Facebook verification changes. We also find platforms’ verification differences do not impact user perceptions of posted content credibility.

Finally, we investigate hypothetical verification policies. We find participants are more likely to perceive posts from verified accounts as credible when only notable accounts are eligible and government document review is required. Payment did not affect credibility decisions, but participants felt strongly that payment for verification was unacceptable.

1 Introduction

Most social media sites, such as Twitter/𝕏[1], Facebook, TikTok, and LinkedIn, support some form of account verification. Each platform reviews accounts [42], then adds a badge (e.g., ✓) next to the *verified account’s* (VA)[2] username to signal the verification process has been completed. VAs were introduced to help users differentiate between accounts belonging to the entity named (often a celebrity or account of public interest) and parodies or impostors [74]. Twitter introduced VAs in 2009 following a rise in impostor accounts [67], and other platforms followed suit [23, 34, 42, 62, 64, 72]. With the rise of disinformation on social media, the value of determining a post’s true source is growing [29, 36–38, 49, 59]. This challenge is exacerbated in emergencies, when users look to social media for real-time information [7, 32, 41, 47, 76]. During terrorist and active-shooter events [4, 7] and natural disasters [41, 46, 54], users look to local authorities, such as police and fire departments, for safety information. Without rigorous account verification, users may trust false information with life threatening consequences [28].

While there is some evidence suggesting users equate account verification with credibility [44], other work has shown, in isolation, users correctly understood the verification badge only indicated authenticity [75]. However, recent changes to verification policies may muddle verification’s purpose. First, the social media ecosystem has splintered, with new and niche platforms growing (e.g., TikTok, Truth Social, etc.). While verification is similar across platforms, some subtle differences should impact the correct interpretation of VAs.

Additionally, some of the largest existing platforms have made significant policy changes. Most notably, Twitter dramatically changed its verification policy after being acquired by Elon Musk in October 2022. Prior to the purchase, Twitter verified notable users’ accounts (e.g., celebrities and public figures or organizations) by requiring proof of identity via a government-issued ID [74]. Twitter then made verification available to any user for a monthly $8[3] subscription fee, and swapped government ID for a verified phone number [73]. This transition was tumultuous, with abrupt changes regularly covered in the media [10, 21, 30, 33, 45, 48]. Some users

[*] The full quote by a participant asked what verification policy changes they would suggest was, “I would require a photo ID. I can say I’m John Travolta and I can give you my email address (which can be almost anything) to confirm me, but I’m not John Travolta.”

[1] Since Twitter’s rebranding to 𝕏 occurred after our survey, we will use “Twitter” in the remainder of the paper.

[2] Terms differ by platform. For consistency, we refer to accounts that have undergone some form of authentication as verified accounts (VAs).

[3] $12 if signing up in-app to account for Apple’s/Google’s service charges.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

took advantage of the new policy to establish impostor accounts [50,66]. To less fanfare, Facebook also adjusted its verification policy, allowing anyone to obtain a VA for a fee, but maintaining the requirement for ID verification, and LinkedIn made verification slightly more open without adding a fee.

We seek to assess the impact of these policy changes on user perceptions of VAs on Facebook and Twitter, as well as how users think verification policies *should* work. Towards that goal, this paper considers three research questions:

RQ1: What are the verification policies used by popular social media platforms and how have they changed over time?

RQ2: What do users think account verification entails? How does it impact perceptions of posted content credibility?

RQ3: How would potential changes to verification policies impact user perceptions of posts from verified accounts and user perceptions of the policies?

RQ1 seeks to understand the VA ecosystem. Due to the fractured landscape, perceptions may vary depending on the platforms used. With the volume of media coverage and rapid policy-making during the Twitter transition, user perceptions may represent a snapshot in time, rather than an accurate depiction of current policy. To understand the impact of these changes, we must first enumerate verification policies.

To address RQ1, we captured the verification policies of eight popular social media sites from April 2022 to August 2023, noting any changes. After enumerating verification policies, we conducted a controlled experiment—using a vignette-based survey of 1600 U.S. Prolific users—to address RQ2 and RQ3 for a U.S. population using text-based social media. Participants were first shown two mock posts containing contradictory information and asked to indicate which they perceived as more credible, to test the VA's impact on their assessment of relative credibility when presented with information from similar accounts—a common challenge when assessing information during emergency events. We varied the platform (Twitter vs. Facebook) and asked participants to indicate how they believed their assigned platform defined verification. Then, we presented participants with a new verification policy and asked them to reevaluate the previously shown mock posts with this new policy in mind. We also asked participants their perceptions of the new policy.

Participants' understanding of Facebook's and Twitter's verification policies was mixed, and they were more likely to correctly perceive Facebook's policy as requiring identity verification. Participants correctly indicated Twitter's policy was open to anyone for a fee. This seems to indicate users have better understood the Twitter policy over time, compared to a similar survey conducted earlier by Xiao et al., which asked participants to identify features of verification [81]. However, participants seemed unaware of Facebook's policy, with many still believing verification was free and only for notable accounts. This is likely due to the newness of Facebook's policy change and lack of broad media coverage.

We did not observe differences in participants' assessments of posted content credibility between assigned platforms. However, after providing participants with a verification policy, they were more likely to find posts from the VA credible when government ID was required and only notable accounts were verified. Participants also perceived these policies as more acceptable (matching Xiao et al. [81]). This difference between initial assessment and re-assessment after reviewing a verification policy suggests participants do not consider the details of the policy fully when assessing posts from VAs.

Finally, while participants strongly disliked paying for verification—corroborating Xiao et al. [81]—payment did not impact participants' credibility decisions before or after reviewing the verification policy. While this indicates verification payment has no direct impact on user assessments of credibility of VAs' posts, the strong dislike of the policy may have downstream impacts that should be considered in future work, especially as several participants reported no longer trusting any verification provided by Twitter.

2 Related Work

Credibility of Online Content. Much work has investigated factors affecting user perception of online content credibility. Wineburg et al. assessed students' ability to judge online source credibility [79]. Fogg et al. found the "design look" of a website impacts perceived credibility [20]. Hilligoss and Rieh found users are more likely to find information legitimate when the source appears "official" [27]. Hassoun et al. performed a qualitative analysis of Gen Z's evaluation of online information, finding three "trust heuristics": credible information was easily accessible, neutral in tone, and "felt right." Their participants reported using number of likes and comments as a form of "crowdsourcing credibility" [25]. This mirrors previous findings that users are more likely to perceive information as credible when they believe others perceive it as credible [9,19,22,27,69], an effect called the *endorsement heuristic*. Familiarity with a source also increases perceived credibility, known as the *reputation heuristic* [43]. We build on prior work, focusing specifically on social media platforms and the effect of verified indicators.

Verification's Impact on Social Media Post Credibility. The verified indicator's purpose is to affirm an account holder's identity, not signal posted content credibility. However, humans' reliance on trust heuristics may lead to an indirect effect on perceived credibility, which may explain conflicting evidence whether users separate *authenticity* and *credibility*.

Early work by Morris et al. suggested the verified indicator highly impacts users' evaluation of credibility [44]. However, their work asked participants to list features they consider when deciding if a tweet is credible, which measures the *conscious* impact of verification badges, not the *behavioral* impact. Conversely, Vaidya et al. conducted a large-scale controlled experiment, measuring the verified indicator's effect on participants' perceptions of post credibility. They found users understood verification indicated the account holder was who they said they were, but does not add credibility to the post [75]. Dumas and Stough conducted a consumer-behavior study where participants were shown influencer-posted content. They found consumers associate VAs with celebrity more than credibility [14]. In this paper, we seek to assess whether user perceptions have changed due to changes to social media verification policies and expand beyond Twitter to consider other platforms.

Most similar to our work, Xiao et al. investigated user understanding of verified indicators on Twitter, Facebook, and TikTok in the wake of Twitter Blue [81]. They surveyed social media platforms and identified dimensions of each verification policy. Using these, they surveyed 299 U.S. adults asking their definitions of verification and whether they found Twitter's policy acceptable. They found participants were more likely to indicate payment was required for Twitter as opposed to other platforms, but most continued to *incorrectly* assume Twitter verified identities of users with verified indicators. They observed users disliked Twitter's policy because it does not verify identity and requires payment. We build on this study in several ways. First, we conducted a more in-depth review of social media platforms by investigating Musk's Twitter posts, which provide valuable context, and monitoring policies over a longer period, which captured policy changes by Meta and LinkedIn. Next, capturing a snapshot after Meta's policy changes allows a useful comparison over time between the works. We also measured how verified indicators impact perceptions of post credibility. Finally, we conduct between-subjects comparisons, randomly assigning participants to define verification for specific platforms instead of asking for general definitions, and test several possible policy designs for their impact on post credibility decisions and policy acceptability. This gives us a more nuanced view of the changing landscape of VAs and its impact on user behaviors.

3 Verification Policy Review

To address RQ1, we reviewed verification policy changes across eight popular social media platforms from April 2022 to August 2023. We outline our collection and review process and describe changing landscape of social media verification.

3.1 Data Collection and Analysis

We collected verification policies from seven of the top eight social media platforms Americans reported getting their news from in 2022 [8], i.e., Twitter, Facebook, TikTok, Snapchat, LinkedIn, Instagram, and YouTube. We excluded Reddit, which does not support account verification, but included Truth Social to represent small, niche platforms.

For each platform, we captured the verification policy on April 14, 2022 and all subsequent policy changes until August 25, 2023. April 14 marked Musk's expression of interest in acquiring Twitter. This date is a significant marker for our analysis, as it potentially influenced changes in the verification policy landscape. We monitored platform policies until our final participant completed our survey (see Section 4) to ensure we captured changes that could affect user perceptions. Details about our web scraping process are in Appendix B.

We also manually reviewed all of Musk's personal tweets about Twitter's verification policy during this period. Musk regularly made policy pronouncements publicly, which drove news coverage [33, 65] and may have influenced perceptions.

To identify common themes across verification policies, we performed an inductive thematic analysis, allowing policy dimensions to arise from the data [68]. Two researchers collaboratively reviewed the initial policies for each platform and subsequent changes as they were collected. Codes were then discussed with the full research team until full agreement was reached. Because we only sought to identify themes and do not attempt to use results for quantitative comparison, we did not assess inter-rater reliability [39].

3.2 Results

We observe several independent *dimensions* of social media verification policy: who can be verified (Eligibility), how accounts are verified (Verification Method), whether users pay a fee, requirements to prevent "deception," and required activity history. Table 1 summarizes the reviewed policies, including any changes occurring during our review.

Further, we observe three distinct *time periods* of social media verification policy:

Before Musk's Twitter takeover (Period 1). From the start of our review (April 14, 2022) until Musk's takeover of Twitter (October 27, 2022), the policies of all eight social media platforms were similar. All allowed verification only for "Notable" users (e.g., celebrities, journalists, public figures). They required users provide government documents to prove identity and did not charge for verification. There was some variation in what platforms considered "deceptive." These policies prevent accounts from changing their account information (e.g., username), having usernames similar to other accounts, posting spam, or attempting to manipulate the platform.

Musk acquired Twitter (October 27, 2022; Period 2). Musk made sweeping verification policy changes by introducing Twitter Blue on November 9, 2022. This program opened verification to any user, removed user identity checks, and required payment [73]. Musk argued open verification would improve conversation quality [15] and reduce bots by creating a barrier to entry [16, 17]. These changes faced broad criti-

Platform	Icon	Eligibility	Ver. Meth.	Payment	Non-Deceptive	Active
Twitter [73, 74]	✓	Notable → Open	Gov ID → Phone	Free → Paid	No profile changes,[1] spam, misleading behaviors, or platform manipulation	Active past 30 days
Facebook [42] Instagram [42]	✓ ✓	Notable → Open Notable → Open	Gov ID Gov ID	Free → Paid Free → Paid	No profile changes,[1] unique No profile changes,[1] unique	Prior posting history Prior posting history
TikTok [72]	✓	Notable	Gov ID	Free	No profile changes[1]	Logged in past 6 months
Snapchat [62]	★	Notable	Gov ID	Free	No misleading behaviors	Regularly post content
LinkedIn [34]	✓	Notable → Open	Gov ID[2]	Free	No profile changes	-
YouTube [23]	✓	Notable	Gov ID[3]	Free	No profile changes[1]	Regularly post content
Truth Social [64]	✓	Notable	Gov ID	Free	No misleading behaviors	Regularly post content

[1] All platforms restricted VAs from changing their username. Some also prevented changes to other profile data, such as profile photos and bios.
[2] LinkedIn's verification is only available to US users (through the CLEAR ID program) or employees of companies participating in LinkedIn's company email verification or Microsoft's Entra Verified ID programs.
[3] YouTube does not verify documentation by default, but reserves the right to request additional documentation if necessary.

Table 1: Summary of verification policy dimensions and verified indicators per platform. → indicates a change in the policy during our review with the left hand side indicating the policy at the start of our review and the right hand side showing the final policy.

cism [10, 21, 30, 33, 45, 48], and verified impostor accounts quickly appeared [50, 66], indicating the changes did not produce Musk's desired effect [70].

Twitter paused Twitter Blue on November 11, 2022 and reintroduced it on December 12, 2022 with modified eligibility requirements to limit impostors. Specifically, users were required to verify a working phone number and must have been active 30 days before verification.[4] Twitter also introduced government (✓) and company (✓) badges which were only available to organizations fitting these descriptions.

Potentially adding to user confusion, users verified under Twitter's original verification policy (Twitter Legacy) maintained their verified indicator. Verification of Twitter Legacy and Blue accounts was indistinguishable when looking at individual posts. The only distinction was an indicator on the Twitter Legacy accounts' profile pages. The Twitter Legacy policy remained in effect until April 1, 2023 [48].

While not directly related to the verification policy, Twitter also began prioritizing posts by VAs (January 5, 2023) [77]. Twitter argued this was to ensure users are most likely to see "content that is relevant, credible, and safe," implying a link between verification and credibility.

During this period, all other platform policies were stable.

Meta and LinkedIn alter policies (February 20, 2023; Period 3). Meta, the parent company of Facebook and Instagram, announced Meta Verified [42]. Like Twitter Blue, this subscription-based verification program was open to all users and required payment. However, Meta continued to require government ID for verification—the most significant difference between Twitter's and Facebook's final policies.

On April 12, 2023, LinkedIn also opened verification eligibility beyond notable users [34]. LinkedIn began allowing U.S. users to verify their identities through the CLEAR ID program and verified users with certain corporate email addresses or through the Microsoft Entra Verified company ID program. While not available to all users, it is more open than previously, and follows Meta's example of maintaining identity verification while increasing eligibility.

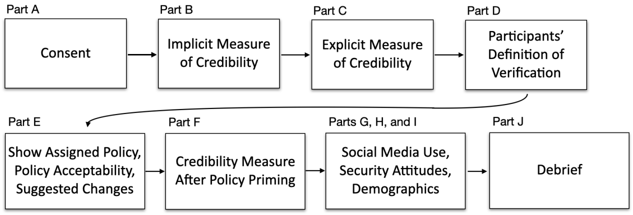

Part A
Consent
Part B
Implicit Measure of Credibility
Part C
Explicit Measure of Credibility
Part D
Participants' Definition of Verification
Part E
Show Assigned Policy, Policy Acceptability, Suggested Changes
Part F
Credibility Measure After Policy Priming
Parts G, H, and I
Social Media Use, Security Attitudes, Demographics
Part J
Debrief

Figure 1: Sections and flow of the user study.

Our identified dimensions of verification align with those outlined by Xiao et al.'s prior review [81], though our results capture changes to Meta's and LinkedIn's policies that occurred after their review. Our full dataset of policy changes is in supplementary materials [2].

4 Survey Methods

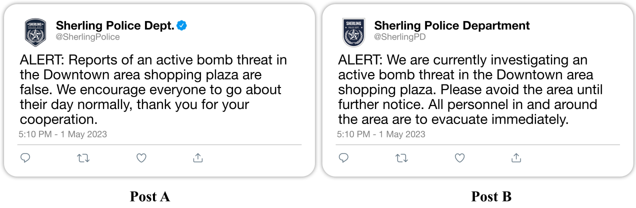
Using the policy dimensions identified in Section 3, we developed an online survey to test participants' understanding of platform policies (RQ2) and their preferences for each policy dimension (RQ3).

4.1 Survey Design

Figure 1 shows the stages of our online survey, which we describe below in turn.

Consent (Part A). We began with a consent form describing the study, potential risks, and data protection procedures. To avoid priming for the verified indicator, which users might

[4] The policy initially added a 90-day activity period on November 24, 2022, but this was relaxed to 30 days prior to Twitter Blue's restart.



Sherling Police Dept. ✓
@SherlingPolice

ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false. We encourage everyone to go about their day normally, thank you for your cooperation.

5:10 PM - 1 May 2023

Post A

Sherling Police Department
@SherlingPD

ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area until further notice. All personnel in and around the area are to evacuate immediately.

5:10 PM - 1 May 2023

Post B

Figure 2: Example Police/Declarative/Twitter condition posts.

otherwise ignore in practice, we used deception when describing the study's purpose, indicating it was to understand how users assess social media posted content credibility.

Implicit effect of verified indicator (Part B, RQ2). Next, participants were shown a pair of posts reporting contradictory information, both from accounts presenting as authorities on the subject. Figure 2 shows an example pair of posts. Posted content details, such as whether they included a verified indicator and the platform for which they were formatted (Twitter or Facebook) varied per condition (see Section 4.2). Participants were asked to indicate which posted content was more likely correct, on a five-point Likert scale. Because the contradictory posts cannot both be true, participants must make some assessment (potentially based on the verified indicator) about account identity to determine which is more credible.

Explicit effect of verified indicator (Part C, RQ2). Next, we asked participants whether the verified indicator affected their posted content credibility choice, on a four-point Likert-type scale from "No effect" to "Major effect." To compare the verified indicator's effect to other account features, participants were asked the same question about the account's picture, name, and handle.[5] The order of account feature questions was randomized to avoid ordering effects [58].

Participants' verification definitions (Part D, RQ2). We then asked participants to define verification to investigate how they understand verification and if this varies by platform.

Assigned verification policy perceptions (Part E, RQ3). We gave a mock verification policy and asked participants to assume their condition-assigned platform adopted this policy. We asked whether they believed it was "acceptable for verifying account owner identity" on a 5-point Likert-type scale from "Unacceptable" to "Acceptable." We also asked them to provide one modification (i.e., addition, deletion, change) to improve the policy. This open-ended question was intended to capture the policy elements participants prefer and prioritize, including those not used on social media platforms.

Credibility perceptions after policy priming (Part F, RQ3). In Part F, we showed participants the original contradictory posts together with their assigned mock verification policy. Then, we repeated Part B's question, asking participants to choose which posted content was more credible, this time assuming verification via the given mock policy. Next, we asked participants to assume a friend was unsure which posted content was more credible, and tell us what advice they would give to help the friend decide. This open-ended question captured an additional perspective into participants' credibility assessment. This section included an attention check to identify and remove inattentive respondents [40].

Social media use (Part G), Security attitudes (Part H), and Demographics (Part I). We concluded with questions about our participants' background and demographics. We asked about their social media use for the two platforms tested, as well as more generally. Participants completed Faklaris et al.'s SA-6 scale [18] to assess their computer security practices.

Debrief (Part J). Because we used deception, we debriefed participants about the study's true nature, providing Twitter's and Facebook's verification policies and links to best-practice guidelines for assessing posted content credibility [35, 60, 61].

4.2 Conditions

Each participant saw two contradictory posts (Parts B and F) and a mock verification policy (Parts E and F). We describe the possible posts and policies defining each *condition*.

Posted content variables. To test the verified indicator's effect, we created four posted content pairs. First, we varied the platform. One of our research questions (**RQ2**) is whether users perceive differences in verification policy between platforms and how this impacts VA credibility perceptions. For this dimension, participants were shown posts using Twitter or Facebook visual cues. This included the posted content design, verified indicator shown (i.e., ✓ vs. ✓), and terminology in survey questions (e.g., "Please answer the following questions considering the two *Twitter* posts above[6]"). We chose these platforms because Twitter changed its verification policy most significantly (see Section 3) and Facebook was the most popular platform with a comparable modality (i.e., YouTube, Instagram, TikTok are mostly image and video-based).

Second, we varied the posted content. Prior work showed content affects users' credibility perceptions [75], so we test multiple content types to avoid bias from a single type.

One pair of posts describes an alleged bomb threat (*Police*), as posted by different accounts (Sherling Police Dept. *@SherlingPolice* or Sherling Police Department *@SherlingPD*) claiming to be the same entity. One post claims the threat is false; the other asserts it is true. The second pair (*Coffee*) appear to be posted by medical doctors (Dr. Samuel Smith, M.D. *@DrSmithMD* or Dr. Alexander Kim, M.D. *@DrKimMD*). The posts contradict about a link between coffee consumption and risk of a disease. Table 2 details the posts. Combining

[5] The account handle question was only included for participants in the Twitter condition because Facebook accounts do not have this feature.

[6] Emphasis not included in survey.

platform and content options produced four posted content conditions. Participants were randomly assigned to one.

The police departments, doctors, and diseases were fictional to eliminate prior knowledge bias. We avoided political topics to prevent polarization effects [31], as prior work showed people distrust evidence contradictory to their beliefs on controversial topics [55]. We chose topics of general importance, where people must rely on expert insights. We chose to use authoritative accounts, as accounts like these could be verified or unverified under all policies reviewed in Section 3, creating a range of reasonable justifications participants could come to in their decision-making. Prior work showed users are more likely to find authoritative accounts credible [75], so we only used authoritative accounts to control for this effect.

To control for potential bias toward declarative or contradictory statements, we randomized which account was verified. We randomized the order the declarative and contradictory posts were shown, to control for ordering effects [58]. To control for other possible credibility indicators, other posted content elements (author profile image, retweets and likes counts, and time since publication) were held constant. Previous research showed these elements significantly affect user perception of posted content credibility [44].

Policy variables. After asking about the pair of posts, we presented participants with mock verification policies to observe how varying policy definitions affect their perception of the verified indicator (**RQ3**). The policies had three variables, representing the three dimensions we observed multiple platforms change in our policy review (Section 3). Table 2b gives the policy text shown for each condition. First, we varied who can be verified (Eligibility). The policy was either *Open*, meaning anyone can apply, or *Notable*, meaning only well-known individuals and organizations are eligible. Next, policies varied in how accounts are verified (Verification Method). That is, accounts must either confirm an email or phone number (*Phone*) or provide government-issued ID (*Gov ID*). Finally, the policy specified whether verification required *Payment*. We used a full-factorial variable combination to create eight policies. Participants were randomly assigned a policy independent of their post and platform condition.

4.3 Recruitment

We conducted our survey on Prolific, a research recruitment service providing high-quality samples [52, 71]. We limited participation to Prolific users at least 18 years old and located in the United States. We used Prolific's census-representative sample feature [56] to ensure a U.S. population-representative distribution by age, gender, and ethnicity. Survey completion time averaged 8.2 minutes, and we paid participants $2.

Content	Position	Posted content Text Summary
Police	Decl.	ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area...
	Cont.	ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false...
Coffee	Decl.	Individuals who consume more than three cups of coffee per day may have a higher risk of developing endothrombocytisis.
	Cont.	There have been no research studies that have established a link between coffee consumption and endothrombocytisis.

(a) Posted Content Variables

Dimension	Option	Policy Text
Eligibility	Open	**Any user** can apply for verification
	Notable	Only **well known, high-profile individuals and organizations** can apply for verification
Verification Method	Phone	Accounts are required to **confirm a phone number or email** with the platform
	GovID	Accounts must **submit government-issued identification** matching the name of the account
Payment	Paid	Accounts **pay a monthly subscription fee** to maintain their verification
	Free	Accounts **do not pay any fee** for verification

(b) Policy Variables

Table 2: Summary of (a) posted content and (b) policy conditions. There were four posted content and eight policy conditions, resulting in 32 total conditions after a full-factorial combination.

4.4 Pilot

We piloted the survey with nine participants—drawn from a convenience sample, selected for varying social media familiarity. Pilot participants were asked to "think aloud" while answering questions. We iteratively updated the survey for clarity after each pilot until further changes were unnecessary.

We also tested a third content type about an E.Coli outbreak in lettuce. We recruited 50 participants on Prolific and assigned them randomly to one of the three content types to test whether any content type behaved unexpectedly (e.g., prior experience bias or unexpected relationship with current events). We did not observe unexpected responses, but saw similar results between the E.Coli and Coffee conditions. Therefore, we dropped the E.Coli condition to increase our analysis power by recruiting more participants per condition.

4.5 Data Analysis

Quantitative analysis. To test verification's effect on participants' posted content credibility perceptions before and after stating a policy, and to assess participants' perception of policy acceptability, we used ordinal logistical regressions.

For the two posted content credibility perceptions questions, the outcome variable is a 5-point Likert-scale response regarding which post was correct (Part B and Part F, respectively). Each response was modified to indicate whether the participant perceived the account with or without the verified indicator as correct, to allow for comparisons; e.g., if a participant shown the posts in Figure 2 selected "Definitely A" from the possible options, because A was the VA, their response was modified to "Definitely the VA." For the policy acceptability regression, the outcome variable was the participant's response to the policy acceptability question in Part E.

In each regression, we include the assigned condition's three elements (platform, content, and position) as explanatory variables. For the policy-related regressions (Part E and Part F), we added the policy variables (Eligibility, Verification Method, and Payment). In all regressions, we include demographic explanatory variables (age, gender, education), amount of time spent using Twitter and Facebook, number of social media platforms used, and SA-6 scores. Table 6 in Appendix D summarizes the variables included per regression.

To select a parsimonious model without overfitting, we constructed initial regression models using all possible explanatory variable combinations. We selected models with the minimum Bayesian Information Criterion, appropriate for testing goodness-of-fit [57, 63].

We also examined the explicit impact of verified indicator on credibility perceptions. We compared responses regarding the verified indicator's impact between Twitter- and Facebook-assigned participants using a Pearson's χ^2 test, appropriate for categorical data [51]. Next, we compared responses across the four[7] account features (verified indicator, account username, photo, and handle) using non-parametric, repeated measures tests, appropriate for multiple Likert-scale responses per participant. We began with an omnibus Friedman test across features to control for Type I error; if the result was significant, we applied the Wilcoxon signed-rank test to planned pairwise comparisons of the verified indicator with every other feature [78]. Comparisons were across content conditions.

Qualitative analysis. We used iterative open coding to analyze free-response questions [68]. As our questions were all related to VAs and verification policies, similar to the free-response questions in Vaidya et al. [75], we began with their codebook. However, as verification policies have changed, we allowed additional codes to arise inductively. Three researchers extended the initial codebook collaboratively by reviewing 10 responses. Two researchers independently coded additional responses in rounds of 100, updating the codebook incrementally. After rounds, the coders met, assessed inter-rater reliability using Krippendorff's alpha [26], and resolved coding differences. After two rounds (200 responses), the coders achieved $\alpha = 0.80$, which represents acceptable agreement. The remaining 1386 responses were divided evenly and coded separately by the two coders [26]. Finally, the two researchers performed an axial coding to identify relationships between codes and produce higher-level groups [12, pg. 123-142]. The final codebook is in supplementary materials [2].

To compare initial verification definitions between participants shown Twitter and Facebook posts, we perform Pearson's χ^2 tests, appropriate for categorical data [51]. For each higher-level code group, we compare a code's presence from this group between Twitter- and Facebook-assigned participants. Because this requires multiple testing, we apply a Benjamini-Hochberg correction to adjust p-values [6].

4.6 Ethical Considerations

Tufts University's IRB approved this study. We obtained informed consent prior to the survey. Because we used deception in our study description, we concluded with a debrief and asked participants to re-consent. To avoid response coercion, participants were told they would be paid for completing the survey even if they refused consent, but their response would be deleted. Three participants withdrew after the debriefing.

Responses through Prolific are provided pseudonymously, with only the participant's Prolific ID identifying their response. We did not request additional identifying information.

4.7 Limitations

We presented mock posts, as this provides the control needed to reason about specific variables' effects on credibility perceptions. However, we are unable to capture other credibility perception influences, such as the author's reputation, the viewer's relationship with the author, or viewer's relationships with others who interact with the posted content (e.g., liked or shared). The types of content and other metadata we test are also limited, meaning we are unable to comprehensively test these factors' influence on posted content credibility. We only test textual content, and so our results may not generalize to verified indicators on video-based platforms (e.g., TikTok). We do not test controversial content, as we expect the introduced bias to overwhelm any effect from the verified indicator. This is an inherent tradeoff to limit the study's scope to a reasonable condition set. Our results establish a baseline of verified indicator effect on perceived credibility, and future work should study how the effect changes in the presence of video and controversial topics. We believe our conditions are sufficient to target our study's research questions.

The study's setting also differs from the real world. Participants may have spent more time reviewing our contradictory posts than when casually browsing social media feeds. Also, presenting contradictory information side-by-side is not representative, as these posts would be interspersed with other posts. Our results are indicative of a best-case situation where users carefully consider all relevant information, which is

[7] Three for participants assigned Facebook because they were not shown a user handle.

likely closer to the truth in emergency situations when finding good information is safety-critical and social media is saturated with posts about an ongoing event.

For open-response questions, we give the percentage of participants who stated each theme. However, not mentioning a theme does not indicate disagreement. Participants may have failed to state an idea or considered other thoughts more relevant. Our open-response results should be viewed as a measure of what was "front of mind" when answering.

We expect non-U.S. populations' views of verified indicators differ due to the ways social media is politicized in the U.S. [11]. Cross-cultural comparisons require a sample size infeasibly large for this study. Instead, we limit our sample and conclusions to a single culture with which we are familiar.

Even though we used Prolific's census-representative sample feature, Prolific users are often more knowledgeable regarding privacy and security and more likely to use multiple social media platforms [71], which may impact generalizability. To account for these differences, we controlled for social media use and security attitudes in our regressions.

As these limitations apply across all conditions, we focus primarily on between-condition comparisons.

5 Survey Results

The majority of our key findings are taken from our regression analyses over initial perceived correctness (Table 4a), perceived correctness after proposing a new policy (Table 4b), and perceived policy acceptability (Table 5). Only variables in the final selected model are shown (as groups of rows). We give the base case first for categorical variables. We selected base cases expected to correlate with the lowest levels of VA perceived correctness and policy acceptability.

For categorical variables, OR is the odds ratio of the outcome (e.g., acceptability) increasing one Likert-scale unit when switching from the base case to the given parameter level. For numeric variables (e.g., SA-6), OR is the odds the outcome increases one Likert-scale unit for each one-point increase in the numeric variable. For example, the OR for *Police* in Table 4a indicates a participant assigned *Police* instead of *Coffee*—holding all other variables equal—would be 1.57× as likely to increase one unit in perception that the VA posted the correct message. Because this effect is greater than one, participants are more likely to report the VA as correct for *Police* than *Coffee*. *Police*'s confidence interval (CI) indicates that if we ran the study many times, we would expect 95% of runs to produce ORs between 1.31 and 1.87. The *p*-value (< 0.001) is less than our significance threshold ($\alpha = 0.05$), indicating a significant difference between *Police* and *Coffee*.

5.1 Participants

1739 participants attempted and 1660 completed the survey. We removed 27 who failed the attention check, 30 who gave nonsensical or obviously AI-generated responses (long paragraphs with distinctive wording) to open-ended questions, and 3 who withdrew after the debrief. Our final dataset contains 1600 responses (50+ per condition).

Metric	%
Age	
18-29 years	23.8%
30-49 years	34.9%
50-64 years	28.9%
65+ years	12.4%
Platform w/Account	
Facebook	82.2%
YouTube	78.9%
Instagram	68.7%
Twitter	66.7%
LinkedIn	42.7%
TikTok	37.0%

Metric	%
Education	
H.S. or below	13.0%
Some college/ Assoc.	32.9%
B.S. or above	53.9%
Prefer not to respond	0.3%
Social Media Use	
<30 mins daily	19.1%
30 mins-1 hr daily	30.6%
1-2 hrs daily	28.7%
2-4 hrs daily	16.6%
5-6 hrs daily	3.3%
>6 hrs daily	1.6%

Table 3: Participant demographics. Percentages may not add to 100% due to non-response or selection of multiple options.

Table 3 summarizes participant demographics. Additional demographics are reported in Appendix D. Our participants' gender and income were similar to the 2020 U.S. Census [1]. Participant ethnicities were similar to the U.S. Census, though White participants were overrepresented and Latino/a participants were underrepresented. Participants were more educated and younger on average than the U.S. population, though similar to estimated Twitter user demographics [80]. Participants' average SA-6 score was 3.61, close to the average score from a U.S. Census-representative sample [18].

Participants most often had accounts with Facebook (82.2%), YouTube (78.9%), Instagram (68.7%), and Twitter (66.7%)—similar to other social media use surveys [5]. They most often used Twitter at least every other day (38.8%), with the majority using it at least once per week (64.9%), and many having no account (35.1%). Participants were more active on Facebook, with most using it at least every other day (56.1%) and only 19.8% not having an account. Facebook use did not vary significantly between participants assigned to the Twitter and Facebook conditions ($\chi^2 = 2.9, p = 0.566$). Twitter usage did vary between platform conditions ($\chi^2 = 9.6, p = 0.047$), but the effect size indicates little if any association ($\phi = 0.08$) [13, pg. 282].

5.2 Initial Impact of Verified Account (RQ2)

Here, we discuss participant perceptions of the contradictory posts' credibility (Part B) and how they perceived the verified indicator impacting their decision-making (Part C) *prior* to being given a verification policy. Figure 3 summarizes initial credibility perceptions divided by experimental condition, and Figure 6 in Appendix D summarizes participants' perceptions of the account features' decision-making impact.

No difference between platforms. Across conditions, par-

Variable	Value	Odds Ratio	CI	*p*-value
Content	Coffee	–	–	–
	Police	**1.56**	**[1.31, 1.87]**	**<0.001***
Position	Contradict.	–	–	–
	Declar.	**1.42**	**[1.19, 1.69]**	**<0.001***
Age	–	–	–	–
	+1	**0.99**	**[0.98, 0.99]**	**<0.001***

– Base case (OR=1, by definition)
*Significant effect

(a) Initial Perceived Verified Account Correctness

Variable	Value	Odds Ratio	CI	*p*-value
Content	Coffee	–	–	–
	Police	**4.13**	**[3.39, 5.02]**	**<0.001***
Availability	Open	–	–	–
	Notable	**1.80**	**[1.50, 2.17]**	**<0.001***
Verification Method	Phone	–	–	–
	Gov ID	**1.30**	**[1.08, 1.56]**	**0.005***
Facebook User	False	–	–	–
	True	**1.54**	**[1.21, 1.95]**	**<0.001***
SA-6	–	–	–	–
	+1	**1.21**	**[1.08, 1.36]**	**<0.001***

– Base case (OR=1, by definition)
*Significant effect

(b) Verified Account Correctness After Policy Given

Table 4: Summary of regression over participants' VA correctness perception (a) before and (b) after being shown a specific policy. Pseudo R^2 measures for (a) were 0.01 (McFadden) and 0.04 (Nagelkerke), and for (b) were 0.07 (McFadden) and 0.17 (Nagelkerke).

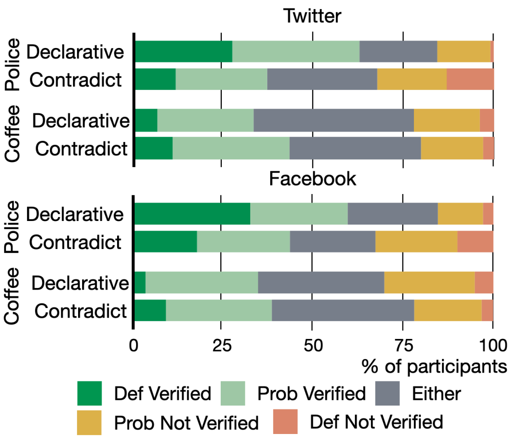


Figure 3: Likert-scale response showing whether participants perceived the VA as more likely credible, organized by assigned social media platform, content type, and the position taken by the VA.

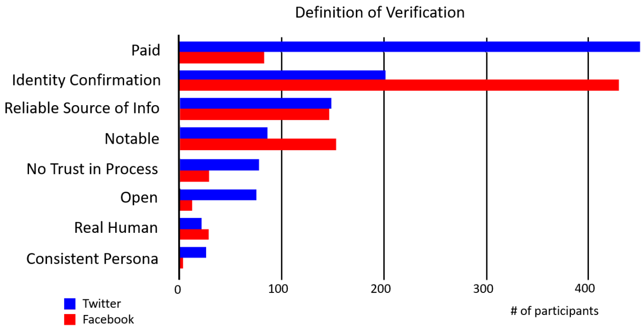


Figure 4: Participants' verification definitions by platform.

ticipant perceptions of the more likely credible post were evenly distributed. Participants most often indicated the VA was "Definitely" or "Probably" credible (43.9%). However, 32.1% indicated "Either the verified or not VA" was credible and 24.1% chose "Definitely" or "Probably" the non-VA. Results were similar whether participants were assigned Twitter (43.8% VA, 33.3% either, 22.9% non-VA), or Facebook (44.0% VA, 30.8% either, 25.3% non-VA). The selected regression (Table 4a) did not include platform, indicating no observed statistically significant difference between platforms.

When asking participants directly about the verified indicator's impact on their decision-making, responses again were split. A slight majority indicted it had no impact (52.0%), while 48.0% reported at least a "Minor effect." Participants were statistically significantly more likely to rank the verified indicator's effect higher than the account picture ($Z = 14.46, p < 0.001$) and handle ($Z = 7.31, p < 0.001$) according to Wilcoxon-Pratt signed rank tests. We did not observe a statistically significant difference between the verified indicator's and account name's perceived impact ($Z = 1.71, p = 0.087$).

Comparing platforms (Figure 6) there is no clear difference: 46.6% of Facebook-assigned participants reported at least a "Minor effect" versus 49.4% for Twitter. No statistically significant difference was observed ($\chi^2 = 4.82, p = 0.186$).

Content had the biggest effect. Participants shown the Police content were statistically significantly more likely to perceive the VA as credible ($OR = 1.56, p < 0.001$). If the VA posted the declarative statement (e.g., there *was* a bomb), participants were statistically significantly more likely to perceive the VA as credible ($OR = 1.42, p < 0.001$). This follows prior work [75], which showed content drives message credibility.

Age has some effect. Grouping participants by decade, we observed a downward trajectory in percentage of participants perceiving the VA as "Definitely" or "Probably" credible (52.1% of <30s to 24.1% of >70s). Older participants more often indicated "Either the verified or not VA" was credible (25.8% of <30s to 42.2% of >70s)—the correct response, as VAs do not necessarily post credible content. With each additional year, participants were 0.99× as likely to find the VA more credible by one point ($p < 0.001$). When comparing an individual one standard deviation older (∼15.75 years), we would expect them to be 0.85× as likely to increase one point on the Likert scale. This contradicts Xiao et al.'s prior observation of no statistically significant relationship [81].

5.3 Verification Policy Definitions (RQ2)

Here, we discuss participants' free-response verification definitions (Part D) prior to priming about a particular policy. These definitions mostly aligned with those found via our policy review (Section 3.2). Because we asked participants about platforms with divergent policies (i.e., Facebook and Twitter), we discuss each separately. Our final codebook is in supplementary materials [2]. Figure 4 summarizes responses by platform. Because participants could describe multiple dimensions, these counts do not sum to the total number of participants. These numbers represent front-of-mind definitions; not mentioning a dimension does not necessarily mean the participant does not believe it applies to the policy.

Participants were more likely to believe Facebook confirms user identity. 54.2% of Facebook-assigned participants stated Facebook confirms the user's identity matches their online persona. As one participant said, "[users] need to submit identification, and Facebook manually reviews it." Only 25.1% of Twitter-assigned participants said the same. This difference was statistically significant ($\chi^2 = 140.58, p < 0.001$). While the share of Twitter-assigned participants who believe Twitter verifies identity is concerning, the majority of participants' perceptions align with each platform's actual policies. While not directly comparable, we note the percentage of participants stating Twitter verifies user identity in our survey is much lower than in Xiao et al.'s [81], potentially indicating user understanding of Twitter's policy has improved.

Many participants focused on measures to ensure accounts were made by real humans, not bots. Instead of ensuring VAs' true identity matched their persona, many perceived verification as simply requiring the user verify personal information (e.g., mailing address, email, phone number), limiting verification of bots (18.4% Twitter; 18.3% Facebook). We did not observe a statistically significant difference ($\chi^2 < 0.001, p = 1$). Both platforms require these checks, though they are Twitter's primary verification mechanism.

Payment is mostly associated with Twitter. More than half of Twitter-assigned participants mentioned payment (56.0%). One participant explained, "You pay $8 and elon gives you the blue checkmark." Conversely, few (10.4%) Facebook-assigned participants believed payment was required. This difference was statistically significant ($\chi^2 = 370.6, p < .001$). Xiao et al. found similar results (i.e., Twitter is paid and Facebook is free) [81], but this was correct at the time, as Facebook switched to a paid model after their survey. We show this perception of Facebook as free is now a misconception, indicating users are unaware of Facebook's policy change.

Facebook-assigned participants were more likely to believe verification was for "notable" accounts. 19.3% of Facebook-assigned participants said only notable accounts could be verified. As one participant said, "they have to be notable enough to where other people want to make fake accounts of them." This misconception was not common, but was more common ($\chi^2 = 21.881, p < .001$) among Facebook-assigned than Twitter-assigned participants (10.8%). Conversely, Twitter-assigned participants (8.9%) were more likely ($\chi^2 = 44.80, p < 0.001$) than Facebook-assigned participants (1.4%) to say anyone could be verified.

Facebook-assigned participants were more likely to be unaware of the platform's policy. Many Facebook participants reported not knowing Facebook's policy (17.2%). One participant said, "I actually don't know what the qualifications are to maintain a checkmark. I kind of blindly trust it has been adequately verified." Some were even unaware Facebook had VAs (2.1%). One participant stated, "Facebook uses blue checkmarks? I thought you were talking about Twitter." Many fewer Twitter-assigned participants (7.6%) reported lacking knowledge ($\chi^2 = 32.988, p < .001$).

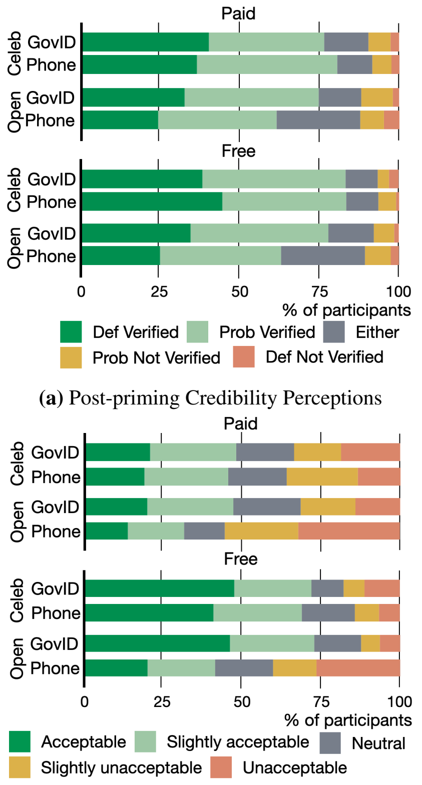
Some people still conflate verification with credibility. Though not many, some participants (3.7% of Facebook-assigned; 2.7% of Twitter-assigned) continue to believe verification indicates the account is a reliable source of information. As one participant explained, "I would think that Facebook's fact checkers would verify the post was legit and gave good information." This mirrors previous work showing a minority of users conflate authenticity with credibility [14, 22, 75, 81].

Participants criticized Twitter more. Some participants mistrusted the verification process, describing it as politically biased ("They must share the same 'opinion' as Facebook's creator/staff"), failing to prevent inauthentic accounts ("there are so many loopholes now for bots to act like humans and falsify information"), or expressed nihilism ("Better to let the [expletive] thing die than waste time on this verification nonsense"). Criticism was more common ($\chi^2 = 23.914, p < .001$) among Twitter-assigned (9.6%) than Facebook-assigned participants (3.4%). These are small fractions, but we note we asked for participants' definition of the process, not their opinion of it.

5.4 Verification Policy Perceptions (RQ3)

We next discuss perceptions of VA posts' credibility after defining a verification policy (Part F) and how acceptable participants consider the policy (Part E). We saw a significant increase in perceptions that the VA's posted content was credible ($Z = 21.69, p < 0.001$ in Wilcoxon Signed Rank test). This was likely affected by our priming participants to focus on verification by asking for a definition (Part D) and giving a specific policy (Part E). Therefore, we do not compare initial and after-priming responses, but only provide between-participant comparisons on the after-priming question.

We focus first on the three varied policy dimensions (Eligibility, Verification Method, and Payment), then discuss other factors. Figure 5a summarizes participant correctness per-



(a) Post-priming Credibility Perceptions

(b) Policy Acceptability

Figure 5: Likert-scale response indicating (a) posted content credibility perceptions and (b) policy acceptability after defining a policy. Both are organized by assigned policy dimensions.

ceptions, divided by dimension, and Figure 5b shows how acceptable participants considered each policy.

Limiting verification to notable accounts and authenticating with a government ID (govID) increases perceived posted content credibility and acceptability. Most govID-assigned participants (78.0%) believed the VA's posted content was "Definitely" or "Probably" more credible. VA posted content credibility perceptions dropped to 72.0% when told accounts were verified via email or phone, with more participants saying "Either" posted content could be credible (18.5%; 12.9% for govID). This difference was statistically significant, with govID-assigned participants 1.30× more likely to increase one point toward the VA ($p = 0.005$, Table 4b). GovID-assigned participants were also statistically significantly more likely to find the policy acceptable ($OR = 1.78, p < 0.001$, Table 5) with a majority finding it "Slightly acceptable" or "Acceptable" (60.0%), but this was a minority opinion for those shown the email or phone policy (46.9%). Requiring govID was the most commonly desired policy change (N=306) with only 33 participants saying govID should not be required. One participant explained, "I would require a photo ID. I can say I'm John Travolta and I can give you my email address (which can be almost anything) to confirm me, but I'm not John Travolta." This aligns with best practices for verification [24], as it is easier to create a new email or phone number than falsify a government document, and there have already been many cases of malicious accounts defeating phone verification [50, 66, 70].

Variable	Value	Odds Ratio	CI	p-value
Eligibility	Anyone	–	–	–
	Notable	**1.57**	**[1.32, 1.88]**	**<0.001***
Verification Method	Phone	–	–	–
	Gov ID	**1.80**	**[1.51, 2.15]**	**<0.001***
Payment	Paid	–	–	–
	Free	**2.53**	**[2.11, 3.03]**	**<0.001***
SA-6	1	–	–	–
	+1	**1.27**	**[1.14, 1.41]**	**<0.001***

*Significant effect — – Base case (OR=1, by definition)

Table 5: Policy acceptability regression summary. The model's pseudo R^2 measures were 0.04 (McFadden) and 0.11 (Nagelkerke).

There was a similar difference when comparing notable-only-assigned participants (80.8% "Definitely" or "Probably" more credible), as opposed to participants assigned an open policy (69.3% "Definitely" or "Probably" more credible). This difference was statistically significant with a slightly larger effect size ($OR = 1.80, p < 0.001$). Participants reported higher acceptability for the notable-only policy (58.5% "Slightly acceptable" or "Acceptable"), compared to an open policy (48.4% "Slightly acceptable" or "Acceptable")—also statistically significant ($OR = 1.56, p < 0.001$). However, when asked for a desired policy change, a greater proportion of participants wanted the policy to be open, not notable. Of the 804 participants shown a notable-only policy, 18.9% wanted it to be open, while only 8.1% of open policy participants wanted verification for notable users only. This sentiment for open policies was driven by concerns of equality; as one participant stated, "I don't believe one has to be well known or high-profile to be verified. That absolutely stinks of elitism." This contradicts our regression results, suggesting participants are split on their preference for Eligibility.

Payment does not affect perceived correctness, but reduces approval. We did not observe a statistically significant impact on participants' VA posted content credibility perceptions based on payment. When shown a free verification policy, 76.8% of participants indicated the VA's post was "Definitely" or "Probably" more credible, compared to 73.2% of participants shown a paid policy. Free verification was the strongest factor increasing policy approval ($OR = 2.54, p < 0.001$). While 64.0% shown a free policy found it at least "Slightly acceptable", only 43.1% said the same of paid policies. Like Xiao et al. [81], we found many participants focused on price when suggesting a policy change (N=342). One participant said, "Money shouldn't be a barrier to doing public good." This indicates payment might not impact users' VA perceptions, but it displeases users, as observed with Twitter [30].

Social media use and security attitudes play a role. Participants who use Facebook were more likely to view the VA's posted content as credible (76.9% said "Definitely" or "Probably" more credible) compared to non-Facebook users (66.3% said "Definitely" or "Probably" more credible) ($OR = 1.54, p < 0.001$). Participants who reported taking more general security actions were more likely to view the VA's posted content as more credible ($OR = 1.21, p < 0.001$) and find the policy acceptable ($OR = 1.27, p < 0.001$). This may suggest the misconception that VAs are "secure," i.e., should be trusted over other accounts. However, prior work contradicts this [22, 75, 81], and few participants said verification indicates credibility (see Section 5.3). This may instead be an effect of the specific contrasting scenarios we chose, where the only major difference was the verified indicator and accounts were authoritative. Security-conscious participants may have been more likely to consider this difference.

6 Discussion

Our results reveal users' understanding of recent verification policy changes, along with their perceptions of the changes and other potential policies. We suggest social media platform verification policy improvements and discuss future work.

Many participants were aware of Twitter's transition to paid, open verification without a required identity check. While the results are not directly comparable, this seems to indicate improved user awareness relative to Xiao et al.'s earlier survey, which found many users believed Twitter performed rigorous identity checks [81]. Conversely, our participants were unaware of Facebook's policy changes, believing it remained free and restricted to notable accounts. This misunderstanding is not as consequential as incorrectly believing accounts undergo identity verification. However, our results suggest participants were more likely to perceive VA posts as credible when only notable accounts are verified, so this misunderstanding still introduces misplaced trust.

To improve trust in the verification process, platforms should employ rigorous ID checks. Participants were more likely to find the VA's posted content more credible when it was verified with a govID, more likely to find govID verification acceptable, and frequently suggested an ID check be added to improve verification. This shows users value identity verification over other requirements for bot prevention or account consistency. If Twitter reverts to rigorous identity checks (as was rumored [53]), future work should consider whether perceptions of Twitter's policy improve, as our hypothetical settings suggest, or if these perceptions represent a one-way-ratchet and are already ingrained in users' minds.

We did not observe any significant difference in the verified indicator's effect between platforms before priming about verification. When primed, participants shown Facebook's policy were statistically significantly more likely to find the VA more credible. This suggests users do not consider policy differences without priming, and because Facebook's policy is less well known, may default to their understanding from Twitter. As social media platforms change verification policies, they must educate users to avoid misunderstandings. This is especially true when changing govID and notability requirements, which had significant impacts, though future work must determine how best to educate users.

Restrictions on account eligibility produced mixed results. Under a notable-only policy, participants were more likely to perceive the VA's posted content as more credible and find the policy acceptable. However, when asked to suggest changes to the platform, participants contradicted this sentiment by saying verification should be open to all users. One remedy suggested by a few participants (N=19) is a tiered approach to verification. As one participant suggested, "I think for public service accounts such as the fire department, police department, federal government, etc. there should be a more rigorous verification process." Similarly, some participants wanted the platform to evaluate users' authoritative credentials (N=62). This could include verifying hospital credentials of medical professionals or press credentials for journalists. Twitter somewhat employs this approach with special indicators for government and business accounts (✓, ✓). Although users may prefer this in theory, prior work found users misunderstood both badges [81]. Future research should consider the impact of these indicators, especially in emergency situations when an account's authority is important (similar to our bomb threat examples) and under various Verification Method regimes to determine the interaction between these variables.

Perhaps the most polarizing verification change is switching to a paid model. Participants found paid policies unacceptable and wanted to remove payment, matching prior work [81]. However, we did not observe an effect from payment on participants' posted content credibility perceptions. We might have expected participants to be less likely to trust paying accounts, since Twitter's verified indicator has been described as a "scarlet letter" [30] and impostor accounts have been created [50]. However, it seems users correctly associate these problems with the lack of identity verification, not payment. This suggests that while payment might annoy users, it does not negatively impact how they evaluate VA posts.

Finally, participants were statistically significantly more likely to find the VA credible after priming about verification. This could be the result of asking participants to consider a hypothetical policy, but appears more likely due to priming effects. This could be problematic for platforms using policies that do not have rigorous identity verification. Malicious users may be able to fool others into believing their posts by drawing attention to their verified indicator. Future work should investigate situations where other information beyond the verified indicator varies between contradictory posts to measure the potential risk of social engineering attacks.

Acknowledgment

We thank the anonymous reviewers who provided very helpful comments on drafts of this paper.

References

[1] Census bureau data. https://data.census.gov/. (Accessed 08-30-2023).

[2] Social media account verification perceptions. https://doi.org/10.17605/OSF.IO/A9Y3J.

[3] Internet Archive. Internet archive: Wayback machine. https://archive.org/web/. (Accessed 09-10-2023).

[4] Associated Press. Students criticize the University of North Carolina's response to an active shooter emergency. https://www.voanews.com/a/awash-in-social-media-how-are-us-police-learning-to-inform-the-public-better-after-shootings-/7100938.html, 2023. (Accessed 09-10-2023).

[5] Brooke Auxier and Monica Anderson. Social media use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/, 2021. (Accessed 08-01-2019).

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[7] Cody Buntain, Jennifer Golbeck, Brooke Liu, and Gary LaFree. Evaluating public response to the boston marathon bombing and other acts of terrorism through twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):555–558, Aug. 2021.

[8] Pew Research Center. Social media and news fact sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/, 2022. (Accessed 08-18-2023).

[9] Shelly Chaiken. The Heuristic Model of Persuasion. In *Social Influence: the Ontario Symposium*, volume 5, pages 3–39, 1987.

[10] Brian X. Chen and Ryan Mac. Twitter's blue check apocalypse is upon us. here's what to know. The New York Times. https://www.nytimes.com/2023/03/31/technology/personaltech/twitter-blue-check-musk.html, 2023. (Accessed 09-10-2023).

[11] Laura Clancy. Americans differ by party, ideology over the impact of social media on U.S. democracy, December 2022.

[12] Juliet Corbin, Anselm Strauss, and Anselm L Strauss. *Basics of qualitative research*. Sage, 2014.

[13] H. Cramér. *Mathematical Methods of Statistics*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1999.

[14] Jazlyn Elizabeth Dumas and Rusty Allen Stough. When influencers are not very influential: The negative effects of social media verification. *Journal of Consumer Behaviour*, 21(3):614–624, 2022.

[15] @elonmusk. Twitter's current lords & peasants system for who has or doesn't have a blue checkmark is bullshit. Power to the people! Blue for $8/month. https://twitter.com/elonmusk/status/1587498907336118274, 2022. (Accessed 09-10-2023).

[16] @elonmusk. Yes, this will destroy the bots. If a paid Blue account engages in spam/scam, that account will be suspended. .... https://twitter.com/elonmusk/status/1587512669359292419, 2022. (Accessed 09-10-2023).

[17] @elonmusk. Given that modern AI can solve any "prove you're not a robot" tests, it's now trivial to spin up 100k human-like bots.... https://twitter.com/elonmusk/status/1640199090112806912?s=20, 2023. (Accessed 09-10-2023).

[18] Cori Faklaris, Laura A. Dabbish, and Jason I. Hong. A Self-Report measure of End-User security attitudes (SA-6). In *Fifteenth Symposium on Usable Privacy and Security*, SOUPS 2019, pages 61–77, Santa Clara, CA, August 2019. USENIX Association.

[19] Andrew J. Flanagin and Miriam J. Metzger. The Role of Site Features, User Attributes, and Information Verification Behaviors on the Perceived Credibility of Web-Based Information. *New Media & Society*, 9(2):319–342, 2007.

[20] B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, page 1–15, New York, NY, USA, 2003. Association for Computing Machinery.

[21] Brian Fung. How Elon Musk transformed Twitter's blue check from status symbol into a badge of shame. CNN Business. https://www.cnn.com/2023/04/24/tech/musk-twitter-blue-check-mark/index.html, 2023. (Accessed 09-10-2023).

[22] Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[23] Google. Verification Badges on Channels. https://support.google.com/youtube/answer/3046484. (Accessed 09-10-2023).

[24] Paul A. Grassi, James L. Fenton, Naomi B. Lefkovitz, Jamie M. Danker, Yee-Yin Choong, Kristen K. Greene, and Mary F. Theofanos. NIST Special Publication 800-63A, Digital Identity Guidelines, Enrollment and Identity Proofing. Technical report, National Institute of Standards and Technology, 06 2017.

[25] Amelia Hassoun, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M. Russell. Practicing information sensibility: How gen z engages with online information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[26] Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[27] Brian Hilligoss and Soo Young Rieh. Developing a Unifying Framework of Credibility Assessment: Construct, Heuristics, and Interaction in Context. *Information Processing & Management*, 44(4):1467–1484, 2008.

[28] Kyle Hunt, Bairong Wang, and Jun Zhuang. Misinformation debunking and cross-platform information sharing through twitter during hurricanes harvey and irma: a case study on shelters and id checks. *Natural Hazards*, 103:861–883, 2020.

[29] Christian Johnson and William Marcellino. Reining in COVID-19 Disinformation from China, Russia, and Elsewhere, November 2021. https://www.rand.org/blog/2021/11/reining-in-covid-19-disinformation-from-china-russia.html.

[30] Alex Kirshner. How Elon Musk Turned the Blue Check Mark Into a Scarlet Letter. Slate. https://slate.com/technology/2023/04/elon-musk-twitter-blue-check-marks-verification-lebron-james.html, 2023. (Accessed 09-10-2023).

[31] Ziva Kunda. The Case for Motivated Reasoning. *Psychological bulletin*, 108(3):480, 1990.

[32] Ian Lamont. Plane Lands on the Hudson, and Twitter Documents It All. Computerworld. https://www.computerworld.com/article/2530453/plane-lands-on-the-hudson--and-twitter-documents-it-all.html, 2009. (Accessed 09-10-2023).

[33] Annabelle Liang. Elon Musk: Twitter boss announces blue tick shake-up. BBC News. https://www.bbc.com/news/business-65095684, 2023. (Accessed 09-10-2023).

[34] LinkedIn. Verified on your linkedin profile. https://www.linkedin.com/help/linkedin/answer/a1359065. (Accessed 09-10-2023).

[35] Megan Loe. 5 VERIFIED Ways You Can Fact-check Online Claims. Verify. https://www.verifythis.com/article/news/verify/fact-sheets-verify/5-tips-fact-check-online-claims-yourself-guide/536-64c58fc6-f17d-42dd-9970-b1b4814f9a87, 2023. (Accessed 09-06-2023).

[36] Gary Machado, Alexandre Alaphilippe, Roman Adamczyk, and Antoine Gregoire. Indian Chronicles: deep dive into a 15-year operation targeting the EU and UN to serve Indian interests. Technical report, EU Disinfo Lab.

[37] Odanga Madung and Brian Obilo. Inside the Shadowy World of Disinformation-for-hire in Kenya. Technical report, Mozilla Foundation. Section: Fellowships & Awards.

[38] Miriam Matthews, Katya Migacheva, and Ryan Andrew Brown. Superspreaders of Malign and Subversive Information on COVID-19: Russian and Chinese Efforts Targeting the United States. Technical report, RAND Corporation, April 2021.

[39] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, nov 2019.

[40] Adam W. Meade and S. Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

[41] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under Crisis: Can We Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 71–79, New York, NY, USA, 2010. Association for Computing Machinery.

[42] Meta. Meta verified | get a verified blue check on instagram, facebook. https://about.meta.com/technologies/meta-verified/. (Accessed 09-10-2023).

[43] Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*, 60(3):413–439, 08 2010.

[44] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[45] Casey Newton and Zoe Schiffer. Elon Musk ignored Twitter's internal warnings about his paid verification scheme. The Verge. https://www.theverge.com/2022/11/14/23459244/twitter-elon-musk-blue-verification-internal-warnings-ignored, 2022. (Accessed 09-10-2023).

[46] Matt Novak. Viral Video Alleging Canadian Wildfires Were 'Set Up' Is Very Misleading. Forbes. https://www.forbes.com/sites/mattnovak/2023/06/09/viral-video-alleging-canadian-wildfires-were-set-up-is-very-misleading/?sh=67e194bb7350, 2023. (Accessed 09-10-2023).

[47] Matt O'Brien. Canada wildfire evacuees can't get news media on facebook and instagram. some find workarounds. AP News. https://apnews.com/article/canada-wildfires-yellowknife-nwt-facebook-instagram-meta-723687efe632884e4eb1172528abb43f, 2023. (Accessed 09-10-2023).

[48] Matt O'Brien and Kathleen Foody. Confusion as Musk's Twitter yanks blue checks from agencies. AP News. https://apnews.com/article/twitter-elon-musk-blue-checkmark-celebrities-544cfd66ed3a62f51a8a80c20e11ac5b, 2023. (Accessed 09-10-2023).

[49] Kari Paul. Russian disinformation surged on social media after invasion of Ukraine, Meta reports. *The Guardian*, April 2022. https://www.theguardian.com/world/2022/apr/07/propaganda-social-media-surge-invasion-ukraine-meta-reports.

[50] Kari Paul. Fake accounts, chaos and few sign-ups: the first day of twitter blue was messy. The Guardian. https://www.theguardian.com/technology/2023/apr/21/elon-musk-twitter-blue-rollout, 2023. (Accessed 09-10-2023).

[51] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1900.

[52] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153 – 163, 2017.

[53] Sarah Perez. Twitter testing government ID-based verification, new screenshots show. TechCrunch. https://techcrunch.com/2023/03/20/twitter-testing-government-id-based-verification-new-screenshots-show/, 2023. (Accessed 08-18-2023).

[54] Karena Phan. Social media videos push baseless conspiracy theory that blue items were spared from maui wildfires. AP News. https://apnews.com/article/fact-check-conspiracy-blue-items-maui-wildfires-118319149774, 2023. (Accessed 09-10-2023).

[55] Monica Prasad, Andrew J. Perrin, Kieran Bezila, Steve G. Hoffman, Kate Kindleberger, Kim Manturuk, and Ashleigh Smith Powers. "there must be a reason": Osama, saddam, and inferred justification. *Sociological Inquiry*, 79(2):142–162, 2009.

[56] Prolific. Representative samples. https://researcher-help.prolific.co/hc/en-gb/articles/360019236753-Representative-samples, 2023. (Accessed 09-10-2023).

[57] Adrian E. Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[58] Harry T. Reis and Charles M. Judd. *Handbook of research methods in social and personality psychology*. Cambridge University Press, 2000.

[59] David E. Sanger and Julian E. Barnes. U.S. Warns Russia, China and Iran Are Trying to Interfere in the Election. Democrats Say It's Far Worse. *The New York Times*, July 2020. https://www.nytimes.com/2020/07/24/us/politics/election-interference-russia-china-iran.html.

[60] Craig Silverman. Verification and Fact Checking. European Journalism Centre. https://datajournalism.com/read/handbook/verification-1/additional-materials/verification-and-fact-checking. (Accessed 09-06-2023).

[61] Craig Silverman and Rina Tsubaki. A guide to verifying digital content in emergencies. Global Investigative Journalism Network. https://gijn.org/2014/03/18/a-guide-to-verifying-digital-content-for-emergency-coverage/, 2014. (Accessed 09-06-2023).

[62] Snapchat. How to verify your public profile. https://businesshelp.snapchat.com/s/article/public-profile-verify?language=en_US. (Accessed 09-10-2023).

[63] Elliott Sober. Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science*, 69(S3):S112–S123, 2002.

[64] Truth Social. Red check verification. https://help.truthsocial.com/moderation/how-to-get-verified. (Accessed 09-10-2023).

[65] Todd Spangler. Elon Musk Says Twitter 'Final Date' for Removing Legacy Blue Check-Marks Is 4/20. Variety. https://variety.com/2023/digital/news/twitter-musk-date-removal-blue-checkmarks-legacy-1235570782/, 2023. (Accessed 09-10-2023).

[66] Mariana Spring and Laura Gozzi. Twitter blue tick: Multiple hillarys and new yorks as verifications disappear. BBC News. https://www.bbc.com/news/technology-65346263, 2023. (Accessed 09-10-2023).

[67] Biz Stone. Not Playing Ball. Twitter. https://blog.twitter.com/official/en_us/a/2009/not-playing-ball.html, June 2009. (Accessed 07-18-2023).

[68] Anselm Strauss and Juliet Corbin. *Basics of qualitative research*, volume 15. Newbury Park, CA: Sage, 1990.

[69] S. Shyam Sundar. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *Digital media, youth, and credibility*, 73100, 2008.

[70] Pete Syme. Elon musk's war against twitter bots isn't going very well. next, you'll have to pay to dm those who don't follow you. Business Insider. https://www.businessinsider.com/elon-musk-war-on-twitter-bots-isnt-working-limits-dms-2023-6, 2023. (Accessed 09-10-2023).

[71] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385, Boston, MA, August 2022. USENIX Association.

[72] TikTok. Verified accounts on tiktok. https://support.tiktok.com/en/using-tiktok/growing-your-audience/how-to-tell-if-an-account-is-verified-on-tiktok. (Accessed 09-10-2023).

[73] Twitter. How To Get the Blue Checkmark on X. Twitter. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts. (Accessed 09-10-2023).

[74] Twitter. Legacy verification policy. https://help.twitter.com/en/managing-your-account/legacy-verification-policy. (Accessed 09-10-2023).

[75] Tavish Vaidya, Daniel Votipka, Michelle L. Mazurek, and Micah Sherr. Does being verified make you more credible? account verification's effect on tweet credibility. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[76] Sarah Vieweg. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. *Computer Supported Collaborative Work*, pages 515–516, 2010.

[77] James Vincent. Twitter says paying blue subscribers now get 'prioritized rankings in conversations'. https://www.theverge.com/2022/12/23/23523845/twitter-blue-paying-priority-replies-conversations, 2022. (Accessed 09-10-2023).

[78] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[79] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating Information: The Cornerstone of Civic Online Reasoning. *Stanford Digital Repository*, 2016.

[80] Stefan Wojcik and Adam Hughes. Sizing Up Twitter Users. Pew Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/, 2019. (Accessed 08-01-2019).

[81] Madelyne Xiao, Mona Wang, Anunay Kulshrestha, and Jonathan Mayer. Account verification on social media: User perceptions and paid enrollment. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3099–3116, Anaheim, CA, August 2023. USENIX Association.

A Overview

In our appendices, we describe our web scraping process for policy collection (Section B), provide our survey text (Section C), and additional tables and figures not included in the main paper for brevity (Section D. The full set of mock posts shown to users in our survey, the full codebook of free-response questions, demographic questions, debrief text, and the timeline of policy changes we observed can be found at https://osf.io/a9y3j/?view_only=d2608dffe87f40c09885c4e55637ddeb.

B Policy Review Web Scraping Process

To capture each platform's verification policies, we created a simple web scraper in Python using the BeautifulSoup4 and Selenium libraries. This script was run daily to pull each policy, compare it to the prior version, and record changes. Because we began our collection in February 2023, we used the Internet Archive's Wayback Machine [3] to collect older changes to the platforms' policies. Therefore, our review could be an under-approximation of changes in the period prior to our direct collection. However, we note that we were able to capture all major changes to Twitter reported in the news, and no other platform had major changes during this period. This process generated a dataset of timestamped verification policy changes for each platform.

C Survey Questionnaire

In this appendix, we provide the full text of our survey for one particular condition (Twitter post with police content with the verified indicatorassigned to the declarative statement). Throughout, we provide heading indicating the section of the survey as shown in Figure 1. These headings were not included in the survey shown to participants and are only included here for readability.

Survey begins

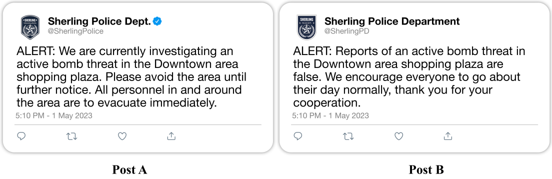
(Consent, Part A)
[Survey Consent presented here]

page break

In this study, we will display a pair of social media posts and ask you questions about the content shared in the posts.

page break

(Implicit Measure of Credibility, Part B)



Sherling Police Dept. ✓
@SherlingPolice

ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area until further notice. All personnel in and around the area are to evacuate immediately.

5:10 PM - 1 May 2023

Sherling Police Department
@SherlingPD

ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false. We encourage everyone to go about their day normally, thank you for your cooperation.

5:10 PM - 1 May 2023

Post A

Post B

Please answer the following questions considering the two Twitter posts above.

1. Post A and Post B contain conflicting information. Which of the posts do you believe is correct?

   (a) Definitely A

   (b) Probably A

   (c) Equally likely to be A or B

   (d) Probably B

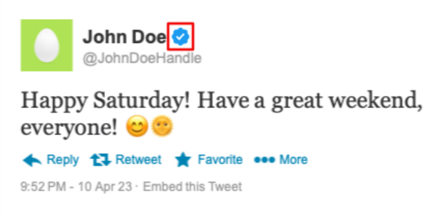
   (e) Definitely B

page break

(Explicit Measure of Credibility, Part C)

In this section, we will ask you some questions about how you determined which Twitter post was more correct in the previous section. Specifically, we will highlight different elements of the post and ask you how much each element influenced your decision. To help you know which visual element we're asking about, we show a different Twitter post, distinct from the posts you saw before, and highlight the element in question.

A VA is denoted by a blue checkmark shown next to the display name, as illustrated within the red box below:



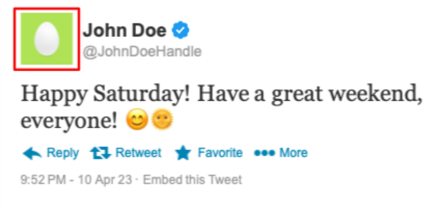
John Doe ✓
@JohnDoeHandle

Happy Saturday! Have a great weekend, everyone! 😊😶

9:52 PM - 10 Apr 23 · Embed this Tweet

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the presence of this verified account indicator (✓) affect your decision?
   (a) No Effect
   (b) Minor Effect
   (c) Moderate Effect
   (d) Major Effect

Every post on Twitter includes the display of the user's profile picture next to their handle or username, as exemplified by the red box in the example post below:



John Doe ✓
@JohnDoeHandle

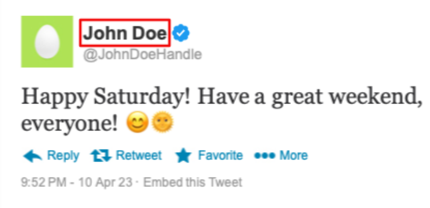
Happy Saturday! Have a great weekend, everyone! 😊😶

9:52 PM - 10 Apr 23 · Embed this Tweet

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the account's *profile picture* affect your decision?
   (a) No Effect
   (b) Minor Effect
   (c) Moderate Effect
   (d) Major Effect

A display name is used to identify the account and can differ from the username. On Twitter, it appears next to the account's profile picture as shown by the red box in the example post below:

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the account's *display name* affect your decision?



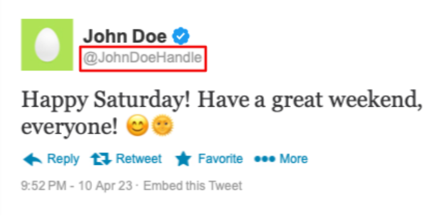
John Doe ✓
@JohnDoeHandle

Happy Saturday! Have a great weekend, everyone! 😊😶

9:52 PM - 10 Apr 23 · Embed this Tweet

   (a) No Effect
   (b) Minor Effect
   (c) Moderate Effect
   (d) Major Effect

On Twitter, a user's handle (also known as their username) is presented next to their profile picture on every tweet they post, and it is marked by the "@" symbol. An example of a user's handle is provided in the red box below:



John Doe ✓
@JohnDoeHandle

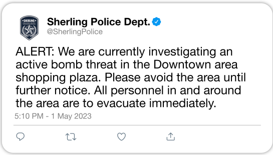
Happy Saturday! Have a great weekend, everyone! 😊😶

9:52 PM - 10 Apr 23 · Embed this Tweet

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the account's handle affect your decision?
   (a) No Effect
   (b) Minor Effect
   (c) Moderate Effect
   (d) Major Effect

- - - - - - - - - - - - - - - - - - - *page break* - - - - - - - - - - - - - - - - - - -

(Participants' Definition of Verification, Part D)

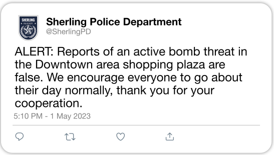


Sherling Police Dept. ✓
@SherlingPolice

ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area until further notice. All personnel in and around the area are to evacuate immediately.

5:10 PM - 1 May 2023

Post A



Sherling Police Department
@SherlingPD

ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false. We encourage everyone to go about their day normally, thank you for your cooperation.

5:10 PM - 1 May 2023

Post B

One of the tweets you were previously shown was by an account with a verification checkmark (✓) indicating that the account has been verified.

1. Based on your understanding of Twitter's account verification, what requirements must an account satisfy to become verified and obtain a verified checkmark?

------------------- *page break* -------------------

(Show Assigned Policy, Policy Acceptability, Suggested Changes, Part E)

Suppose Twitter adopted a verification policy in which the account had to meet all of the following criteria:

- **Any user** on the platform is allowed to apply for verification Accounts must submit government-issued identification that matches the name of the account being verified
- Any user on the platform is allowed to apply for verification Accounts must **submit government-issued identification** that matches the name of the account being verified
- Accounts **pay a monthly subscription fee** to maintain their verification checkmark

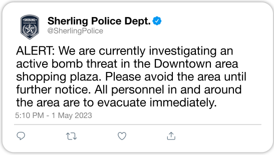
1. To what level do you believe these verification requirements are acceptable for verifying account owner identity?
    (a) Unacceptable
    (b) Slightly Unacceptable
    (c) Neutral
    (d) Slightly Acceptable
    (e) Acceptable
2. If you could suggest one thing to add, remove, or change in this policy to improve its ability in verifying the account owner is who they say they are, what would it be? Please explain why.

------------------- *page break* -------------------

(Credibility Measure After Policy Priming, Part F)

We will now ask you to revisit the Twitter posts you were shown previously, and answer the following questions assuming this new policy was used for verification.

We display the Twitter posts and the new verification policy below for you to reference while you answer the questions.

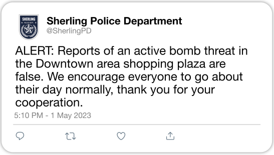


Sherling Police Dept. ✓
@SherlingPolice

ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area until further notice. All personnel in and around the area are to evacuate immediately.

5:10 PM - 1 May 2023

Post A



Sherling Police Department
@SherlingPD

ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false. We encourage everyone to go about their day normally, thank you for your cooperation.

5:10 PM - 1 May 2023

Post B

- **Any user** on the platform is allowed to apply for verification Accounts must submit government-issued identification that matches the name of the account being verified
- Any user on the platform is allowed to apply for verification Accounts must **submit government-issued identification** that matches the name of the account being verified
- Accounts **pay a monthly subscription fee** to maintain their verification checkmark

1. Which of the following most closely resembles the subject matter of the two posts?
    (a) Police investigating a bomb threat
    (b) Effects of coffee on health
    (c) Food recall due to E. coli outbreak
2. After reviewing the criteria required for an account to receive a verification checkmark, which of the posts do you believe is correct?
    (a) Definitely A
    (b) Probably A
    (c) Equally likely to be A or B
    (d) Probably B
    (e) Definitely B
3. If a friend of yours was unsure about which post to trust, what would you say to this friend to help them decide?

------------------- *page break* -------------------

(Social Media Use, Part G)

Now we will end the survey with several short questions concerning your social media use and and demographics.

1. Which of the following social media platforms do you currently have an account with? Select all that apply.
    - Twitter
    - Facebook
    - Instagram
    - LinkedIn
    - TikTok
    - YouTube
    - Other (please specify)
2. How often do you use Twitter in any given week?
    (a) Daily
    (b) Every other day
    (c) Every two days
    (d) Once a week
    (e) I do not use Twitter
3. How often do you use Facebook in any given week?
    (a) Daily
    (b) Every other day
    (c) Every two days
    (d) Once a week
    (e) I do not use Facebook
4. How much time do you spend on social media sites per day?
    (a) Less than 30 minutes

(b) 30 minutes-1 hour

(c) 1-2 hours

(d) 2-4 hours

(e) 5-6 hours

(f) Greater than 6 hours

- - - - - - - - - - - - - - - - - - *page break* - - - - - - - - - - - - - - - - - -

(Security Attitudes, Part H)

Each statement below describes how a person might feel about the use of security measures. Examples of security measures are laptop or tablet passwords, spam email reporting tools, software updates, secure web browsers, fingerprint ID, and anti-virus software.

Please indicate the degree to which you agree or disagree with each statement. In each case, make your choice in terms of how you feel right now, not what you have felt in the past or would like to feel.

1. I seek out opportunities to learn about security measures that are relevant to me

   (a) Strongly disagree
   (b) Somewhat disagree
   (c) Neither disagree nor agree
   (d) Somewhat agree
   (e) Strongly agree

2. I am extremely motivated to take all the steps needed to keep my online data and accounts safe.

   (a) Strongly disagree
   (b) Somewhat disagree
   (c) Neither disagree nor agree
   (d) Somewhat agree
   (e) Strongly agree

3. Generally, I diligently follow a routine for security practices.

   (a) Strongly disagree
   (b) Somewhat disagree
   (c) Neither disagree nor agree
   (d) Somewhat agree
   (e) Strongly agree

4. I often am interested in articles about security threats.

   (a) Strongly disagree
   (b) Somewhat disagree
   (c) Neither disagree nor agree
   (d) Somewhat agree
   (e) Strongly agree

5. I always pay attention to experts' advice about the steps I need to take to keep my online data and accounts safe.

   (a) Strongly disagree
   (b) Somewhat disagree
   (c) Neither disagree nor agree
   (d) Somewhat agree
   (e) Strongly agree

6. I am extremely knowledgeable about all the steps needed to keep my online data and accounts safe.

   (a) Strongly disagree
   (b) Somewhat disagree
   (c) Neither disagree nor agree
   (d) Somewhat agree
   (e) Strongly agree

Factor	Description	Baseline
Posted content Variables		
Platform	The assigned visual design used to display posts	Twitter
Content type	The assigned content condition	Coffee
Position	The side of the argument the verified indicator was assigned to	Contradict.
Policy Variables[1]		
Availability	Who can become verified?	Open
Verification Method	How are accounts verified?	Phone
Payment	Is payment required to become verified?	Paid
Social Media Experience		
Twitter experience	Does the participant report using Twitter (binary)	False
Facebook experience	Does the participant report using Facebook (binary)	False
Social Media Accts.	Number of social media platforms participants use	–
Demographics		
SA-6	Participant's score on Faklaris et al.'s SA-6 scale [18]	–
Age	Age of participant	–
Gender	Gender of participant	Male
Education	Does the participant hold a B.S. or higher degree (binary)	False

[1] Policy variables were only included when considering participants' policy acceptability rating (Part E) and their credibility perceptions after providing them with a mock policy (Part F).

Table 6: Factors used in regression models. Categorical variables are compared individually to the given baseline.

D Additional Tables and Figures

Finally, we provide tables and figures excluded from the main text for brevity. This includes a summary of the variables in the initial model for each regression (Table 6), additional participant demographics information (Table 7), and a summary of participants' responses regarding perceive impact of each account feature (Figure 6).

Metric	%
Gender	
Woman	49.9%
Man	48.4%
Non-binary	1.2%
Transgender/ Agender	0.3%
Other	0.2%
Race/Ethnicity	
White	73.9%
Black	11.6%
Asian	6.0%
Hispanic or Latino/a	4.9%
Indigenous	0.7%
Two or More Races	2.0%
Other	0.2%
Prefer not to respond	0.6%

Metric	%
Income	
<$10k	10.6%
$10k-$25k	14.8%
$25k-$50k	25.1%
$50k-$75k	19.1%
$75k-$100k	11.9%
$100k-$150k	10.4%
$150k+	5.1%
Prefer not to respond	3.1%

Table 7: Additional participant demographics.

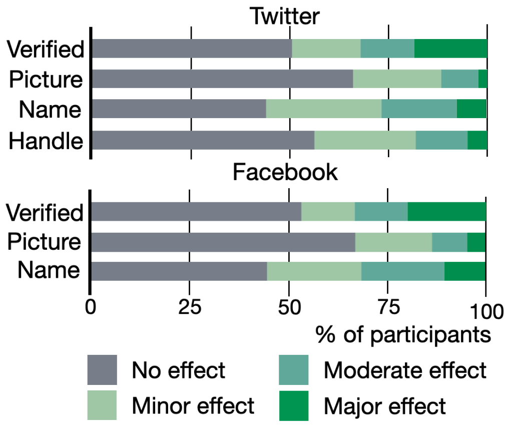


Figure 6: Likert-scale response indicating how much participants perceived each account feature impacted their credibility decision, organized by assigned social media platform.

E Demographics Questions & Debrief

(Demographics, Part I)

1. What is your age?
2. How do you describe your gender identity?
   (a) Female
   (b) Male
   (c) Agender
   (d) Non-binary
   (e) Gender-queer
   (f) Not sure
   (g) Not listed above [with text entry]
   (h) Prefer not to respond
3. Do you identify as Hispanic and/or Latino?
   (a) Yes
   (b) No
   (c) Prefer not to respond
4. What level of education have you attained?
   (a) Less than high school
   (b) High School graduate (high school diploma or equivalent such as GED)
   (c) Some college, but no degree
   (d) Associate Degree
   (e) Bachelor's Degree
   (f) Master's Degree
   (g) Professional Master's Degree (JD, MD)
   (h) Doctorate Degree
   (i) Prefer not to respond
5. What was your 2020 taxed income?
   (a) Less than $10,000
   (b) $10,000-$24,999
   (c) $25,000-$49,000
   (d) $50,000-$74,999
   (e) $75,000-$99,999
   (f) $100,000-$149,000
   (g) $150,000 and greater
   (h) Prefer not to respond
6. Do you get the majority of your earnings from Prolific or similar platforms?
   (a) Yes
   (b) No
   (c) Prefer not to respond

------------------- *page break* -------------------

(Debrief, Part J)

Throughout this study you were shown social media posts showing conflicting reports about a particular event or research findings. These events are completely fictional and not based on any true events or findings. For the purpose of this study, these were made up to avoid bias in participant responses.

You were also given a set of criteria used for social media verification. Although the verification criteria we used for this study was based on the verification criteria Twitter and Facebook use to verify accounts on their platforms, the criteria you saw does not reflect the true criteria Twitter and Facebook use for their verification policies.

The verification process Twitter uses can be viewed in full by following this link. In this policy, verification is open to anyone but requires the owner of the account to pay a monthly fee to maintain the verification checkmark. The account must have a display name and profile photo. This display name and profile photo cannot be modified once the account has been verified. The account owner also must confirm a phone number with Twitter. Additionally, the account must show no signs of engaging in platform manipulation or spam, and show no signs of being misleading or deceptive.

The verification process Facebook uses can be viewed in full by following this link. This process is used for verifying accounts owned by public figures, celebrities, or notable brands. Notable brands are those that represent well-known, often searched for brands that are unique (i.e. be the only presence of this business), authentic (i.e. registered business), and have a complete Facebook Page or Facebook Profile (i.e. the account has a completed "About" section, has shared at least one post, and show recent activity.

Facebook also offers account profile verification for all accounts via Meta Verification. To be eligible for Meta Verification the account owner must be at least 18 years of age, have a public or private Facebook profile with the account owner's full name and a profile picture that matches a government issued ID. Additionally, the account must have a prior posting history, have two-factor authentication enabled. You can learn more about Meta Verification and its process here.

It can be difficult to determine whether information garnered online is true or false. However, there are steps you can take to help confirm if the information you read online is true or meant to mislead you. We provide links to several guides below for verifying digital content and fact checking information online below:

- 5 Ways You Can Fact-Check Online Claims
- A Guide to Verifying Digital Content in Emergencies
- Verification and Fact Checking - A General Guide