**Evaluating Misinformation Warnings Across Facebook, Instagram, X and Threads**

Riley Howard Gokul Gopakumar Siddharth Mishra

*Pennsylvania State University* *Pennsylvania State University* *Pennsylvania State University*

# Abstract

Misinformation has emerged as a critical challenge on social media platforms, particularly during pivotal events like elections. This study investigates the effectiveness of misinformation warnings on four major platforms—Facebook, Instagram, Threads, and Twitter/X—focusing on the lead-up to the 2024 U.S. Presidential Election. The primary research questions address how design features of warnings influence user trust and engagement, how perceptions of platform bias shape these responses, and the impact of political ideology and social media habits. Using a survey of 19 participants from Penn State University, we analyzed user interactions with static screenshots of misinformation posts accompanied by platform-specific warnings. Results revealed that Instagram's warnings were rated highest in clarity and effectiveness, while Threads and Twitter/X faced challenges with user trust and perceived bias. Our findings highlight the importance of transparency, context-rich warnings, and neutral designs in fostering trust and reducing misinformation engagement. Overall, this research contributes actionable insights for designing effective misinformation warnings and underscores the role of platform reputation in content moderation strategies.

# 1 Introduction

The increasing abundance of misinformation on social media poses significant risks to public discourse, especially during critical moments such as elections and public health crises. Despite platforms’ attempts to curb misinformation through warning systems, their effectiveness remains inconsistent and influenced by user perceptions, platform design, and political biases.

This study aims to provide a comprehensive analysis of misinformation warnings across four major platforms—Facebook, Instagram, Threads, and Twitter/X—in the context of the 2024 U.S. Presidential Election. The objective of this work is to assess how warning design features, platform trustworthiness, and perceived bias influence the effectiveness, trust, accuracy, and clarity of misinformation warnings. Existing research often focuses on individual platforms or specific types of misinformation, leaving a gap in comparative analyses across multiple platforms and political contexts. This study addresses that gap by exploring user perceptions of trust, clarity, accuracy, and engagement in response to platform-specific warnings. Key findings from this research reveal that while design elements like clarity and contextual information significantly improve warning effectiveness, user trust is deeply tied to perceptions of platform bias. The findings underscore the need for transparent and user-centered approaches to misinformation, offering practical recommendations for enhancing warning systems. By focusing on user behavior and platform dynamics, this study provides valuable insights for researchers and practitioners aiming to combat the spread of misinformation on social media.

# 2 Related Work

Research on misinformation warnings has primarily focused on specific domains, such as health misinformation or single platforms. Studies addressing health misinformation warnings exploring their impact on reducing the spread of misleading claims about vaccines or health practices [2, 3, 9, 10, 13, 14]. These works highlight the importance of clear, actionable warnings, though they often limit their scope to isolated platforms or specific public health crises, such as COVID-19. For instance, an analysis by Krishnan et al. [2021] showed that most social media platforms, except Twitch, Tumblr, Messenger, and WhatsApp, had policies banning COVID-19 misinformation. Yet, it also highlighted serious inconsistencies and a general lack of transparency across the platforms, which raises the need for more clarity and transparency from platforms to combat misinformation, especially in fast-changing public health emergencies like the COVID-19 pandemic. Existing studies also demonstrate that the presence of warnings is not the only factor, rather, the clarity and quality of the corrective information also act as main moderators in shaping users’ perceptions and behaviors related to health-related misinformation [2, 13, 14].

Existing research on political misinformation warnings often focus on a single platform, such as Facebook or Twitter [1, 6, 11, 14]. These studies assess metrics like engagement reduction, behavioral change, and user perceptions of the platforms' credibility. For instance, Margolin et al. [2018] and Grinberg et al. [2019] assessed the impact of these warnings in relation to political ideology and user engagement. Margolin et al. [2018] highlighted that the acceptance of the correction to misinformation may depend on the underlying social structure of the misinformation warning. Based on their analysis on Twitter, they showed that a correction is more likely to be accepted if it originates from people with whom the participants share stronger social connections. Additionally, Grinberg et al. [2019]found that Twitter's misinformation warnings were more effective for users who already possessed moderate to high levels of trust in the platform.

Existing research indicates that misinformation warnings on social media can have varying levels of effectiveness depending on their design and implementation [11]. For instance, Guo et al. [2023] conducted a study on video-sharing platforms and found that while contextual warnings do not always lead to behavioral adherence, they can increase user vigilance towards misinformation. They also found that users’ perceptions of warnings are influenced by the warning’s explicitness and the risk level of the misinformation. Other studies also explore the diversity of user preferences for misinformation warnings and the factors that influence them, including political orientation and cognitive style. Enhanced warning tags that provide meaningful context and standard iconography are generally preferred, especially among left-leaning and moderate users [15]. However, users with conservative leanings may exhibit distrust towards these systems [2].

What has yet to be explored is a comprehensive comparison of misinformation warnings across Facebook, Instagram, Twitter/X, and Threads and the impact of perceived platform bias and participant ideology on the effectiveness of warnings. This study broadens the scope by assessing multiple platforms and focusing on political misinformation in the lead-up to the 2024 Presidential Election, creating a more comprehensive analysis of misinformation warning effectiveness. This study bridges that gap by conducting a comparative analysis, focusing on design effectiveness and user-centric factors like perceived platform bias and trustworthiness. Understanding user perceptions of trust and bias allows this study to investigate not only how warnings can be more effective in reducing misinformation but also how platform reputation, warning trust, and perceived impartiality shape the effectiveness of these warnings. Notably, to the best of our knowledge, we are the first to conduct an empirical analysis of warning effectiveness on Threads.

# 3 Experiment

**3.1 Research Questions**

To evaluate the impact of different misinformation warning designs on user perceptions and engagement, we present four research questions that will be addressed in the study:

**RQ1**: What design features of misinformation warnings impact user’s perceived trust in the warning, perceived accuracy, clarity, and effectiveness of the warning?

**RQ2:** Which platform's misinformation warning is rated the highest/lowest for perceived accuracy, clarity, effectiveness, and the likelihood of engaging with misinformation?

**RQ3:** How do perceptions of platform bias affect users’ perceived trust in the warning, perceived accuracy, clarity, effectiveness of the warning, and likelihood to engage with the content (like, share, comment)?

**RQ4:** How dosocial media use and political ideologyimpact users’ trust in each platform and warning, perceived accuracy of each platform’s warning, and likelihood to engage with the content (like, share, comment)?

**3.2 Participants**

Participants were recruited from Penn State University via text or email to complete the online survey hosted on Qualtrics. No compensation was provided, and participation was voluntary. Of the 35 participants who began the survey, 19 completed it. Excluding the one outlier of five hours, the median time for participants to complete the survey was about 11 minutes. The average age of participants was 22.7. Almost all of our participants were from the United States (94.7%), with one participant from Taiwan. Generally, the most common area of study among our participants is Information Sciences and Technology (63.2%). Within the field of Information Sciences and Technology, participants reported the following majors: Security and Risk Analysis, Cybersecurity, and Informatics. Our participant pool is skewed Democratic, with 63.1% identifying as either “Consistently Democratic” or “Mostly Democratic”. Meanwhile, 21.1% identified as “Moderate/Independent,” and 15.8% identified as “Mostly Conservative,” with no participants identifying as “Consistently Conservative.” Additionally, we ran a pilot study with three participants, which gave us feedback on the survey design and allowed us to improve before officially launching the survey. Improvements included rewording questions to improve clarity and reducing the number of open-ended questions.

**Table 1:** Demographics of Participants in the Online Survey

|  |  |  |
| --- | --- | --- |
| **Item** | **Option** | **Percentage** |
| Gender | Female | 63.2% |
| Male | 36.8% |
| Age | 20  21  22  23  24  25  28 | 5.3%  10.5%  42.1%  21.1%  10.5%  5.3%  5.3% |
| Nationality | United States of America  Taiwan, Province of China | 94.7%  5.3% |
| Major | Security and Risk Analysis  Cybersecurity  Informatics  Nursing  Geology  Education  Law  Public Relations  Agribusiness Management | 31.6%  15.8%  15.8%  10.53%  5.3%  5.3%  5.3%  5.3%  5.3% |
| Political Ideology | Consistently Democratic  Mostly Democratic  Moderate/Independent  Mostly Conservative  Consistently Conservative | 26.3%  36.8%  21.1%  15.8%  0% |

**3.3 Stimuli**

For this study, we used a set of social media posts accompanied by misinformation warnings from four platforms: Facebook, Instagram, Threads, and Twitter/X (See Figure 1). These posts focused on political misinformation relevant to the 2024 U.S. Presidential Election. To keep the experience authentic, we preserved the native warning designs and platform-specific interface elements in each post. Participants viewed these as static screenshots embedded within a Qualtrics survey, allowing us to create a realistic environment for evaluation. To simulate the real social media environment, each post included the platform’s native warning design, interface elements, and a link to the original post embedded in the image. To minimize ideological bias, we included two posts from each platform: one containing Democratic-leaning misinformation and one containing Republican-leaning misinformation. This balanced approach helped ensure that participants evaluated the warnings based on their design and content rather than personal political ideologies. Participants viewed the posts as static screenshots embedded in a Qualtrics survey, where they were prompted to evaluate each post and its misinformation warning. They answered questions arranged on a assessing the trustworthiness, accuracy, effectiveness, and clarity of the warnings, as well as their likelihood to engage with the flagged content by liking, sharing, or commenting. Additionally, before viewing the posts, we asked general questions about participants’ perceptions of platform trustworthiness and bias to explore how their pre-existing beliefs influenced their evaluations. To collect further insights, we included optional open-ended questions at the end of the survey that encouraged participants to provide qualitative feedback on what they found effective or ineffective about each warning and how they would improve the warning design.

A screenshot of a social media page

Description automatically generated

**Figure 1**: Misinformation Warning Design Per Platform

Since we did not want to overload the participants, we did not evaluate a baseline for each platform (e.g., showing posts without warning) because that would require participants to evaluate twice as many posts, increasing the amount of time participants have to dedicate to the survey, which could reduce the number of responses we receive. Eliminating a baseline condition allows for a more streamlined, efficient study design that minimizes participant fatigue and keeps their focus on assessing each warning. Additionally, prior studies have generally established that posts with no warnings tend to generate more trust and engagement compared to posts with warnings [2, 4, 10]. This allows our study to build on existing findings by focusing on relative differences between warnings rather than reiterating the already well-documented effect of “no warning” scenarios.

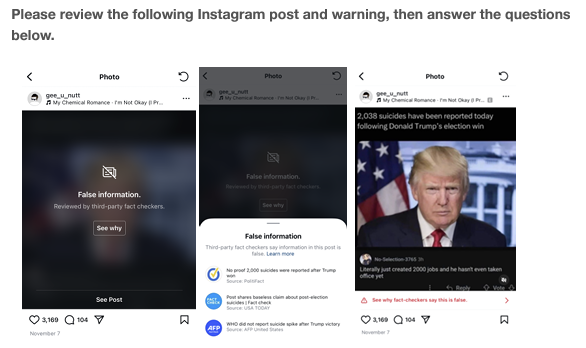
**3.4 Procedure**

The entire process was conducted online using a Qualtrics survey and included the following steps: informed consent, pre-survey, stimuli presentation, evaluation of stimuli, and post-survey.

**Informed Consent.** Participants were first presented with an introduction detailing the study’s purpose, their rights, and the voluntary nature of participation. They were informed about potential risks, such as discomfort from viewing political misinformation, and the confidentiality of their responses. Participants proceeded only after providing their consent.

**Pre-Survey.** Participants answered a series of demographic questions, including their age, gender, academic major, country of citizenship, and political ideology. They also answered questions about their social media usage frequency and their levels of trust in information posted on each platform (Facebook, Instagram, Threads, and Twitter/X).

**Stimuli Presentation.** Participants reviewed eight static screenshots of social media posts, with two posts from each platform. Each platform featured one post containing misinformation about presidential candidate Kamala Harris and one containing misinformation about presidential candidate Donald Trump to minimize ideological bias. Each screenshot displayed the post content and the platform's native misinformation warning (See Figure 2). We also embedded a clickable link to the original post to enhance authenticity and allow the participant to view the post on social media in a new window. The order of the eight posts was randomized to reduce the likelihood of order effects influencing participant responses.



**Figure 2:** Example Demonstrating How Each Post was Presented to Participants in the Survey

**Evaluation of Stimuli.** After viewing one of the posts, participants answered a set of standardized Likert-scale questions evaluating their perceived trustworthiness of the misinformation warning, accuracy of the warning in identifying misleading content, effectiveness of the warning in preventing the spread of misinformation, clarity of the warning in providing context about the flagged post, and likelihood to engage with the post (e.g., liking, sharing, or commenting). The screenshots of the corresponding post for evaluation were located at the top of each page, followed by the five Likert-scale questions. When the participants completed the five questions, they navigated to the next page to view the next post and set of questions. They did this for each of the eight posts. To ensure further randomness, the order of these five questions was also randomized for each participant. Separating each post and its associated questions onto individual pages in Qualtrics ensured that participants focused on one post at a time, reducing cognitive load and potential confusion.

**Post Survey.** After reviewing all eight posts, participants completed post-survey questions evaluating the overall quality and effectiveness of misinformation warnings on each platform. They also rated their trust in each platform after viewing the warnings, indicating whether or not viewing the warning impacted their trust in the platform. Additionally, we included optional open-ended questions for participants to provide qualitative feedback on the warning designs, including suggestions for improvement and insights into what made the warnings more or less effective.

**3.5 Results**

***RQ1: What design features of misinformation warnings impact user’s perceived trust in the warning, perceived accuracy, clarity, and effectiveness of the warning?***

At the end of the survey, we asked participants three optional open-ended questions to gain qualitative insights into how viewing the warnings impacted users’ trust in the platform, what made each warning more or less effective in making them question the truthfulness of the post, and changes participants would make to improve the warnings. Across these three open-ended questions, we received twenty responses from 9 participants, which we used for our thematic analysis. By conducting a thematic analysis, we identified two major themes:

**Design and Presentation of Effective Warnings.** One of the main themes we identified in the qualitative data is that warnings should be visually distinct, prominently displayed, and integrated directly into the user experience. For example, four participants highlighted the effectiveness of the blurring technique used by Facebook and Instagram, which requires users to take an additional step to view the content and makes the warning much more difficult to overlook. One participant specifically indicated, *“It must be disruptive and cut into the user experience because misinformation should never be integrated into the user experience.”* Additionally, two participants highlighted how misinformation warnings should be visually distinct from other types of warnings, such as explicit or AI-generated content. Finally, four participants highlighted how warnings should minimize user effort by making factual information about the misinformation readily accessible. Extra steps, like requiring users to access fact-checking websites, can reduce the effectiveness of warnings. For instance, one participant suggested *“Providing further explanation as to why the user’s post was flagged for false information.”* This finding is consistent with existing research on the use of interstitial warnings, which interrupt the user experience to present a warning. These warnings are found to be highly effective in prompting users to seek information from alternative sources and reducing the perceived credibility of misinformation [8].

**Transparency and Independent Verification.** The other major theme we identified in the qualitative data is providing multiple, clearly identified sources and links to trusted, unbiased third-party organizations enhances the credibility of warnings. Three participants specifically indicated that using multiple reputable sources increased their perceived effectiveness of warnings. For instance, one participant stated, *"The addition of multiple fact-checking sources as well as a clear description of the context was more effective in making me question the truthfulness."* Additionally, two participants indicated how detailed explanations about why a post was flagged, along with accessible links to verified information, could help counter skepticism and ensure that warnings feel justified and credible. For example, one participant suggested, *“Show people the facts from the third-party checks.”*

***RQ2: Which platform's misinformation warning is rated the highest/lowest for perceived accuracy, clarity, effectiveness, and likelihood of engaging with misinformation?***

To evaluate and rank the platforms for each metric, we first assigned numerical values to each Likert scale item (e.g., "Not trustworthy at all" = 1, "Somewhat untrustworthy" = 2, “Neutral” = 3, "Somewhat trustworthy" = 4, "Completely trustworthy" = 5). For each participant, we calculated an average rating for each metric (e.g., perceived accuracy, clarity, effectiveness) across both Democratic and Republican misinformation warnings for each platform. This step ensures that within-subject differences are accounted for. Finally, the averages for all participants were aggregated to compute the overall mean rating for each platform and metric.

**Table 2:** Mean Weighted Ranking of Likert Scale Responses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Trust** | **Accuracy** | **Clarity** | **Effectiveness** | **Engagement Likelihood** |
| **IG** | 3.76 | 3.92 | 4.08 | 3.82 | 1.47 |
| **FB** | 3.76 | 3.95 | 3.63 | 3.63 | 1.42 |
| **TH** | 3.84 | 3.95 | 3.82 | 3.39 | 1.58 |
| **X** | 3.73 | 3.97 | 3.87 | 3.52 | 1.58 |

**Warning Trust.** To evaluate warning trust, we asked participants, “How trustworthy do you find the false information warning that comes with the post?” when presented with each misinformation post. Participants responded on a Likert scale, which we coded to add numerical values for the following options: "Not trustworthy at all" = 1, "Somewhat untrustworthy" = 2, “Neutral” = 3, "Somewhat trustworthy" = 4, "Completely trustworthy" = 5. Conducting a weighted average ranking, we found that Threads had the highest average trust rating (3.84), followed by Facebook and Instagram (3.76), then Twitter/X (3.73). Conducting a one-way ANOVA test, we did not find any significant variance between any of the platforms.

**Warning Accuracy.** To evaluate warning accuracy, we asked participants, “How accurate do you think the false information warning is in identifying false or misleading information in this post?” when presented with each misinformation post. Participants responded on a Likert scale, which we coded to add numerical values for the following options: "Not accurate at all" = 1, "Somewhat inaccurate" = 2, “Neutral” = 3, "Somewhat accurate" = 4, "Very accurate" = 5. Conducting a weighted average ranking, we found that Twitter/X had the highest average warning accuracy rating (3.97), closely followed by Facebook and Threads (3.95), then Instagram (3.92). Conducting a one-way ANOVA test, we did not find any significant variance between any of the platforms.

**Warning Clarity.** To evaluate warning clarity, we asked participants, “How clearly do you think the false information warning provides context about the content of the post?” when presented with each misinformation post. Participants responded on a Likert scale, which we coded to add numerical values for the following options: "Not clear at all" = 1, "Somewhat unclear" = 2, “Neutral” = 3, "Somewhat clear" = 4, "Very clear" = 5. Conducting a weighted average ranking, we found that Instagram had the highest warning

**Table 3:** Descriptive Statistics for Each Metric and Platform

A table with numbers and text

Description automatically generated

clarity (4.08), followed by Twitter/X (3.87), then Threads (3.82), and lastly Facebook (3.63). Conducting a one-way ANOVA test, we did not find any significant variance between any of the platforms.

**Warning Effectiveness.** To evaluate warning effectiveness, we asked participants, “How effective do you think the false information warning is at preventing the spread of false or misleading information?” when presented with each misinformation post. Participants responded on a Likert scale, which we coded to add numerical values for the following options: "Not effective at all" = 1, "Somewhat ineffective" = 2, “Neutral” = 3, "Somewhat effective" = 4, "Very effective" = 5. Conducting a weighted average ranking, we found that Instagram had the highest ranking for warning effectiveness (3.82), followed by Facebook (3.63), then Twitter/X (3.52), and lastly Threads (3.39). Conducting a one-way ANOVA test, we did not find any significant variance between any of the platforms.

**Engagement Likelihood.** To evaluate engagement likelihood, we asked participants, “How likely are you to engage with this post (such as by liking, sharing, or commenting)?” when presented with each misinformation post. Participants responded on a Likert scale, which we coded to add numerical values for the following options: "Extremely unlikely" = 1, "Somewhat unlikely " = 2, “Neutral” = 3, "Somewhat likely" = 4, "Extremely likely" = 5. Conducting a weighted average ranking, we found that the platform participants were most likely to engage with was Threads and Twitter/X (1.58), followed by Instagram (1.47), then Facebook (1.42). The low engagement likelihood across platforms is consistent with prior work [12]. Conducting a one-way ANOVA test, we did not find any significant variance between any of the platforms.

**Overall Platform Ratings.** At the end of the survey, we asked participants to rate the quality of each platform’s false information warning design on a 5-point Likert scale (“Major improvements needed” = 1, “Some improvements needed” = 2, “Neutral/no opinion” = 3, “Minor improvements needed” = 4, and “Excellent design, no improvements needed” = 5). We found that Instagram had the highest quality false information warning design (3.79), followed by Facebook (3.00), then Twitter/X and Threads (2.21). Conducting a one-way ANOVA test, we found a significant variance between treatments (*F = 6.88776, p = 0.000383*) at *p < 0.05*. Using the post hoc Tukey HSD test, we found a significant variance between Threads and Instagram (*p = 0.0130*) and Instagram and Twitter/X *(p =* *0.0130).* We did not find a statistically significant difference between other platforms.

**Table 4:** Participants Evaluation of the Quality of Each Warning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **IG** | **FB** | **X** | **TH** |
| **No Improvements Needed** | 26.3% | 15.8% | 5.3% | 0% |
| **Minor Improvements**  **Needed** | 47.4% | 31.6% | 10.5% | 26.3% |
| **Neutral/No Opinion** | 10.5% | 15.8% | 10.5% | 5.3% |
| **Some Improvements**  **Needed** | 10.5% | 10.5% | 47.4% | 31.6% |
| **Major Improvements**  **Needed** | 5.3% | 26.3% | 26.3% | 36.8% |

***RQ3: How do perceptions of platform bias affect users’ perceived trust in the warning, perceived accuracy, clarity, effectiveness of the warning, and likelihood to engage with the content (like, share, comment)?***

Before showing participants the misinformation warning posts, we asked them, *“How biased do you think each platform is in its treatment of misinformation related to the presidential election?”* on a Likert scale, which we coded with values for analysis (“Not biased at all” = 1, “Slightly biased” = 2, “Somewhat biased” =3, “Mostly biased” = 4, and “Extremely biased” = 5). We found the average of the coded responses for each platform and found that participants found Twitter/X the most biased in its treatment of misinformation related to the presidential election (3.95), followed by Facebook (3.32), then Instagram (2.89), and lastly Threads (2.84). We then conducted a Spearman-ranked correlation analysis to determine if there is a relationship between platform bias and each warning metric. We conducted this analysis using both the average ranking for each participant, for each metric and platform and the individual rankings for each post (Democratic and Republican) for each platform and metric. The data indicate that perceptions of platform bias impact users’ evaluations and engagement differently across platforms. The most notable effects were observed on Instagram and Twitter/X, while no significant relationships were found on Facebook or Threads. Overall, we found platform bias did not consistently influence user perceptions of warning attributes (trust, accuracy, effectiveness, clarity, and engagement) across all platforms.

**Table 5:** Relationship Between Platform Bias and Each Metric for Each Post and Combined Average

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Warning Trust | | Warning Accuracy | | Warning Clarity | | Warning Effectiveness | | Engagement Likelihood | |
|  | ***rs*** | ***p*** | ***rs*** | ***p*** | ***rs*** | ***p*** | ***rs*** | ***p*** | ***rs*** | ***p*** |
| IG Avg | **-0.502** | **0.029** | **-0.534** | **0.019** | -0.103 | 0.674 | -0.246 | 0.311 | -0.152 | 0.533 |
| IG (D) | -0.495 | 0.031 | **-0.470** | **0.043** | 0.072 | 0.771 | -0.148 | 0.545 | -0.205 | 0.399 |
| IG (R) | -0.352 | 0.139 | -0.454 | 0.051 | -0.379 | 0.109 | -0.036 | 0.884 | 0.114 | 0.643 |
| FB Avg | -0.230 | 0.343 | -0.099 | 0.685 | -0.317 | 0.186 | -0.052 | 0.834 | -0.013 | 0.959 |
| FB (D) | -0.198 | 0.417 | -0.135 | 0.581 | -0.259 | 0.500 | -0.139 | 0.570 | -0.040 | 0.871 |
| FB (R) | -0.186 | 0.446 | 0.165 | 0.500 | -0.379 | 0.109 | 0.167 | 0.495 | 0.101 | 0.681 |
| TH Avg | -0.039 | 0.875 | -0.032 | 0.175 | 0.221 | 0.363 | -0.056 | 0.821 | 0 | 1 |
| TH (D) | 0.0032 | 0.898 | -0.160 | 0.514 | 0.187 | 0.444 | 0.047 | 0.849 | -0.069 | 0.779 |
| TH (R) | -0.169 | 0.488 | 0.145 | 0.348 | 0.388 | 0.100 | 0.101 | 0.680 | 0.103 | 0.675 |
| X Avg | 0.237 | 0.328 | 0.242 | 0.319 | 0.145 | 0.554 | 0.207 | 0.395 | **-0.561** | **0.012** |
| X (D) | 0.104 | 0.673 | 0.143 | 0.560 | 0.178 | 0.467 | -0.029 | 0.905 | -0.175 | 0.474 |
| X (R) | 0.149 | 0.542 | 0.409 | 0.082 | -0.258 | 0.287 | 0.150 | 0.539 | **-0.562** | **0.012** |

**Instagram Key Findings.** A statistically significant relationship exists between platform bias and the average warning trust on Instagram (*rs = -0.502, p = 0.029*). This indicates that higher perceived bias reduces trust in the warnings overall for Instagram. Additionally, platform bias is negatively correlated with warning trust in the Democratic post on Instagram (*rs = -0.495, p = 0.031)*. This indicates that when users perceived more bias on the platform, they trusted the warning for Democratic posts less. However, we found no statistically significant relationship between platform bias and warning trust in the Republican post on Instagram (*rs = -0.352, p = 0.139)*. Additionally, we found a statistically significant relationship between platform bias and the average warning accuracy rating on Instagram (*rs = -0.5336, p = 0.0186).* Specifically, there is a negative correlation between platform bias and the warning accuracy of the Democratic post on Instagram (*rs = -0.470, p = 0.0425)*. This indicates that as perceived platform bias increased, participants rated warning accuracy lower. Meanwhile, the relationship for Republican posts is marginally non-significant (*rs = -0.454, p = 0.051).*

**Twitter/X Key Findings.** A statistically significant relationship exists between platform bias and the average engagement likelihood rating on Twitter/X (*rs = -0.562, p = 0.0123).* This finding indicates higher bias perceptions reduce users’ likelihood to engage with misinformation posts.For Republican posts, this effect is particularly strong*.* We found that platform bias is strongly negatively correlated with post-engagement likelihood on the Republican post (*rs = -0.562, p = 0.0122)*. This indicates that when users perceive more bias on the platform, they were less likely to engage with Republican content. However, we found no statistically significant relationship between platform bias and the post-engagement likelihood for the Democratic post on Twitter/X. Our findings indicate on Twitter/X, perceived bias discourages user engagement, especially with Republican posts, highlighting a potential interaction between platform perceptions and political content. This aligns with responses to our open-ended questions. For instance, one participant stated, “*A platform’s stance can be seen by the design choices made through their misinformation warnings.”* Another participant cited explicit concerns with Twitter/X since Elon Musk purchased the company and the impact of his personal bias.

***RQ4: How do social media use and political ideology impact users’ trust in each platform and warning, perceived accuracy of each platform’s warning, and likelihood to engage with the content (like, share, comment)?***

**Platform Frequency.** Using a ranked correlation test, we found a statistically significant relationship between platform frequency and the average likelihood of engaging with the misinformation for Threads (*rs = 0.6126, p = 0.0053).* Additionally, we performed chi-squared tests to evaluate the relationship between platform frequency and engagement with each post on Threads. We found a strong statistically significant relationship between platform frequency and engagement likelihood with the Democratic post on Threads (*Cramér’s V = 0.618, p = 0.0246)*. We also found a statistically significant relationship between the platform frequency and the likelihood of engaging with the Republican misinformation post on Threads (*Cramér’s V = 0.588, p = 0.0412).* We found no other statistically significant relationships between platform frequency and warning ratings across each metric.

**Political Ideology.** We found there is a strong statistically significant relationship between ideology and the effectiveness rating of the Democratic post on Instagram (*p = 0.0328, Pearsons r = 0.491*). However, we found no other statistically significant relationships between political ideology and the warning rating metrics. We also conducted Spearman’s rho tests for correlation between each post and metric for each platform.

**Instagram.** We found engagement with the Republican misinformation post is strongly positively correlated with engagement with the Democratic post for Instagram *(p= 0.0078, rs = 0.590)*. Engagement with Republican and Democratic posts is strongly correlated, revealing that users are unlikely to engage with flagged content on Instagram regardless of ideological alignment. We found no statistically significant relationships between the ratings for trust, accuracy, clarity, or effectiveness for the Democratic and Republican posts on Instagram.

**Facebook.** We found warning trust for the Republican Facebook post is positively correlated with warning trust for the Democratic post (*p = 0.0358, rs = 0.484)*. We also found that the warning accuracy rating for the Democratic post is strongly positively correlated with the Republican post warning accuracy rating (*p = 0.0209, rs = 0.525)*. Since trust and accuracy ratings for Republican and Democratic posts are positively correlated, it suggests that users assess these metrics consistently across political alignments on Facebook. However, we found no statistically significant relationships between the ratings for clarity, engagement, or effectiveness for the Democratic and Republican posts on Facebook.

**Threads.** We found engagement with the Republican misinformation post is strongly positively correlated with engagement with the Democratic post for Threads (*p = 0.0001, rs = 0.771).* Engagement with Republican and Democratic posts is strongly correlated, revealing that users are extremely unlikely to engage with flagged content on Threads regardless of ideological alignment. One explanation for this is the significant proportion of participants who never use Threads (78.9%). This supports our finding mentioned above between platform frequency and engagement with Threads. Additionally, we found warning effectiveness for the Democratic post is strongly positively correlated with the warning effectiveness for the Republican post (*p = 0.0106, rs = 0.571)*. However, we found no statistically significant relationships between the ratings for trust, accuracy, or clarity for the Democratic and Republican posts on Threads.

**Twitter/X.** We found that the warning clarity of the Democratic misinformation post was positively correlated with the warning clarity of the Republican misinformation post on Twitter/X (*p = 0.0487, rs = 0.458)*. However, we found no statistically significant relationships between the ratings for trust, accuracy, effectiveness, or engagement for the Democratic and Republican posts on Threads.

# 4 Discussion

This study sought to investigate the impact of platform bias perceptions, social media use frequency, and political ideology on user evaluations of misinformation warnings across four major platforms (Twitter/X, Facebook, Instagram, and Threads). The findings contribute to the broader discourse on misinformation management by shedding light on how these factors shape trust in misinformation warnings, perceived accuracy, effectiveness, engagement likelihood, and clarity. Our findings align with existing literature suggesting that perceptions of platform bias can significantly affect trust and engagement with misinformation warnings. Notably, Twitter/X was perceived as the most biased platform, with higher perceptions of bias linked to reduced warning trust, accuracy, and engagement likelihood. This corroborates prior studies highlighting the role of platform reputation and user perceptions in shaping responses to content moderation measures. The statistically significant relationships observed for Twitter/X suggest that platforms perceived as biased face greater challenges in gaining user trust. On Twitter/X, perceived bias underscores the platform's reputational challenges under its new ownership and reveals how bias perceptions and platform design choices may interact with user behavior.

Moreover, our findings related to platform use frequency and political ideology offer new insights. Threads, a relatively new platform, demonstrated strong correlations between usage frequency and engagement likelihood, highlighting the influence of platform familiarity on behavior. Similarly, ideological alignment was shown to affect evaluations of misinformation warnings, though inconsistently across platforms. This suggests that while ideology matters, other factors, such as platform reputation, post content, and warning design, are equally critical in shaping user responses. Our study advances the current understanding of how platform bias, user behavior, and political ideology intersect in evaluating misinformation warnings. By directly comparing user perceptions across multiple platforms, we highlight factors influencing trust, clarity, effectiveness, accuracy, and engagement. The findings also provide actionable insights for platforms aiming to design and implement effective misinformation warnings, emphasizing the need to address user bias perceptions and foster trust through transparent and credible sources.

**4.1 Limitations**

Despite its contributions, this study has several limitations. First, the use of a sample of 19 college students as participants may limit the generalizability of the findings. College students are not fully representative of the broader population, particularly in terms of social media use habits and political diversity. The sample skewed more liberal, which may have influenced the evaluation of platforms and warnings, particularly for politically charged content. Including a larger sample of conservative participants in future studies could provide a more balanced perspective. Additionally, the participant pool also reflected significant disparities in platform usage frequency, with a majority of participants regularly using Instagram and never using Threads, which could bias perceptions and the outcome of our analysis. Since Threads had a disproportionately low usage rate among participants, it may limit the quality of the findings related to this platform. Finally, while we tried to simulate the natural platform environment through static images, the study design involved artificial scenarios rather than organic user experiences on live platforms, which may not fully capture real-world behaviors.

**4.2 Future Work**

Future research should replicate this study with refined methodologies to ensure greater consistency between the misinformation posts used. Differences in content, tone, or visual design across posts may inadvertently influence participants' evaluations of the warnings or their engagement with the content. By creating highly standardized posts that are identical in structure, formatting, and design—varying only by political affiliation or platform-specific features—future studies can minimize the potential confounding effects of content differences. Future research should also recruit a more diverse participant pool, including more participants and a broader range of ages, political ideologies, and social media habits. Additionally, experimental designs incorporating real-time user interactions with misinformation warnings in naturalistic settings could yield more ecologically valid insights. Future research can expand the scope of the study to include other popular platforms such as TikTok, YouTube, and Reddit. These platforms differ significantly in user demographics, content formats, and community norms, which may influence the effectiveness and perception of warnings. For instance, TikTok’s younger audience and short-form video content present unique challenges and opportunities for misinformation warnings and including a wider scope of platforms can provide a more comprehensive understanding of misinformation warning designs.

# 5 Conclusion

This study highlights the multifaceted challenges and opportunities in designing effective misinformation warnings on social media platforms. Beyond merely assessing user perceptions of trust, accuracy, clarity, effectiveness, and engagement, the findings underscore the intricate relationship between platform reputation, user bias, and design features. While Instagram’s success in clarity and effectiveness highlights the value of user-centric, visually distinct warnings, the struggles of Threads and Twitter/X to garner trust reveal deeper issues tied to platform bias and user skepticism. Importantly, this research highlights that the effectiveness of misinformation warnings extends beyond their immediate design. A collection of user experiences, political ideologies, and platform reputations collectively shape how warnings are perceived and acted upon. These insights suggest that platforms must move toward more transparent and impartial practices, not just in warning design but also in their overall approach to content moderation. Effective interventions require addressing both the technical elements of warning design and the social factors, such as trust and ideological alignment, that influence user behavior. By integrating these dimensions, platforms can not only mitigate misinformation more effectively but also rebuild trust and credibility in an era of increasing digital skepticism. This work advances the field by integrating quantitative analyses with user-centered perspectives, offering a foundation for future research to enhance generalizability, address ideological diversity, and develop personalized, transparent warning systems. Ultimately, the findings emphasize that platforms need to carefully design misinformation warnings in a manner that is not only clear and informative but also fosters trust and credibility. Understanding how user perceptions are shaped by the characteristics of the platform, ideological alignment, and how the warnings are presented will help social media platforms better address the challenges brought about by misinformation and work toward improving the effectiveness of their content moderation efforts.

# References

1. Pablo Barberá. 2018. Explaining the Spread of Misinformation on Social Media: Evidence from the 2016 U.S. Presidential Election. Retrieved from <https://api.semanticscholar.org/CorpusID:221335315>
2. Dipto Barman and Owen Conlan. 2023. Does Explanation Matter? An Exploratory Study on the Effects of Covid–19 Misinformation Warning Flags on Social Media. 2023 10th International Conference on Behavioural and Social Computing (BESC) (2023), 1–7. Retrieved from <https://api.semanticscholar.org/CorpusID:263136478>
3. Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Valensise, Emanuele Brugnoli, Ana Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. Scientific reports 10, (October 2020). DOI:https://doi.org/10.1038/s41598-020-73510-5
4. Melanie Freeze, Marybeth Baumgartner, Peter Bruno, Jacob R. Gunderson, Joshua Olin, Morgan Quinn Ross, and Justine Szafran. 2020. Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect. Political Behavior (2020), 1–33. Retrieved from <https://api.semanticscholar.org/CorpusID:214269774>
5. R. Kelly Garrett and Shannon Poulsen. 2019. Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings. Journal of Computer-Mediated Communication 24, (September 2019), 240–258. DOI:https://doi.org/10.1093/jcmc/zmz012
6. Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. Science 363, (January 2019), 374–378. DOI:https://doi.org/10.1126/science.aau2706
7. Chen Guo, Nan Zheng, and Chen Guo. 2023. Seeing is Not Believing: A Nuanced View of Misinformation Warning Efficacy on Video-Sharing Social Media Platforms. Proceedings of the ACM on Human-Computer Interaction 7, (2023), 1–35. DOI:https://doi.org/10.1145/3610085
8. Ben Kaiser, Jerry W. Wei, Elena Lucherini, Kevin Lee, J. N. Matias, and Jonathan R. Mayer. 2020. Adapting Security Warnings to Counter Misinformation. arXiv: Human-Computer Interaction (2020).
9. Nandita Krishnan, Jiayan Gu, Rebekah Tromble, and Lorien C. Abroms. 2021. Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review* 2, 6, (December 15, 2021). https://doi.org/10.37016/mr-2020-85
10. Kevin T. Leicht, Joseph Yun, Brant Houston, Loretta Auvil, and Eamon Bracht. 2022. The Presentation of Self in Virtual Life: Disinformation Warnings and the Spread of Misinformation Regarding COVID-19. RSF 8, 8 (December 2022), 52–68. DOI:https://doi.org/10.7758/RSF.2022.8.8.03.
11. Drew B. Margolin, Anikó Hannák, and Ingmar Weber. 2018. Political Fact-Checking on Twitter: When Do Corrections Have an Effect? Political Communication 35, (2018), 196–219. Retrieved from <https://api.semanticscholar.org/CorpusID:53965636>
12. Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio Arechar, and Dean Eckles. 2021. Shifting attention to accuracy can reduce misinformation online. Nature 592, (April 2021), 1–6. DOI:https://doi.org/10.1038/s41586-021-03344-2
13. Emily Saltz, Claire Leibowicz, and Claire Wardle. 2020. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (2020). Retrieved from <https://api.semanticscholar.org/CorpusID:227162294>
14. Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. 2022. Misinformation warnings: Twitter’s soft moderation effects on COVID-19 vaccine belief echoes. Comput. Secur. 114, C (Mar 2022). <https://doi.org/10.1016/j.cose.2021.102577>
15. Filipo Sharevski, Amy Devine, Peter Jachim, and Emma Pieroni. 2022. Meaningful Context, a Red Flag, or Both? Preferences for Enhanced Misinformation Warnings Among US Twitter Users. Proceedings of the 2022 European Symposium on Usable Security (2022). Retrieved from https://api.semanticscholar.org/CorpusID:252123359

# Appendices

# A Survey Questions

**Presurvey**

A screenshot of a survey

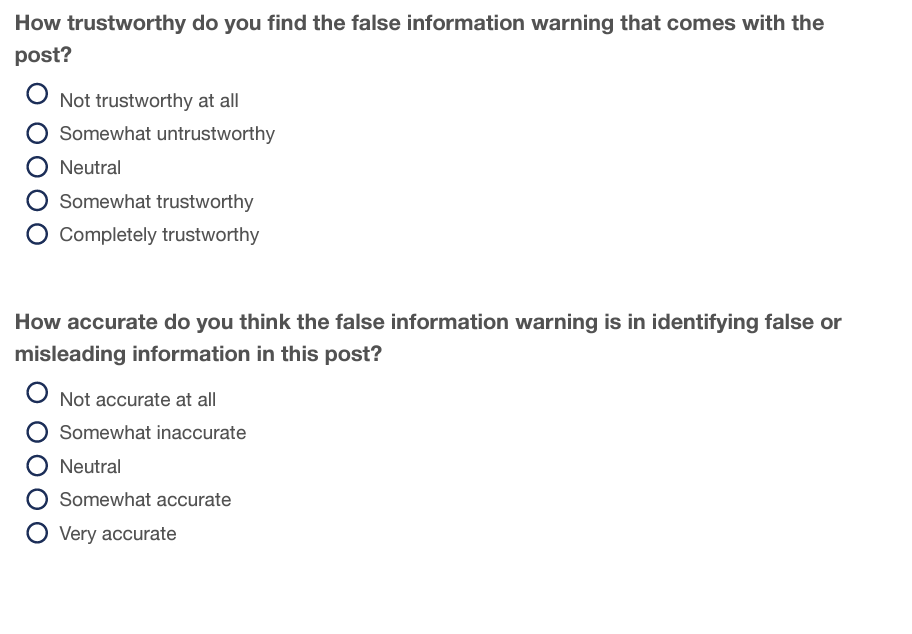
Description automatically generatedA questionnaire with a few circles

Description automatically generated with medium confidence

**Stimuli Evaluation**

The following Likert-scale questions accompanied each post:

A screenshot of a questionnaire

Description automatically generated

**The following posts were used for evaluation in the survey:**

|  |  |  |
| --- | --- | --- |
|  | **Republican** | **Democratic** |
| **Instagram** | Screens screenshots of a social media account  Description automatically generated | Screens screenshots of a social media account  Description automatically generated |
| **Facebook** | Screens screenshot of a social media post  Description automatically generated | A screenshot of a social media post  Description automatically generated |
| **Threads** | Screenshots of a social media post  Description automatically generated | A screenshot of a social media account  Description automatically generated |
| **Twitter/X** | A screenshot of a phone  Description automatically generated | A screenshot of a person's ear  Description automatically generated |

**Post Survey Questions:**

A screenshot of a survey

Description automatically generated

A screenshot of a survey

Description automatically generatedA screenshot of a form

Description automatically generated