The Journal of Social Media in Society
Fall 2024, Vol. 13, No. 2, Page 131-150
thejsms.org

Examining the Effectiveness of Social Media Warning Labels: The Role of Worldview Inconsistency and Psychological Reactance

Bingbing Zhang
School of Journalism and Mass Communication, University of Iowa, Iowa City, IA
bingbing-zhang@uiowa.edu, 319-335-1760, Twitter @Blingblingzhang

Social media platforms frequently employ warning labels to identify potentially misleading information, yet research has yielded inconsistent findings regarding the efficacy of these labels. This study aimed to assess the effectiveness of two prominent types of Twitter (which is rebranded as X) warning labels in the specific context of identifying misinformation related to immigration. The results revealed that the warning labels had no direct impact on diminishing perceived credibility and the intention to share the misinformation post. However, the findings did highlight a moderating effect of worldview inconsistency. The research findings provide valuable insights for effective designs for corrective messaging on social media platforms, emphasizing the role of worldview inconsistency and reactance on processing various warning label types in discrediting misinformation posts.

Keywords: misinformation, warning label, worldview inconsistency, psychological reactance, perceived credibility, intention to share

---

Given the prevalence and persistence of misinformation posing as "factual news" on social media, both industry and academia are urging the implementation of more effective corrective measures to address this problem (Allcott et al., 2019). Scholars have proposed two primary corrective approaches: debunking misinformation through corrections from third-party expert organizations or ordinary social media users (Bode & Vraga, 2015; Lewandowsky et al., 2017), and preventive actions such as enhancing media literacy (Vraga, Bode, & Tully, 2020) or inoculating users with games that teach strategies for recognizing fake news distribution (Maertens et al., 2021). One common debunking strategy adopted by social media platforms to combat misinformation is attaching labels to posts containing suspicious information, thereby flagging potential misinformation (Mena, 2020; Oeldorf-

Hirsch et al., 2020). This approach, initially seen on news websites using rating scales or flags from third-party fact-checking organizations (Amazeen et al., 2018; Mena, 2020), was later adopted by social media platforms.

For instance, Facebook introduced "disputed news" labels beneath posts to indicate fact-checking results (Oeldorf-Hirsch et al., 2020). Following in Facebook's footsteps, Twitter (which is rebranded as X currently) implemented similar flagging actions in 2020. Given that the data collected for this study predates Twitter's rebranding, the study will use X's previous brand name. Unlike Facebook, Twitter provides not only a disputed label but also a spectrum of labels. Twitter offers both a flagging false warning label that directly informs users the information in the post is false and a neutral label that doesn't confirm the information's falseness but warns users about potential suspicious content (Roth & Pickles, 2020). Twitter has expanded the types of warning labels, incorporating soft moderation methods to debunk misinformation through neutral and general warning labels (Sharevski et al., 2021).

Research on the persuasiveness of corrections has produced inconsistent results. While a meta-analysis found that corrections had a moderate effect on correcting belief in misinformation (Walter & Murphy, 2018), other studies revealed a backfire effect to corrections (Nyhan & Reifler, 2010). Motivated reasoning theories have been used to explain correction ineffectiveness, asserting that prior worldviews impact correction effects because individuals may resist information inconsistent with their attitudes (Taber & Lodge, 2006). Some studies argue that analytic thinking is crucial for processing information in the face of suspicious content, challenging the notion of motivated reasoning (Pennycook & Rand, 2019). Additionally, research debunked the backfire effect by testing correction effects across various issues, finding that individuals pay attention to factual information even when it contradicts their prior worldviews (Wood & Porter, 2019).

Given the mixed results regarding the impact of worldview (in)consistency on correction effects, this study aims to test the interactions of worldview consistency and different types of warning labels adopted by Twitter in combating misinformation. Specifically, it compares the effectiveness of two major types of warning labels—soft moderation measures (neutral warning labels) and flagged false warning labels—examining how individuals process them. This study contributes to the current literature

by addressing the lack of research comparing the effectiveness of different warning labels, especially the new soft moderation flagging label adopted by Twitter. Moreover, while prior research has started to examine resistance to corrective measures such as flagging labels (Garrett & Poulsen, 2019), the psychological mechanisms behind it have not been fully explored. This study investigates how worldview inconsistency contributes to resistance to flagging labels and how reactance impacts perceived credibility and the intention to share misinformation posts. Importantly, the study empirically explores the interaction effect of worldview consistency and warning labels on perceived credibility and the intention to share misinformation posts on Twitter.

LITERATURE REVIEW

Misinformation & Correction on Social Media

The term "misinformation" refers to false or misleading information (Berinsky, 2015). The identification of misinformation often hinges on the correction process, where information initially presented as true is later retracted or corrected (Lewandowsky et al., 2012; Ecker et al., 2015). This study adopts a temporal perspective, treating misinformation as information initially presented as factual but subsequently debunked. Social media platforms have implemented corrective measures to combat misinformation. For instance, Facebook distributes related stories debunking misinformation posts, a strategy shown to be effective in reducing misperceptions (Bode & Vrage, 2015). However, correcting misinformation on social media faces challenges due to a lack of professional gatekeeping in content creation and distribution (Bessi et al., 2015). Social media users, acting as cognitive misers, may opt to quickly scan posts rather than delving into each one for more details, even if additional information or links are provided (Moravec et al., 2018). Warning labels can potentially capture the attention of social media users during their quick scans, prompting them to reconsider the information they encounter.

Research testing the effectiveness of flagging or warning labels on social media, particularly on Facebook and Twitter, has produced mixed results. While Mena (2020) found that flagging labels reduce the sharing of false information by diminishing the perceived credibility of misinformation posts, Oeldorf-Hirsch et al. (2020) reported that third-party disputed fact-checking labels on Facebook were ineffective. Garrett and

Poulsen (2019) investigated various types of flags on Facebook, revealing that fact-checker flags and peer-generated flags had no impact on reducing misbeliefs and sharing intentions related to misinformation posts. Conversely, self-identified humor flags were effective in diminishing prior misbeliefs (Garrett & Poulsen, 2019). Lee et al. (2023) found that fact-checking labels on Twitter effectively reduced belief in vaccination misinformation but did not increase vaccination intention. Similarly, Lees et al. (2022) found that Twitter's disputed labels reduced the intention to share fake news among Democrats and Independents. However, other scholars have found that Twitter labels have limited effectiveness in mitigating the spread of misinformation (e.g., Papakyriakopoulos & Goodman, 2022; Sanderson et al., 2022; Sharevski et al., 2022).

Despite prior research on the impact of warning labels on combating misinformation on social media, minimal attention has been given to the effectiveness of different types of warning labels. For example, Clayton et al. (2020) found that, compared to specific warning labels indicating that a message is rated false or disputed, general warnings broadly stating that media messages might contain inaccurate information were more effective in reducing the perceived accuracy of false headlines. In contrast, Freeze et al. (2020) reported that general warning labels were not as effective as specific warning labels in helping users reject misleading information. They found that specific warning labels directly stating that the information is disputed were more effective than soft moderation types of warnings (Freeze et al., 2020). Nassetta and Gross (2020) examined warning labels that flagged state media and found that such labels can mitigate belief in misinformation. However, they also found that warning labels worked only when users noticed their existence.

In summary, past research has provided valuable theoretical foundations for the current study to examine the effectiveness of different types of warning labels. Building on the outlined research, this study aims to investigate the effects of both general warning labels, encouraging users to learn more about the topics depicted in the post, and specific warning labels, explicitly stating that the information contained in the post is false. The following hypotheses are thus proposed:

H1: Individuals will perceive message with (a) neutral warning label and (b) specific warning label (rated false) more credible than posts without warning label.

H2: Individuals will be more likely to share misinformation post without warning label than post with (a) neutral warning label and (b) specific warning label (rated false).

Worldview Inconsistency & Psychological Reactance to Warning Labels

Prior worldview or attitudes are frequently considered crucial factors influencing how individuals process correction and fact-checking information. Walter and Tukachinsky's (2020) meta-analysis revealed that corrective messages are less effective when they do not align with the audience's prior worldview. Resistance to corrective messages is often driven by motivated reasoning (Taber & Lodge, 2016), where individuals avoid information that contradicts their existing beliefs (Berinsky, 2015). According to Taber and Lodge (2006), people engage in attitude congruency bias and confirmation bias when exposed to information inconsistent with their worldview. Specifically, attitude congruence bias means individuals are more likely to perceive arguments supporting their prior worldview in misinformation posts as stronger than arguments in corrective messages opposing their prior attitude. Confirmation bias means people are more likely to avoid information in corrections incongruent with their prior attitude. Building on this research, the current study utilizes individuals' prior worldview to determine whether warning labels and corrections align with their worldview.

Worldview inconsistency can also make individuals perceive a threat to their freedom when consuming media messages (Kahan, 2010). For example, during the pandemic, people exhibited reactance to masking mandates and vaccine mandates based on their political beliefs and prior worldviews (Taylor & Asmundson, 2021). Psychological reactance theory posits that overt persuasive messages are perceived as a threat to one's freedom, leading individuals to reject such messages (Brehm, 1966; Brehm & Brehm, 1981). Corrective messages may induce psychological reactance easily due to their persuasive intent. Past research has found that correction serving as an overt persuasive appeal may lead to a backfire effect when the message contradicts individuals' prior attitudes (Lewandowsky et al., 2012). Consequently, attaching warning labels to social media posts may induce psychological reactance when it contradicts individuals' prior worldviews.

Reactance to warning labels might lead to message derogation. Lewandowsky et al. (2012) argued that debunking is prone to be ineffective or even backfire, leading people to believe in falsehoods more. When individuals feel their worldview is challenged, they might engage in message derogation, actively resisting warning labels and refusing to change their attitudes (Fransen, Smit, & Verlegh, 2015). Moreover, reactance to warning labels might result in increased attention to the false information contained in social media posts. Garrett and Poulsen (2019) found that warning flags from peers induced a higher level of reactance than humor flags when people held stronger prior beliefs in false information. However, Wood and Porter (2019) debunked the backfire effect to correction, showing that people pay attention to factual information even when it contradicts their prior beliefs across 52 issues.

Taken together, competing hypotheses emerge. On one hand, warning labels correcting misinformation aim to reduce perceived credibility and the intention to share misinformation posts. On the other hand, the presence of warning labels might induce reactance to the labels, where individuals' worldviews are offended, leading to resistance that could hinder the process of reducing trust in misinformation posts. Reactance, as a form of arousal induced when individuals perceive their freedom is threatened, is the focus of the current study, emphasizing cognitive and motivated counterarguing rather than the feeling of anger towards warning labels. Therefore, the following hypotheses are proposed:

H3: Perceived credibility will mediate the relationship between warning labels exposure and intention to share the misinformation post under the conditional effects of worldview inconsistency.

H4: Worldview inconsistency will moderate the effect of (a) neutral warning label and (b) specific warning label (rated false) on perceived credibility of misinformation post.

H5: Reactance to warning label will increase perceived credibility of misinformation post which will lead to the increase of intention to share misinformation post.

METHODS

This study conducted an online between-subjects experiment using Amazon MTurk. IRB approval was secured before initiating data collection in March 2022. Amazon MTurk facilitated the recruitment of a substantial panel of American adults through verified sources, establishing itself as a reliable platform for data collection in the communication field (Sheehan, 2018). To ensure data quality, participants were required to have a percentage of approved HITs around 95%. Each participant received compensation for completing a 10-minute online survey questionnaire.

Participants

A priori power analysis using G*Power indicated that a minimum sample size of 207 is needed to detect fixed effects in one-way ANOVA (effect size $f = 0.25$, $\alpha = .05$, power = .90, and number of groups = 3). An original sample of 237 participants were recruited. A manipulation check question asked participants to indicate the topic of the Twitter post they have read about, eventually 8 people did not pass the manipulation check. Thus, the final sample was 229 people. Each condition group has around 76 participants. Among the samples, 76.4% of them were male while 23.6% of them were female. 57% of them were Caucasians, followed by 26.1% Asian, 9.7% Black/African American, and 4.2% Hispanic/Latino. The average age of our participants was 33.08 years old ($SD = 8.41$). The median education level of the participants was 4-year college degree. The average income level of the participants was between $40,000 to under $60,000.

Design & Procedure

After reporting their attitudes toward immigrants, participants were exposed to stimuli presented on a simulated Twitter page for external validity. Initially, participants were randomly exposed to one of three stimuli: (1) a misinformation post with a false warning label; (2) a misinformation post with a neutral warning label; and (3) a misinformation post with no warning label. Participants were required to view the stimuli for a minimum of 10 seconds before advancing to the next page. The misinformation post focused on the false claim that immigrants are responsible for job theft from Americans, leading to a high unemployment rate. Subsequently, participants reported their attitudes toward the warning labels, perceived credibility of the misinformation post, intention to share the misinformation post, and demographic information.

Stimuli

Three different messages were designed specifically for this study. The misinformation post claimed that immigrants are stealing 20% of the jobs in America and the unemployment rate is increasing every year for that. This claim is false because fact-checking organizations have already debunked this claim (Novak, 2010). Two different types of warning label were designed to attach with the misinformation Twitter post. Flagging false label comes with the text "This claim is flagged false, and then information in this Tweet is false." Disputed label comes with the text "Learn about U.S. immigration and relevant government policy" without mentioning whether this post is false information or not (see Table 1 for stimuli details).

Table 1

Warning Label Stimuli

		
Misinformation Post Without Label (control)	Flagging False Warning Label	Disputed Warning Label

Measures

Worldview inconsistency. This measure was created by recoding the measure of prior attitude towards immigrants adapted from Hameleers and van der Meer (2020). A 10-item scale was used to examine individuals' attitudes toward immigrants. Participants were asked to indicate their agreement (1 = *strongly disagree* to 7 = *strongly agree*) to the following statements such as (a) Immigrants coming to the U.S are responsible for major increases in crime rate; (b) Immigrants are increasingly involved in crimes targeted at U.S. citizens; (c) Increases in crime rates are caused by the government's failing immigration policies; (d) Our country border controls are too weak; (e) Our country law enforcement is too weak. All items were recoded and combined into an index (M = 4.94; *SD*

= 1.20; *a* = .90). A higher score indicates that participants have more negative attitudes towards immigration which means that their worldview is more inconsistent with the warning labels and correction.

Psychological reactance. Adapted from Moyer-Guse & Nabi (2009), participants were asked to indicate their attitude towards the warning labels attached to the Twitter post (1 = *strongly disagree* to 7 = *strongly agree*): (a) Twitter warning labels are meant to keep me from reading or sharing important news stories; (b) Flagging news stories is just a way to pressure people to think a certain way; and (c) Twitter is using warning labels to force its opinions on me. Psychological reactance is only measured on the conditions that people who are exposed to warning labels. All items were combined into an index (*M* = 4.91; *SD* = 1.44; *a* = .75).

Perceived source credibility. Adapted from Sundar et al. (2017), a seven-item scale was used to measure perceived credibility of the misinformation post. Participants were asked how well (1 = *describes very poorly* to 7 = *describes very well*) the following adjectives describe the previous Tweet they just read: (a) Accurate; (b) Authentic; (c) Believable; (d) Biased; (e) Fair; (f) Objective; and (g) Sensationalistic. Items d and item g were reverse coded. The reverse coded items and the rest of the items were combined to an index (*M* = 4.43; *SD* = 0.85; *a* = .84).

Intention to share. Adapted from Garrett and Poulsen (2019), two items were used to assess individuals' intention to share the misinformation post. Participants were asked to rate their agreement (1 = *strongly disagree* to 7 = *strongly agree*) to the following statement: I would 'like' the Tweet if it showed up on my "feed" on Twitter. Then participants were also asked to answer the following question (1 = *Extremely unlikely* to 7 = *Extremely likely*): How likely are you to share the Tweet with others on social media? These two items were combined to an index (*M* = 3.64; *SD* = 1.03; *r* = .87). The zero-order correlations between variables of interest and demographics are presented in Table 2.

Table 2
Pearson's Correlations Between Variables of Control and Interest

	1	2	3	4	5	6	7	8	9
1.Worldview Inconsistency	--								
2.Label Reactance	.598**	--							
3.Perceived Credibility	.529**	.708**	--						
4. Intention to Share	.624**	.648**	.692**	--					
5.Sex(Male)	-.141	-.134	-.079	-.14	--				
6.Age	.158*	.045	.210**	.162*	-.124	--			
7.Race (White)	-.035	.01	.049	-.025	.006	.162*	--		
8.Education	.246**	.240*	.207**	.265**	.14	-.028	-.041	--	
9. Income	.011	.009	.046	.028	.011	.083	.137	.161*	--

*Notes. N = 229. Cell entries are two-tailed correlation coefficients. *p < .05; **p < .01; ***p <.001*

RESULTS

H1 proposed that individuals would perceive message with (a) neutral warning label and (b) specific warning label (rated false) more credible than posts without warning label. A one-way ANOVA was conducted to test whether perceived credibility of the misinformation post vary as a function of warning labels. This analysis revealed no main effect of warning label on perceived credibility, $F(2, 227) = 0.55$, $p = .58$, partial $\eta^2 = .007$. This indicated that whether exposure to warning labels or not did not have significant difference in perceived credibility of the misinformation post. Therefore, H1 was not supported.

H2 proposed that individuals would be more likely to share misinformation post without warning label than with (a) neutral warning label and (b) specific warning label (rated false). One-way ANOVA was also conducted to test whether intention to share misinformation post vary as a function of warning labels. This analysis revealed no main effect of warning labels on intention to share was found, $F(2, 227) = 2.28$, $p = .17$. Therefore, H2 was not supported.

H3 proposed that perceived credibility would mediate the relationship between warning labels exposure and intention to share the misinformation post under the conditional effects of worldview consistency. A moderated-mediation model specifically PROCESS Model 7 (Hayes, 2018) was used to estimate the degree to which worldview consistency moderate the effects of flagged false and neutral labels on perceived credibility and the mediating role of perceived credibility on intention to share the misinformation

post. First, exposure to flagged false and neutral labels or no label was entered as a multicategorical independent variable in the model. Also, the condition without warning label was set to be the reference group in the PROCESS model. Perceived credibility of the misinformation post was entered as mediator and intention to share was entered as dependent variable while worldview inconsistency was entered as moderator in the model. Age, gender, race (White), education, and income were entered as covariates.

Results showed that the indirect effect was contingent on the level of worldview consistency for paths from neutral warning label (moderated mediation index: .24, *se* = .13, 95% *CI* = [.012, .511]) to intention to share through perceived credibility but not for flagging false warning label (moderated mediation index: .18, *se* = .13, 95% *CI* = [-.055, .448]). More specifically, results (see Figure 2) showed that exposure to neutral warning label significantly negatively predicted perceived credibility of the post (*b* = -1.53, *se* = .62, *t* = -2.47, *p* = .02). Importantly, perceived credibility significantly positively predicted intention to share the misinformation post (*b* = .79, *se* = .07, *t* = 11.06, *p* < .001, 95% CI [.645, .925]). Results also showed that there is no significant effect of exposure to neutral label (*b* = -.24, *se* = .07, *p* = .10) on intention to share the post when perceived credibility served as a mediator. This indicated that perceived credibility served as a mediator for the relationship between exposure to neutral warning labels and intention to share the misinformation post under the conditional effect of worldview consistency. Therefore, H3a was not supported but H3b was supported.

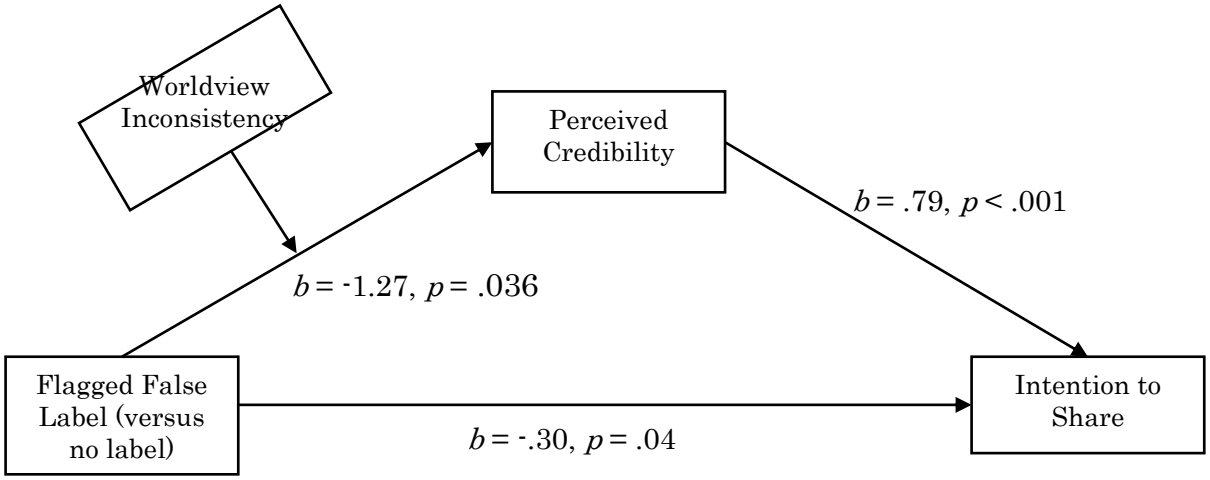

Figure 1. Moderated-mediation Model for Flagging False Label. Sample size = 229. Dotted line means the path is not significant while the straight line means the path is significant at $p < .05$. Path cells are unstandardized coefficients. Bootstrap samples for CI: 5000 simulations. The index of the *moderated mediation* is .183, *se* = .13, 95% *CI* = [-.055, .448].

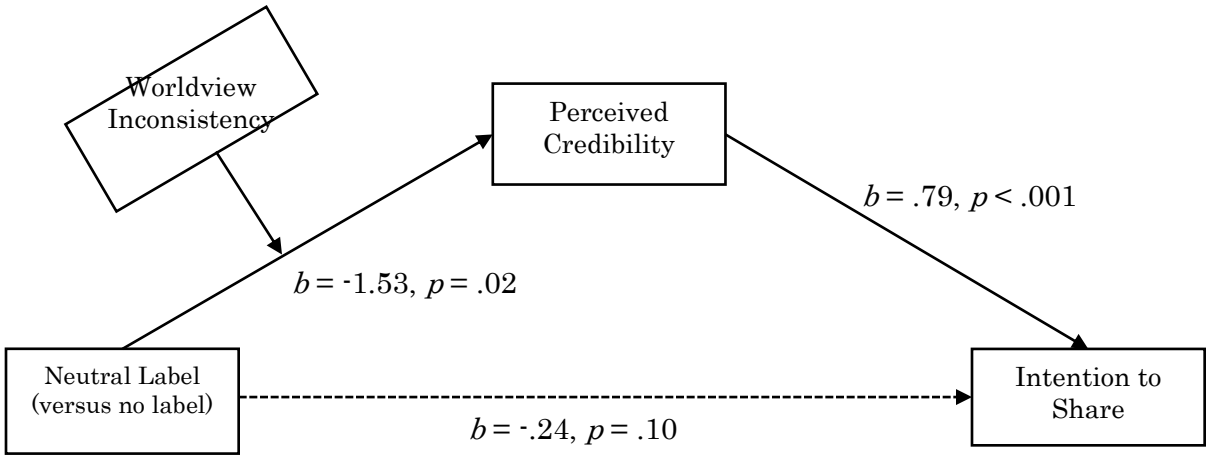

Figure 2. *Moderated-mediation Model for Neutral Label*. Sample size = 229. Dotted line means the path is not significant while the straight line means the path is significant $p < .05$. Path cells are unstandardized coefficients. Bootstrap samples for CI: 5000 simulations. The index of the *moderated mediation* is .24, $se = .13$, 95% $CI = [.012, .511]$.

H4 proposed that worldview consistency will moderate the effect of (a) neutral warning label and (b) specific warning label (rated false) on perceived credibility of the misinformation post. Results showed that the interaction effects of exposure to both types of warning labels and worldview inconsistency on perceived credibility were statistically significant. More specifically, compared to no label condition, those with higher level of worldview inconsistency who were exposed flagging false label perceived misinformation post to be more credible ($b = .23$, $se = .12$, $t = 1.98$, $p = .04$, 95% CI [.0001, .466]). Similarly, those with lower level of worldview inconsistency who were exposed to neutral label perceived misinformation post to be less credible compared to those who were not exposed to label ($b = .31$, $se = .12$, $t = 2.55$, $p = .012$, 95% CI [.070, .551]). Therefore, H4 was supported. The conditional effect of the focal predictor was visualized with the aid of PROCESS model syntax output (see Figure 3).

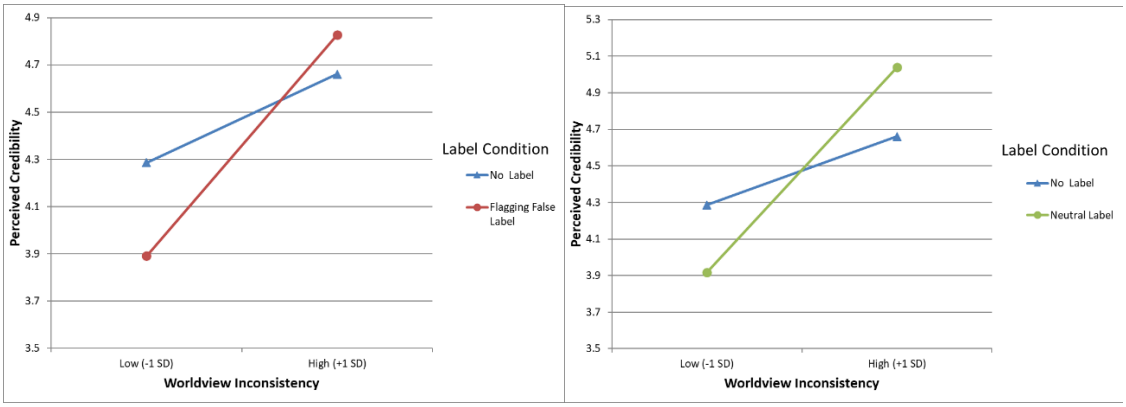

Figure 3. Interaction Between Worldview Inconsistency and Label Conditions on Perceived Credibility. Values for worldview inconsistency are given at the -1SD, mean, and +1SD. Higher values represent more negative value toward immigration.

H5 proposed that reactance to warning label will increase perceived credibility of misinformation post which will lead to increase intention to share misinformation post. A moderated-mediation model specifically PROCESS Model 4 (Hayes, 2018) was used to examine how label reactance impacts intention to share misinformation post indirectly through perceived credibility. Perceived credibility of the misinformation post was entered as mediator and intention to share was entered as dependent variable while label reactance was entered as independent variable in the model. Age, gender, race (White), education, and income were entered as covariates.

Results indicated a significant indirect effect of perceived credibility on relationship between label reactance and intention to share misinformation post ($b = .31$, $se = .06$, 95% $CI = [.194, .431]$). More specifically, reactance to warning labels significantly positively predicted perceived credibility of the misinformation post ($b = .43$, $se = .04$, $t = 10.04$, $p < .001$). Perceived credibility significantly positively predicted intention to share misinformation post ($b = .73$, $se = .11$, $t = 6.80$, $p < .001$, 95% CI [.516, .940]). Results also showed that there is significant effect of reactance to warning labels ($b = .14$, $se = .07$, $t = 2.17$, $p = .03$, 95% CI [.013, .274]) on intention to share the post when perceived credibility served as a mediator. This indicated that perceived credibility served as a mediator for the relationship between exposure to warning labels and intention to share the misinformation post. Therefore, H5 was supported.

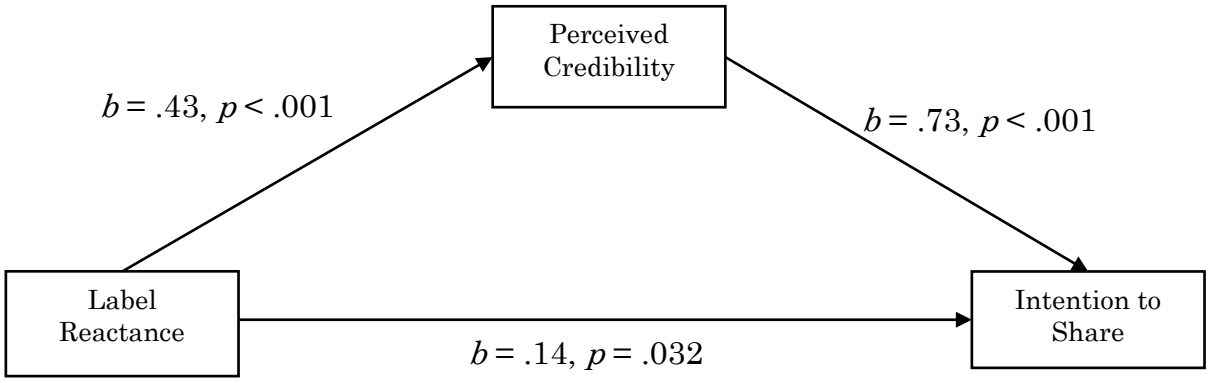

Figure 4. Mediation Model for Relationship Between Label Reactance & Intention to Share. Sample size = 229. Dotted line means the path is not significant while the straight line means the path is significant $p < .05$. Path cells are unstandardized coefficients. Bootstrap samples for CI: 5000 simulations. The estimate of indirect effect is $b = .31$, $se = .06$, 95% $CI = [.194, .431]$.

DISCUSSION

The pervasive presence of misinformation on social media has prompted the exploration of debunking strategies in both academia and industry (Amazeen, 2020). This study delves into one prevalent debunking strategy employed by social media platforms – flagging posts with potential misleading information – to discredit misinformation. While the results indicate that both neutral and specific warning labels have limited direct effects in reducing the perceived credibility and intention to share misinformation on social media, they are still effective among groups for whom warning labels are perceived as less inconsistent with their worldview. Reactance to warning labels influences the intention to share misinformation through perceived credibility, underscoring the importance of minimizing individuals' reactance to warning labels. In this study, warning labels did not appear to be effective in directly diminishing the perceived credibility of misinformation posts or reducing the intention to share misinformation, aligning with prior research (Oeldorf-Hirsch et al., 2020). However, the type of warning label may factor into the effectiveness of correction (Clayton et al., 2020; Freeze et al., 2020). This study investigated both neutral and specific warning labels (rated false) but found no main effects on perceived credibility and intention to share misinformation posts. This does not imply that warning labels are entirely ineffective; instead, it suggests that

social media users might not pay adequate attention to warning labels when scanning through posts (Moravec et al., 2018). State media warning labels across platforms like Facebook, YouTube, and Twitter were effective in mitigating the impact of viewing election misinformation from the Russian media channel RT, but only when users noticed the warning labels (Nassetta & Gross, 2020). Therefore, for practical effectiveness, social media platforms should enhance the visibility of warning labels to ensure user engagement.

The study reveals that worldview inconsistency plays a significant role in how individuals perceive warning labels. This study examined the effectiveness of warning labels on Twitter concerning immigration. Immigration is a complex and increasingly polarized political issue. Misinformation in this context is particularly harmful because the complexity of immigration issue makes it easier for false information to spread, potentially undermining democratic processes (Hameleers et al., 2020). Worldview inconsistency occurs when individuals feel that their prior worldview and beliefs are challenged by encountered information (Kahan, 2010). Individuals tend to prefer information aligning with their existing beliefs (Sears & Freedman, 1967; Stroud, 2017; Taber & Lodge, 2016). The results show that worldview inconsistency moderates the effect of both general and specific warning labels on perceived credibility. Among those experiencing lower worldview inconsistency, warning labels are more effective in changing the perceived credibility of misinformation posts. While warning labels may not shift the attitudes of hardcore believers in misinformation, they can influence individuals with moderate attitudes or less extreme ideologies.

Reactance to warning labels increases the perceived credibility of misinformation posts and subsequently heightens the intention to share misinformation. Individuals, when perceiving a threat to their freedom, tend to have the intention to restore freedom through actions (Brehm, 1966; Brehm & Brehm, 1981). Higher psychological reactance, driven by warning labels inconsistent with their worldview, might lead to the boomerang effect where individuals ignore the warning label or believe the falsehood more (Fransen et al., 2015; Lewandowsky et al., 2012). Designing warning labels requires attention to decreasing psychological reactance, potentially achieved through visual cues and less threatening language, as suggested by persuasion literature (Quick et al., 2013).

Furthermore, the moderated mediation analysis found that neutral warning labels can decrease the intention to share misinformation under the conditional effects of worldview consistency. Neutral warning labels worked better when individuals perceived less worldview inconsistency, influencing their intention to share misinformation posts. This mechanism did not occur with specific (rated false) warning labels, indicating that neutral warning labels offer more potential to impact the intention to share misinformation posts compared to false labels. Research suggests that neutral or general warning labels might work better than specific labels in correcting misinformation (Clayton et al., 2020). As a softer moderation measure, neutral warning labels could induce less reactance (Sharevski et al., 2021).

In summary, this study conducted a rigorous examination and found that warning labels were ineffective in directly reducing perceived credibility and curbing the intention to share misinformation posts. Nonetheless, the investigation delved into the nuanced dynamics of how warning labels operate under the conditional effects of worldview inconsistency and the impact of psychological reactance on their effectiveness. Several limitations in the study findings should be acknowledged. Firstly, the primary focus of this study was on whether fact-checking labels could prevent misinformation spread on social media, without exploring their potential to alter behaviors in supporting immigration policies. Future research could delve into whether fact-checking labels influence individuals' actual behaviors related to supporting immigration policies. Secondly, it is crucial to exercise caution, considering that politicized misinformation is typically more challenging to debunk compared to other types, such as health-related misinformation. The meta-analysis by Walter and Murphy (2018) highlighted that corrective messages are most effective against health misinformation but least effective against political misinformation. Given the increasing politicization of the immigration issue, especially its polarization (Hameleers et al., 2020), testing this study in less controversial contexts might yield different results. It is also worth noting that the current study sample has a slightly more negative attitude about immigration, making them more susceptible to immigration misinformation and more reactive to warning labels. Lastly, the study exclusively examined one type of immigration misinformation, namely, the claim that immigrants are stealing jobs from Americans. This singular focus may pose a case-

category confounding issue (Jackson et al., 1994). Future research could diversify stimuli by incorporating different types of immigration misinformation, such as claims that immigrants are contributing to increased crime rates in the U.S., to provide a more comprehensive understanding of the impact of warning labels across various contexts.

Despite these limitations, this study contributes to the existing field of research by examining various factors that could impact the effectiveness of warning labels, specifically worldview inconsistency and psychological reactance. Theoretically, it delves into the impact of warning labels on the intention to share misinformation posts and individuals' reactance to these labels. Furthermore, it enhances our understanding of the influences of prior beliefs and worldviews on individuals when exposed to corrective messages. From a practical standpoint, this study offers valuable insights for fact-checking research and message design. These findings can guide the development of effective designs for corrective messaging on social media platforms, emphasizing the differential effectiveness of various warning label types in discrediting misinformation posts. Additionally, social media platforms should take into account contextual factors, such as worldview inconsistency, when crafting corrective messages.

References

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 1-8. https://doi.org/10.1177/2053168019848554

Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly*, *95*(1), 28-48. https://doi.org/10.1177/1077699016678186

Amazeen, M. A. (2020). Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism, 21*(1), 95-111. https://doi.org/10.1177/1464884917730217

Berinsky, A. J. (2015). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science, 47*(2), 241-262. https://doi.org/10.1017/s0007123415000186

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, *10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619-638. https://doi.org/10.1111/jcom.12166

Brehm, J. W. (1966). *A theory of psychological reactance*. New York, NY: Academic Press.

Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control*. San Diego, CA: Academic Press.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073-1095. https://doi.org/10.1007/s11109-019-09533-0

Fransen, M. L., Smit, E. G., & Verlegh, P. W. (2015). Strategies and motives for resistance to persuasion: an integrative framework. *Frontiers in Psychology, 6*, 1201. https://doi.org/10.3389/fpsyg.2015.01201

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2020). Fake claims of fake news: political misinformation, warnings, and the tainted truth Effect. *Political Behavior*, *43*, 1-33. https://doi.org/10.1007/s11109-020-09597-3

Garrett, R. K., & Poulsen, S. (2019). Flagging facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, *24*(5), 240-258. https://doi.org/10.1093/jcmc/zmz012

Hameleers, M., & van der Meer, T. G. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers?. *Communication Research*, *47*(2), 227-250. https://doi.org/10.1177/0093650218819671

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, London: Guilford Press.

Jackson, S., O'Keefe, D. J., & Brashers, D. E. (1994). The messages replication factor: Methods tailored to messages as objects of study. *Journalism Quarterly*, *71*(4), 984-996. https://doi.org/10.1177/107769909407100421

Kahan, D. (2010). Fixing the communications failure. *Nature*, *463*, 296-297. https://doi.org/10.1038/463296a

Lee, J., Kim, J. W., & Yun Lee, H. (2023). Unlocking conspiracy belief systems: how fact-checking label on twitter counters conspiratorial MMR vaccine misinformation. *Health Communication*, *38*(9), 1780-1792. https://doi.org/10.1080/10410236.2022.2031452

Lees, J., McCarter, A., & Sarno, D. M. (2022). Twitter's disputed tags may be ineffective at reducing belief in fake news and only reduce intentions to share fake news among Democrats and Independents. *Journal of Online Trust and Safety*, *1*(3). https://doi.org/10.54501/jots.v1i3.39

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131. https://doi.org/10.1177/1529100612451018

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal

experiments. *Journal of Experimental Psychology: Applied, 27*(1), 1-16. https://doi.org/10.1037/xap0000315

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet*, *12*(2), 165-183. https://doi.org/10.1002/poi3.214

Moravec, P., Minas, R., & Dennis, A. R. (2018). Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper*, (18-87). http://dx.doi.org/10.2139/ssrn.3269541

Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign misinformation. *Harvard Kennedy School Misinformation Review*. Accessed at https://misinforeview.hks.harvard.edu/article/state-media-warning-labels-can-counteract-the-effects-of-foreign-misinformation/

Novak, V. (2021). *Does Immigration Cost Jobs?*. Retrieved from https://www.factcheck.org/2010/05/does-immigration-cost-jobs/

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303-330. https://www.jstor.org/stable/40587320

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Vraga, E. K., Kim, S. C., Cook, J., & Bode, L. (2020). Testing the Effectiveness of Correction Placement and Type on Instagram. *The International Journal of Press/Politics*, *25*(4), 632-652. https://doi.org/10.1177/1940161220919082

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350-375. https://doi.org/10.1080/10584609.2019.1668894

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The ineffectiveness of fact-checking labels on news memes and articles. *Mass Communication & Society*, *23*(5), 682-704. https://doi.org/10.1080/15205436.2020.1733613

Papakyriakopoulos, O., & Goodman, E. (2022). The impact of Twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets. In *Proceedings of the ACM web conference 2022* (pp. 2541-2551).

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Quick, B. L., Shen, L., & Dillard, J. P. (2013). Reactance theory and persuasion. In *The SAGE handbook of persuasion: Developments in theory and practice* (pp. 167-183).

Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-77

Smith, J. (2017, December 20th). *Designing Against Misinformation*. Retrieved from https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2021). Misinformation warning labels: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. *arXiv preprint arXiv:2104.00779*. https://doi.org/10.48550/arXiv.2104.00779

Sharevski, F., Devine, A., Jachim, P., & Pieroni, E. (2022). Meaningful context, a red flag, or both? preferences for enhanced misinformation warnings among us twitter users. In *Proceedings of the 2022 European Symposium on Usable Security* (pp. 189-201).

Sheehan, K. B. (2018). Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs*, *85*(1), 140-156. https://doi.org/10.1080/03637751.2017.1342043

Stroud, N. J. (2017). Selective exposure theories. In *The Oxford handbook of political communication*.

Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology*, *58*(3), 366–378. http://dx.doi.org/10.1002/asi.20511.

Sears, D. O., & Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, 31(2), 194–213. https://doi.org/10.1086/267513

Roth, Y. & Pickles (2020, May 11th). *Updating our approach to misleading information*. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

Taylor, S., & Asmundson, G. J. G. (2021). Negative attitudes about facemasks during the COVID-19 pandemic: The dual importance of perceived ineffectiveness and psychological reactance. *PLoS One, 16*(2), e0246317. https://doi.org/10.1371/journal.pone.0246317

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460-480. https://doi.org/10.1080/10584609.2015.1102187

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423-441. https://doi.org/10.1080/03637751.2018.1467564

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135-163. https://doi.org/10.1007/s11109-018-9443-y

Funding and Acknowledgements

The author declares no funding sources or conflicts of interest.

Online Connections

X handle @Bingblingzhang