universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„The Effect of Social Media Warning Labels on Reducing the Spread of Misinformation"

verfasst von / submitted by

Juxhina Malaj

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021/ Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 550
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Communication Science
Betreut von / Supervisor:	Univ.-Prof. Dr. Sophie Lecheler

Abstract

From threatening democracy and peace to risking people's health, the rapid spread of false news and misinformation has become one of the greatest challenges of the last decade. Strict measures from regulatory entities and, in particular, from social media giants such as Facebook and Twitter, which have been singled out as "incubators of fake news and propaganda" (Persily & Tucker, 2020), are necessary to stop or prevent the spread of false news and misinformation. The use of warning labels is one of the latest strategies introduced by social media platforms such as Facebook and Twitter to counter the spread of misinformation. However, research on testing which types of warning labels are more effective is still limited.

This study evaluates the effectiveness of Twitter warning labels on message credibility and willingness to share and tests which types of warning labels are more effective in preventing the spread of misinformation. Data from an online survey experiment (N = 208) indicate that unambiguous warning labels such as "This tweet is misleading" significantly lowered the credibility level of misleading posts, compared to other warning labels such as "This tweet may be misleading" or "Find more information here."

The study findings indicate that message credibility mediates the effect of warning labels on willingness to share and that posts without warning labels are perceived as more credible compared to posts with warning labels, but only when the comparison is with an unambiguous warning label. In addition, the study finds that the more credible a post is perceived (no matter if it has a warning label attached or not), the more it will be shared. Online trust and social media self-disclosure did not affect willingness to share. The findings of this experimental study contribute to research on finding effective strategies to prevent misleading posts from being shared or going viral.

Table of Contents

Introduction

To share or not to share a post that contains a warning label attached to it? That is the question for whoever has come across social media posts containing a warning label attached to them (Mena, 2019; Koch et al., 2021). Following the United States presidential election in 2016, the phenomenon of 'fake news' took a life on its own and became a force with significant reach and impact in the media world (Nelson & Taneja, 2018; Clayton et al., 2019; Allcott & Gentzkow, 2017; Pennycook et al., 2018). It also became a hot topic in the public and scientific discourse (Egelhofer & Lecheler, 2019).

After receiving huge criticism for the role it played in "spreading a deluge of political misinformation during the US presidential election" (Heath, 2016, para. 2), Facebook introduced fact-checking and other measures to combat the spread of fake news in December 2016 (Heath, 2016). The rapid spread of misinformation and fake news about the COVID-19 pandemic in 2020 led social media platforms such as Facebook and Twitter to implement new measures by attaching warning labels and fact-checkers to posts and news that may be or are confirmed to be false or misleading by the public health authorities (Sardarizadeh, 2020; Sharevski et al., 2021).

Additional warning labels were attached to any news posts about the Coronavirus, which were contested by third parties for not being accurate, credulous, or trustful (Sardarizadeh, 2020). Since social media is one of the leading platforms where false and misleading news can easily go viral (Waszak et al., 2018), during the pandemic, the world also had to fight against an infodemic (Courtney, 2020). As WHO's director-general, Tedros Adhanom Ghebreyesus said at the 2020 Munich Security Conference, "we're not just fighting an epidemic; we're fighting an infodemic" ("The COVID-19 infodemic", 2020, para. 1).

"Fake, misleading and over-interpreted health news in social media is the potential threat for public health" (Waszak et al., 2018). Considering the fact that a great deal of fake news was spread about COVID-19 cures during the pandemic (Persily & Tucker, 2020), it is crucial to test how effective warning labels are when it comes to reducing the spread of misinformation on delicate topics that have a direct effect on the daily lives of billions of people worldwide. Several empirical studies have tested the effect of warning labels on reducing the spread of false news and misinformation.

However, as the measures introduced by Twitter are relatively new, the majority of previous studies have focused on testing Facebook warning labels (Mena, 2019; Pennycook et al., 2020; Buchanan & Benson, 2019). Only a few studies have focused on testing Twitter warning labels (Starbird et al., 2014; Sanderson et al., 2021) or how to detect and identify fake news spread on Twitter (Reddy et al., 2021; Hamdi et al., 2019). This study aims to fill this research gap by focusing on Twitter warning labels and testing how effective they are.

Moreover, prior research has focused on warning labels attached to posts related to the 2016 U.S. election campaign (Allcott & Gentzkow, 2017; Nelson & Taneja, 2018; Clayton et al., 2019; Pennycook et al., 2018), and there is a research gap when it comes to researching the effect of these warning labels in other contexts. This study will fill this gap by shifting the focus to other important topics and contexts (i.e., climate change and health communication).

Lastly, and more importantly, the majority of previous studies have focused on the effect of warning labels as a whole concept and not on the separate impact of different types of warning labels (Starbird et al., 2014; Sanderson et al., 2021; Reddy et al., 2021; Hamdi et al., 2019). Therefore, this study will also compare different warning labels and find out which types of

warning labels are more effective in reducing the credibility of misleading posts and deterring people from sharing posts with a warning label.

Taking into consideration all the factors mentioned above, this experimental study aims to add to this field of research and further test the impact of social media warning labels on message credibility and willingness to share. What distinguishes this study from previous ones is testing different warning labels and finding out which Twitter warning labels are more effective when it comes to reducing the spread of false news and misinformation. Since labeling initiatives are expected to increase in the near future (Newman, 2019), the results of this study will be important for regulatory entities and social media giants such as Facebook and Twitter to reevaluate their current and future warning labels and take further measures to prevent the spread of misinformation.

2. Literature review

2.1 Social media as an "incubator" of fake news

Whether we realize it or not, social media platforms are defining and shaping our lives on a daily basis, not only online but also offline (Gillespie, 2018). The plethora of social media and social network platforms that we use nowadays has given us a "personal stage" where we have the right to speak and be heard. Facebook, Twitter, LinkedIn, Instagram, Snapchat, Pinterest, and Reddit, all headquartered in California, USA, have been ranked as the seven biggest social media sites in 2020 (Kellogg, 2020). The top two on the list, namely Facebook and Twitter, have been ranked among the most powerful ones for empowering vulnerable groups of people worldwide and pushing the agenda of social change (Guo & Saxton, 2013).

The internet and social media are great innovative tools used to inform and mobilize communities worldwide as well as to supplement advocacy efforts (Scott & Maryman, 2016).

However, these same tools are being used by companies and governments to quietly corrode civil liberties every day (MacKinnon, 2013). In her book *"Consent of the Networked: The Worldwide Struggle for Internet Freedom,"* MacKinnon (2013) says that the time has come for us to stop "arguing over whether the Internet empowers people, and address the urgent question of how technology should be governed to support the rights and liberties of users around the world," (MacKinnon, 2012, para. 1).

Fake news is "any form of falsehood, including rumours, hoaxes, myths, conspiracy theories and other misleading or inaccurate (purposely or not) shared or published content" (Apuke & Omar, 2020; Wang et al., 2019). Recent studies indicate that social media platforms such as Facebook and Twitter have become powerful disseminating mechanisms of misinformation and fake news (Persily & Tucker, 2020; Lyons, 2018; Allcott & Gentzkow, 2017). Terms like "false news," "fake news," and "misinformation" have been used to generally refer to "false and misleading information spread on social media platform" (Mena, 2019; Lyons, 2018). While not all seven silicon valley giants are used as means of spreading fake news, Facebook and Twitter "are more likely to be discussed as incubators of "fake news" and propaganda than as tools for empowerment and social change" (Persily & Tucker, 2020, p. 10).

Once the COVID-19 pandemic hit in 2020 ("Coronavirus disease (COVID-19)", 2020), the number of rumors and misleading stories related to the pandemic started to circulate online (Huynh, 2020). Recent research has shown that most false news and misleading stories circulating in the past two years have been related to the COVID-19 pandemic (Pennycook et al., 2020). The reason why people were willing to share fake news or misleading claims about COVID-19 was "partly because they simply fail to think sufficiently about whether or not the content is accurate when deciding what to share" (Pennycook et al., 2020). "Fake news arises in

equilibrium because it is cheaper to provide than precise signals, because consumers cannot costlessly infer accuracy, and because consumers may enjoy partisan news" (Allcott & Gentzkow, 2017, p. 212). That is why content moderation is essential (Gillespie, 2018).

While social media is regulating our speech in overt and hidden ways, there is still a lot to do when it comes to social media regulation regarding the phenomenon of 'fake news' and the never-ending struggle of eliminating or, at the very least, minimizing it. In his book, *Custodians of the Internet*, Tarleton Gillespie (2018) explains why social media content moderation is not just another function of it. According to Gillespie (2018), content moderation is key to the very existence of social media. Web add-on correction was also found to be effective in decreasing belief in misinformation, especially for "those who use social media for social interaction" (Lee, 2020).

2.2 The "birth" of warning labels on Facebook and Twitter

"After the 2016 election, many feared that fake news articles spread on Facebook swayed the results of the election" (Kurtzleben, 2020, para. 3). Succeeding this, in December 2016, Facebook started "fact-checking, labeling, and burying fake news and hoaxes in its News Feed" (Heath, 2016, para. 1). However, instead of doing the fact-checking on its own, Facebook announced that it would rely on its community and trusted third parties, especially fact-checking organizations that have committed to the International Fact-Checking Code of Principles (Stelter, 2016).

To prevent the spread of fake news about the COVID-19 pandemic in 2020, Facebook and Twitter implemented new measures by attaching warning labels and fact-checkers to posts and news that contain misinformation (Rosen et al., 2019). According to Facebook's VP of Integrity, "since March 2020 the company put "warning labels" on 180 million pieces of content

— meaning a third-party fact-checker reviewed and debunked them" (Kraus, 2021, para. 2). In March 2020, Twitter broadened its policy guidance (Roth & Pickles, 2020). Then, in May 2020, it introduced its own labels and warning messages aiming to "address content that goes directly against guidance on COVID-19 from authoritative sources of global and local public health information" (Roth & Pickles, 2020, para. 2). Twitter used this same strategy during the 2020 U.S. presidential election since its initial plan was to continue using this new form of censure for other situations "to provide additional explanations or clarifications in situations where the risks of harm associated with a Tweet are less severe but where people may still be confused or misled by the content" (Roth & Pickles, 2020, para. 2). Analogous to Facebook, Twitter monitors the content with the help of trusted third parties besides its own teams (Roth & Pickles, 2020).

2.3 The "backfire effect" of warning labels and the "implied truth"

While warning labels are a good strategy for discouraging people from sharing news or posts that are flagged, this strategy can create a "backfire effect" (Lewandowsky et al., 2012). A decade ago, several studies found out that the "flagging" of fake news or misinformation may reinforce the belief that the flagged posts are actually to be trusted (Mena, 2019). "Such backfire effects, also known as "boomerang" effects, are not limited to the correction of misinformation but also affect other types of communication" (Lewandowsky et al., 2012, p. 119).

For instance, if a campaign aims to reduce smoking rates, the same campaign may lead to a spike in smoking rates (Lewandowsky et al., 2012). Similarly, when it comes to correcting messages that aim to lower the belief in such posts, the same labels may cause people to continue relying on misinformation (Lewandowsky et al., 2012; Nyhan & Reifler, 2010). Moreover, studies have found that backfire effects can be manifested not only in the form of rejecting the message of the warning label but also the source behind it (Brehm & Brehm, 1981). This

"backfire effect" has been found not only for political misinformation (Nyhan & Reifler, 2010) but also when it comes to false and misleading posts on health communication such as vaccine promotion (Nyhan et al., 2014). Furthermore, the "implied truth" phenomenon is becoming increasingly prominent (Pennycook et al., 2020). While fact-checkers and warning labels are gaining popularity and being used as a strategy by social media platforms to reduce the spread of misinformation, unlabeled posts and news that might share false or misleading information are seen as more accurate simply because they do not contain any warning labels attached to them (Pennycook et al., 2020).

Despite the persistent backfiring effects of warning labels, several studies have found that certain warning labels successfully reduce the spread of misinformation without causing a backfire (Wood & Porter, 2018; Young et al., 2017). "Video is an effective way of correcting misperceptions while also generating interest in fact-checking information and reducing viewer confusion compared with print-based fact-checks" (Young et al., 2017, p. 71). A recent study comparing video with online messaging apps also found that users perceive the content of a story as more credible when it was in video format than audio and text" (Sundar et al., 2021). These studies indicate that the format of the content and that of the fact-checkers is crucial when it comes to how credible such content is perceived by users (Young et al., 2017; Sundar et al., 2021).

3. Research questions and hypotheses

3.1 The effect of warning labels on willingness to share

When it comes to social media, not all users engage with the type of content they see in their feed or receive via personal messages in the same way. Some users are more active than others and willing to share more than others; some users act only as 'observers' and do not

engage with any type of content; while others reduce their sharing frequency once they get a stable social media following (Weller, 2016). People are also found to share content that aligns with what they themselves believe is true (An et al., 2014; Pennycook et al., 2020).

A study by Pennycook (2020) titled *"The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings"* found other factors that affect users' willingness to share on social media, such as the "desire to signal one's group membership" and reputational consequences. According to a recent study, when sharing content, "people were considering not only whether they believed a headline to be true but also what their friends would believe—in an effort to avoid potential reputational consequences of being seen to share false content" (Pennycook et al., 2020, p. 4955).

The rapid spread of fake news and the "birth" of warning labels on social media platforms such as Facebook and Twitter have also affected how people interact with social media content, how often they share, and what they share. Even though Facebook has not been keen to share whether its warning labels are successful or not when it comes to fighting the spread of false news (Mantzarlis, 2017; Kraus, 2021; Pennycook et al., 2020), a study by Paul Mena (2019) titled *"Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook"* found that false news that is flagged is less likely to be shared on social media compared to other news that is not flagged.

Unlike Facebook (Mantzarlis, 2017), Twitter has been open about the effectiveness of its warning labels. In November 2020, Twitter claimed that its warning labels "slowed [the] spread of false election claims" (Wells, 2020). A study by Pennycook et al. (2020) found that giving people a simple reminder about the accuracy of the headlines they are about to read improved their level of truth discernment and subsequently helped them make better decisions regarding

their willingness to share. In the world of social media, warning labels serve as such reminders to help people make more informed decisions. Other studies have also found that warning labels help reduce the spread of misinformation by decreasing people's willingness to share (Lewandowsky et al., 2012; Pennycook et al., 2018; Schul, 1993). "Misinformation effects can be reduced if people are explicitly warned up front that information they are about to be given may be misleading" (Lewandowsky et al., 2012, p. 116).

Considering the findings of previous studies, it is expected that posts with a warning label will be less shared than posts without warning labels. Therefore, the following main effect hypothesis was formulated:

H1: Posts with warning labels are less likely to be shared compared to posts without warning labels.

The color and positioning of social media warning labels have been found to have an important role when it comes to the effectiveness of warning labels and how they are perceived (Nassetta & Gross, 2020). Since there are different Twitter warning labels (Roth & Pickles, 2020; Gao et al., 2018) depending on the content being shared, it is expected that not all warning labels will have the same effect on people's willingness to share. Therefore, the study explores the following research question:

RQ 1: What type of warning label is more effective in deterring people from sharing posts?

For this study, three Twitter warning labels will be taken into consideration: Negative warning labels such as "This tweet is misleading" and "This tweet may be misleading" and neutral or positive warning labels such as "Find more information here." Since social media posts and headlines that have tags like "Disputed" or "Rated false" have been found to be perceived as less accurate (Clayton et al., 2019), it is expected that warning labels with a negative connotation will

be less likely to be shared compared to posts that have a warning label that says “Find more information here.” Therefore the following hypothesis was formulated:

H2: People are less likely to share posts with negative warning labels compared to posts with a positive or neutral warning label

Besides exploring which type of warning label is more effective in deterring people from sharing false news, this study also aims to find out which kind of warning label is more effective in lowering the credibility of the message perceived from the post that contains such a warning label. Therefore, this study asks the following research question:

RQ 2: What type of warning label reduces credibility the most?

In order to be more effective, warning labels need to be as unambiguous as possible (Ecker et al., 2010). “... to be effective, such warnings need to specifically explain the ongoing effects of misinformation rather than just generally mention that misinformation may be present” (Lewandowsky et al., 2012, p. 116). Recent studies also found that people are less likely to share posts that contain clear labels (Gao et al., 2018). Since the warning label “This tweet is misleading” is the most unambiguous one compared to “This tweet may be misleading” and “Find more information here,” it is expected that the warning label “This tweet is misleading” will lower the credibility level the most. Therefore, the following hypothesis was formulated:

H3: The more unambiguous the warning label, the lower the message credibility level

3.2 The mediating effect of message credibility

The majority of Twitter posts are found to be truthful and credible by previous studies. However, as with other social media platforms, Twitter is also used to spread misinformation (Castillo et al., 2011). To better understand the relationship between Twitter warning labels and their effect on people’s willingness to share, it is also essential to explore the indirect effect of

warning labels on willingness to share. Research has identified message credibility as one of the "key ingredients in the online information assessment" (Li & Suh, 2015) and as a mediator variable that explains the relationship between warning labels and willingness to share (Mena, 2019; Pennycook et al., 2020).

"An important aspect of understanding social media users' behaviors concerning the sharing of false news is identifying how credible misleading messages are perceived to be" (Mena, 2019, p. 170). Twitter's addition of warning labels to misleading posts in the past years (Roth & Pickles, 2020) has made it easier for people to assess the message credibility of tweets containing warning labels before deciding to share them. Therefore, the following hypothesis was formulated:

H4: The effect of warning labels on willingness to share is mediated by message credibility.

This study will add to the body of literature (Mena, 2019; Pennycook et al., 2020) by testing the indirect effect of warning labels on willingness to share through message credibility. Warning labels have been found to affect the message credibility of posts (Oremus, 2019) while message credibility is found to be one of the key factors determining people's willingness to share on social media (Pennycook et al., 2020). Given the literature discussed above, H4 is split into two parts that explore each of the paths:

H4a: Posts without warning labels are perceived as more credible compared to posts with warning labels

H4b: The more credible a post is perceived, the more it will be shared

3.3 Online Trust and Self-Disclosure

A recent study suggested that five elements are linked with the phenomenon of sharing fake news on social media: online trust, self-disclosure, social comparison, fear of missing out,

and social media fatigue (Talwar et al., 2019). For this study, two of those elements have been selected to be tested and included in the research design: online trust and self-disclosure. According to a recent study, a "high level of online trust encourages people to provide more social support and take risks in sharing information" (Talwar et al., 2019, p. 74). Therefore, the following hypothesis was formulated:

H5a: The effect of warning labels on willingness to share is moderated by online trust in that participants with lower online trust will be less affected by warning labels.

Besides trusting online information shared on social media, online trust also includes trust in warning labels attached to posts since they also appear online and are added by third parties or social media platforms themselves (Roth & Pickles, 2020). "Self-disclosure represents sharing of personal information with others" (Talwar et al., 2019, p. 75), and it is motivated by the need for popularity and "the need to increase mutual understanding, strengthening relationships, and enhancing bonds among group members" (Talwar et al., 2019, p. 75).

This need to win popularity points and create a stronger bond with online connections leads people with high social media self-disclosure to easily share fake news without ensuring the post is trustworthy before sharing (Talwar et al., 2019) and disregarding warning labels. Therefore, the following hypothesis was formulated:

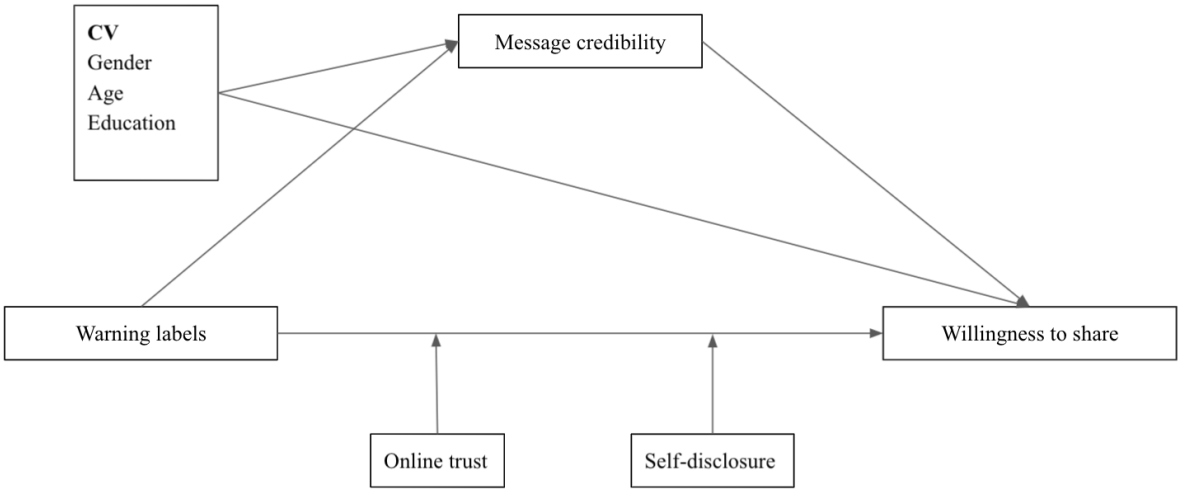
H5b: The effect of warning labels on willingness to share is moderated by self-disclosure, in that participants with low self-disclosure will be more affected by warning labels.

4. Research Design and Methodology

4.1 Theoretical Model

Figure 1

Theoretical Model



As seen in Figure 1, the theoretical model consists of one independent variable, one dependent variable, one mediator variable, two moderator variables, and three covariates. The independent variable (warning labels) is manipulated into four levels for each group. Twitter takes different measures and attaches various warning labels depending on the content being shared (Roth & Pickles, 2020). If a post contains misleading information or has a disputed claim and the harm is moderate, Twitter attaches a label to it, mentioning that the Tweet may be misleading. If the post includes a disputed claim and the harm is severe, Twitter will attach a "harsher" warning label to it. In case a post contains misleading information and the harm is severe, Twitter deletes the post. However, for posts that contain unverified claims, Twitter takes no action.

For this study, three types of warning labels inspired by Twitter warning labels were tested: Label 1 ("This tweet is misleading"), Label 2 ("This tweet may be misleading"), and Label 3 ("Find more information here"). Group one received three tweets without any warning labels; group two received three tweets with the warning label "This Tweet is misleading"; group

three received three tweets with the warning label “This Tweet may be misleading”; and group four received three tweets with the warning label “Find more information here.”

The independent variable in this study is “willingness to share,” which depends on the warning label of the tweet or the lack thereof. To further explain the relationship between the independent and independent variables, the effect of two moderator variables, online trust and social media self-disclosure was explored. This study also tested the indirect effect of warning labels on willingness to share through a mediator variable, which is message credibility. Since demographic characteristics such as gender, age, and education might also have an effect on the direct relationship between the independent variable and the dependent variable (Allcott & Gentzkow, 2017), the study will control for these three covariate variables.

4.2 Research Method

An online survey experiment was used as the research method for this study. This methodological approach was the best method to empirically test the effect of social media warning labels on people's willingness to share and test which warning label is more effective when it comes to decreasing message credibility, and in turn, reducing the spread of misinformation. Another reason this research method was chosen was to give people from all over the world, from different cultural backgrounds, the opportunity to easily participate in an online survey experiment, especially since the data had to be collected during a global pandemic. Reaching a broad audience through an online survey experiment will ensure that the results of this study have high external validity.

Online Survey Experiment

Since the participants of this survey were not paid, the survey was kept as concise as possible so it would not take participants more than seven minutes to complete on average. All of

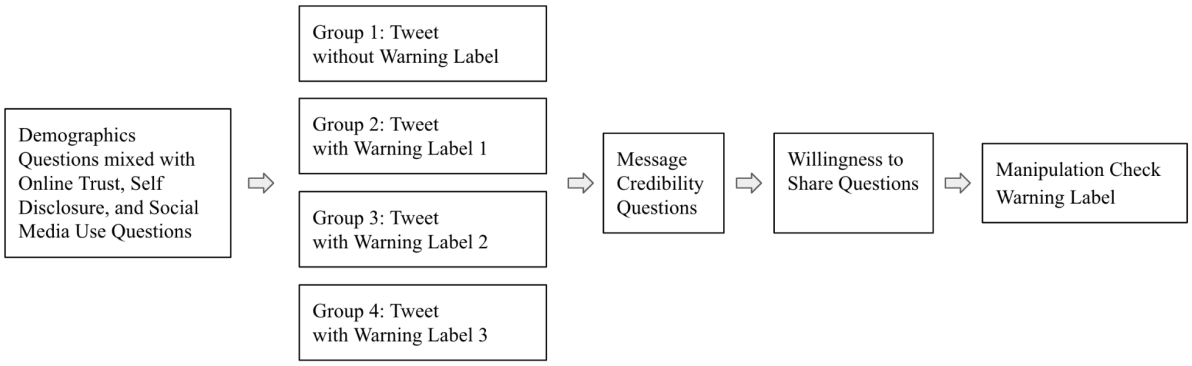
the tweet screenshots (see Appendix C) were kept precisely the same for all groups. The only difference was the addition of a warning label for groups two, three, and four. Group one, which was used as a control group, only received three screenshots of tweets without any type of warning label.

Several tweet topics were chosen. One was related to climate change, one to health, and one to a political issue. Different topics were selected for each set of tweets per group to eliminate the possibility of one specific topic affecting the test results. Since the study aims to test the effect of warning labels and not which type of topics people share the most, having a variety of topics ensured the focus would remain on the warning labels attached to the tweets.

All of the tweets used in this study were based on actual tweets (see Appendix B) published from existing accounts; however, all of them were modified for the purpose of the study. All of the tweets were modified using the "Tweet Generator: Tweetgen" (2018) tool. The profile picture, Twitter handle, date, time, and tweet engagement numbers were among the details that were modified. The text of the original tweets was also changed slightly, a technique used by people and accounts that aim to spread false information (Mena, 2019). “It is usual that false news creators blend real and false information to craft false news stories” (Mena, 2019, p. 173). The photos attached to the original tweets were also removed so that the focus remained on the message of the tweet and not the picture accompanying it. A warning label was added to the tweet screenshots for groups two, three, and four, using the Tweet Generator: Tweetgen (2018) tool.

Figure 2

Online Survey Experiment



As shown in Figure 2, the online survey experiment design contained three main sections to explore this study's research questions and hypotheses. The first section consisted of demographics questions mixed with several questions on online trust, self-disclosure, and social media use. The demographic questions were asked at the beginning of the survey instead of the end to "act" as a distractor so that the study participants would not be pre-conditioned to think what the study would be about right away. When mixed with other questions regarding the two moderator variables, demographic questions helped create a balance and an easy opening section for the participants. After completing the first section, the participants were randomly assigned to one of the four groups. SoSci survey's PHP code (Leiner, 2021) was used for the randomization generator, which placed people into one of these groups:

Group 1: 3 posts without any labels (no warning, just the tweet)

Group 2: 3 posts with label 1- "This Tweet is misleading" (misinformation)

Group 3: 3 posts with label 2 - "This Tweet may be misleading" (disputed tweet)

Group 4: 3 posts with label 3 - "Find more information here" (neutral/positive)

In the second section, participants were shown three different screenshots of tweets about health, climate change, and politics. After each screenshot of each tweet, they were asked to evaluate the tweet's credibility from '-3' to '3', where '-3' was "not credible at all," and '3' was "very credible." After each question on message credibility, the participants were asked to evaluate how willing they are to share the post on their social media feed and how willing they are to share the tweet with a colleague, a friend, or a family member.

The dependent variable was split into two items to better evaluate respondents' willingness to share. Furthermore, previous studies have found that people are more likely to share fake news with family members than on their social media feeds (Sundar et al., 2021). Therefore, this distinction in willingness to share would help understand the sharing process of false news better and give some directions on which future studies can focus. Participants had to evaluate their "willingness to share from '-3' (not likely at all) to '3' (very likely.)

The last section consisted of a few questions regarding detecting tweets that contain misleading or false information, manipulation check, and warning label. The participants had to answer whether they noticed any tweets that contained misleading or false information and select the tweet or tweets which they thought included such information. In the end, to conduct a manipulation check, the participants were asked whether they noticed any warning labels attached to the tweets shown to them. If they answered "Yes," they were asked to evaluate how helpful and informative the warning label was for them. After the participants completed the survey, they were informed that all the tweets they saw were based on actual tweets but modified for this study.

4.3 Scale Measurements

The scales used for measuring Online Trust and Self-Disclosure (see Appendix A) were based on the scale used by Talwar et al. (2019). The results of the study by Talwar et al. suggested that "online trust, self-disclosure, fear of missing out (FoMO), and social media fatigue are positively associated with the sharing fake news (intentionally)....The study findings also indicate that online trust has negative association with authenticating news before sharing" (2019). To make the study less complicated and shorten the questionnaire, only two constructs were selected as moderator variables for this study: Online Trust and Self-Disclosure.

Respective measurement items from Talwar et al. (2019) were used for this study. The only modification to the questions was replacing "WA" (WhatsApp) with "Social Media." The reason why WhatsApp was not replaced with Twitter was due to two reasons. The first reason is the fact that there was no guarantee that all participants were active on Twitter. As the test results showed, only 23.6% of the participants had a Twitter account. Therefore the questions were kept for social media channels in general. The second reason was to avoid giving away too much information by asking participants about Twitter early on. This was important not to affect the results of their answers since these questions were asked early on.

Validity and reliability

Online Trust has a total of two measurement items, while Self-Disclosure has a total of three measurement items. A reliability test was conducted using SPSS Statistics, Version 27 (IBM Corp, 2021) to ensure that both scales for measuring Online Trust and Self-Disclosure had a high internal consistency. Cronbach's Alpha for the Online Trust scale was equal to .868, indicating that the scale has a high internal consistency. Cronbach's Alpha for the Self-Disclosure

scale was equal to .873, indicating the scale has a high internal consistency and is reliable as well.

The scale used for willingness to share was also based on the same study (Talwar et al., 2019), but the wording of both measurement items was changed to better fit the purpose of this study and measure the dependent variable, which is "willingness to share." The Cronbach's Alpha for willingness to share was equal to .875, indicating that the scale measuring willingness to share has a high internal consistency and reliability.

5. Data Collection

5.1 Pre-test

A pre-test using an online survey experiment was conducted in early August 2021. A total of 24 participants (six participants per group) participated in the pre-test. After the pre-test, several minor changes to the online survey were made based on the feedback received from respondents, university colleagues, and the professor. More options were added to the alternatives regarding 'social media use' and social media channels. Initially, the participants were shown the tweet screenshots and then asked to answer with either a "Yes," "No," or "Maybe" about their willingness to share after each post. For the final test, a second question was added regarding their willingness to share, and the participants had to choose from '-3' to '3' regarding willingness to share. This made it possible to test the willingness to share variable in more depth and get more info using a 7-point Likert scale.

Initially, participants were asked to evaluate the credibility of all the tweets shown to them by answering three questions based on the credibility scale developed by Paul Mena (2019). The problem with this initial approach was that the credibility questions were asked after all the tweets were shown to participants. However, for the main test, it was necessary that

besides asking participants whether they would share the post, they would also be asked about message credibility for each tweet separately.

Asking three questions for message credibility and two questions for willingness to share would be quite long for the participants. Therefore, the credibility scale was disregarded for the main test. Participants were only asked one question regarding credibility and two questions regarding willingness to share per each tweet shown to them. Lastly, all 5-point Likert scales (from 1 to 5) used for the pre-test were changed to 7-point Likert scales (from -3 to 3) for the main test so that the participants had a greater understanding of the extreme points used in the survey. The answers were coded from '1' to '7' on SPSS, where '-3' was coded as '1', and '3' was coded as '7'.

5.2 Test and sample size

A sample of (N =328) participants participated in the survey from August 15, 2021, until September 8, 2021. The questionnaire was made available for anyone 18 years or older and who understands English. Out of 328 participants, 243 finished the survey. After cleaning the data and removing the participants who completed the survey too fast based on high degradation points (removing the ones with degradation points over 72), the dataset consisted of 208 valid cases. For the online survey experiment, participants were randomly assigned into four different groups. Therefore, each group consisted of 52 participants.

All participants took part in this study voluntarily, and their participation was kept anonymous and confidential. They were also asked to consent before starting the survey. SoSci Survey platform (Leiner, 2021) was used to compile the survey. The survey was shared with students of the University of Vienna, and on different online platforms such as SurveyCircle as well as LinkedIn and Facebook groups tailored explicitly for survey exchange. The survey was

also shared with European Forum Alpbach 2021 scholarship holders and alumni, who come from over 60 countries all over the world. Considering that the study was targeted at various people from different countries and backgrounds, the collected sample represents a diverse group of participants.

5.3 Demographics

There were 70.7% female respondents [N = 147] , 26.9% male participants [N = 56], 0.5% [N = 1] Non-Binary, and the rest (about 2% [N = 3]) selected “other” or did not disclose their gender.

Meanwhile, 41% of the participants [N = 85] belonged to the age range 23-27; 28% of the participants [N = 58] to the age range 28-32; 10% to the age range 18-22 [N = 21], and 7.7% to the age range 33-37. 10% belonged to the age range 38 and above, and about 1% preferred not to disclose their age.

Regarding the level of education, the majority of the participants [N = 186] had either a Bachelor's degree (35.6%), a Master’s degree (48%), or a Doctorate (5.8%). The rest of the participants (around 10.5%) had completed some college coursework. (4.3%) completed high school (3.8%), and 2% had a Technical or occupational certificate.

6. Results

6.1 Data analysis methods

SPSS Statistics, Version 27 (IBM Corp, 2021), was used to analyze the collected data. PROCESS version 4.0 by Hayes (2021) was also used for mediation and moderation analysis. The research model was tested using a three-step approach. First, several reliability tests were done to confirm that the scales used to measure the moderator variables and the dependent variable had a high internal consistency and were reliable. Second, ANOVA tests were used to

test the first three hypotheses. Third, Process by Hayes (2021) was used to test the last hypotheses about mediation and moderation effects.

6.2 Main Effect

H1 predicted that posts with warning labels would be less likely to be shared than posts without warning labels. Since the independent variable is a categorical predictor, a one-way between-group ANOVA was run to determine the main effect of warning labels on willingness to share. This study found that most of the respondents, no matter in which group they belonged, leaned towards the "not likely to share this tweet at all" alternative from the 7-point Likert Scale. The respondents who received tweets with the warning label "This tweet is misleading" had a lower willingness to share the post on their social media feed or with a colleague, a friend, or a family member compared to all the other three groups. However, this difference was not statistically significant. Therefore, no support was found for H1.

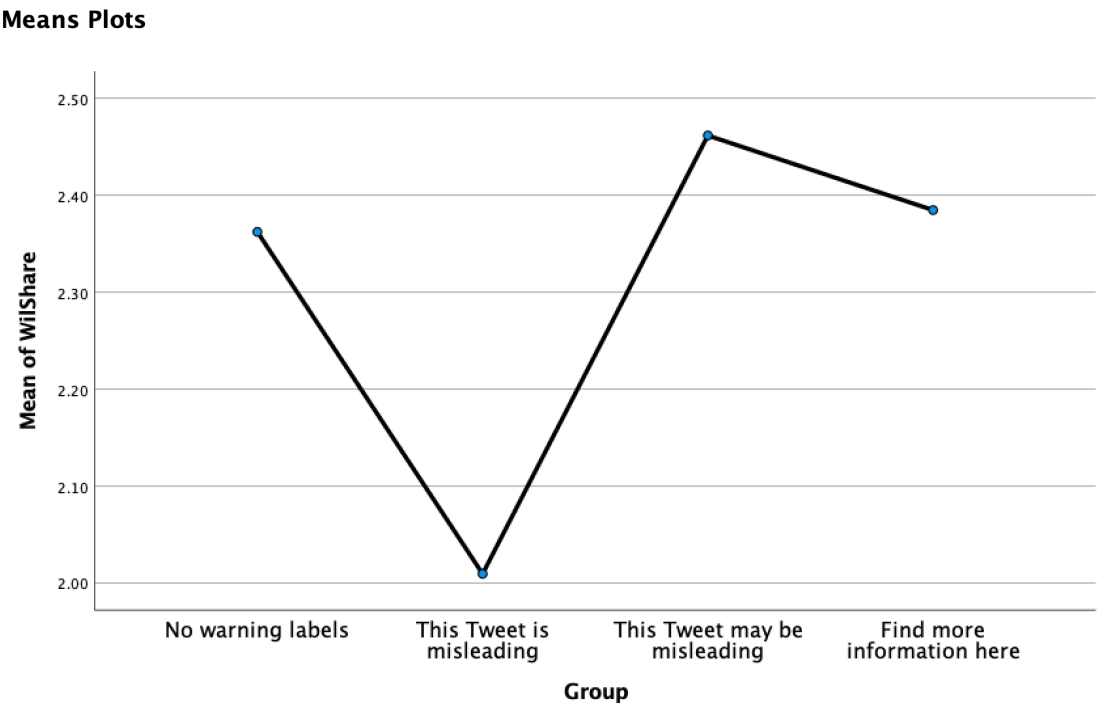
The test found that the mean for willingness to share for all groups was around 2 out of 7: Group 1 without warning labels ($M = 2.36$, $SD = 1.24$), group 2 with warning label 1 "This Tweet is misleading" ($M = 2.01$, $SD = 0.97$), group 3 with warning label 2 "This tweet may be misleading" ($M = 2.46$, $SD = 1.5$), and group 4 with warning label 3 "Find more information here" ($M = 2.38$, $SD = 1.26$). The difference between groups was not statistically significant, $F(3,204) = 1.32$, $p = .27$. LSD and Bonferroni post hoc tests also confirmed that the difference between groups was not statistically significant.

H2 predicted that people are less likely to share posts with negative warning labels such as warning label 1 "This tweet is misleading" and warning label 2 "This tweet may be misleading" compared to posts with a positive or neutral warning label such as warning label 3

"Find more information here." Since the difference between groups was not statistically significant, no support was found for H2.

Figure 3

Willingness to share

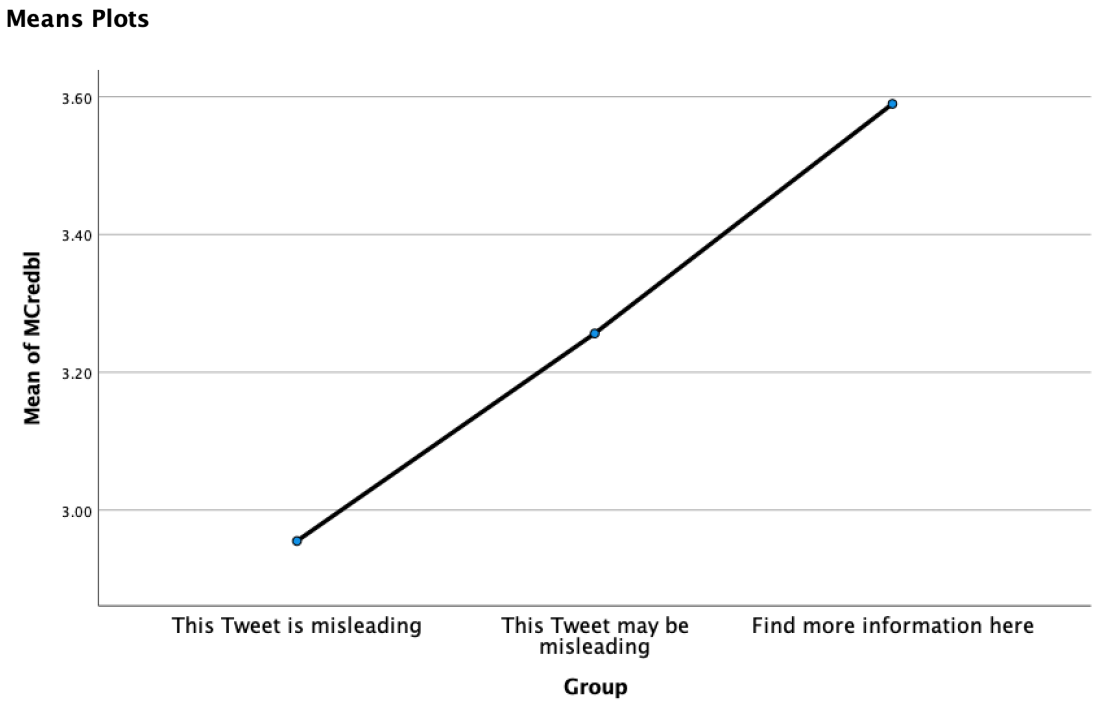


As seen in Figure 3, the second group, which received tweets containing warning label 1, "This Tweet is misleading," is the least likely to share the tweets compared to the other groups that contain no warning labels or neutral labels such as "Find more information here." The mean for willingness to share for tweets containing warning label 1- "This Tweet is misleading" (M = 2.0) is lower compared to the means of warning label 3, "Find more information here." However, the difference is not statistically significant. Group 3, which received the other "negative" warning label "This tweet may be misleading," was more likely to share the tweets (M = 2.46) compared to Group 4 (M= 2.38), which received tweets containing a more neutral warning label "Find more information here." However, the difference was not statistically significant.

H3 predicted that the more unambiguous the warning label, the lower the level of message credibility would be. Out of the three warning labels, the most unambiguous warning label was warning label 1, “This tweet is misleading.” A one-way between-group ANOVA was run to determine the main effect of warning labels on message credibility and find out what type of warning label reduces credibility the most. The test was run only for groups 2, 3, and 4, leaving out group 1, which did not receive any warning labels.

Figure 4

Message credibility



As seen in Figure 4, the mean of message credibility for Group 2, which received three tweets with Label 1 “This tweet is misleading,” is equal to 2.95. The mean of message credibility for Group 3, which received three tweets with Label 2 “This tweet may be misleading,” is equal to 3.25, which means that group 3 scored a higher credibility level than group 2. The mean of

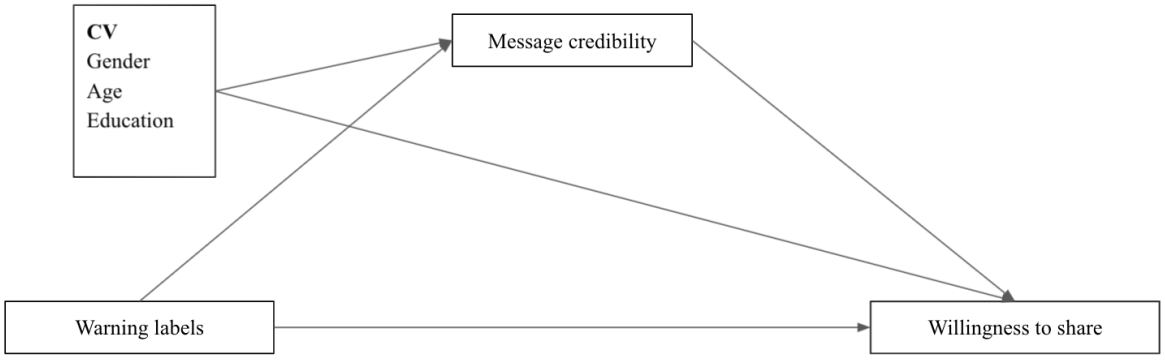
message credibility for Group 4, which received three tweets with Label 3 “Find more information here,” is equal to 3.6, indicating Group 4 scored a higher level of credibility than the other two groups, which received a more “negative” warning label.

The Homogeneity of Variances tests ($p > .05$) showed that the homogeneity variances are not significantly different. Therefore equal variances are assumed in this case. The ANOVA test, $F = 3.1$, $p = .028$, showed that the difference between groups regarding the level of message credibility was statistically significant. Therefore, the study found support for H3: The more unambiguous the warning label, the lower the message credibility level. The LSD post hoc test showed that there is no statistically significant difference between warning label 1 “This tweet is misleading” and warning label 2, “This tweet may be misleading” ($p = .203$). However, the difference between warning label 1, “This tweet is misleading,” and warning label 3, “Find more information here,” was statistically significant ($p < .005$).

6.3 Mediating Effect

Figure 5

Mediating effect



Model 4 in PROCESS macro (Hayes, 2021) was used to test whether the effect of warning labels on willingness to share was mediated by message credibility. Since Model 4 in Process is used to test mediation, it was the relevant one to test mediation effects. Model 4 is also perfect for assessing the statistical significance of the direct effect of warning labels on willingness to share as well as the indirect effect of the independent variable to the dependent variable through message credibility, which is the mediating variable.

The same model in Process was run for all three comparisons to test the mediation effects when comparing the control group (no warning labels) with one of the other groups. Dummy variables were made to compare Group 1 with each of the other groups. The control group was always coded as 0, while the other group (for each of the comparisons) was coded as 1. Three covariates (gender, age, and education) were added to the model for all groups. For all comparisons, 95% bias-corrected bootstrap confidence intervals (CIs) with 10000 bootstrap samples were computed.

Group 1 vs. Group 2

According to the generated Model 4 matrix, the entire model explained 36% of the variance in willingness to share, $R^2 = .36$, $F(5, 98) = 10.9$. Therefore, 36% of the variance can be attributed to all our predictors to Y. The model is also statistically significant, $p<0.001$. Thus, H4: The effect of warning labels on willingness to share is mediated by message credibility was supported. The effect of covariates was not statistically significant regarding the entire model. H4a predicted that posts without warning labels are perceived as more credible than posts with warning labels (in this case, compared to warning label 1, "This tweet is misleading." Model 4 analysis showed that H4a was supported, $b= -.47$, $t(4, 99)= -2.1$, $p<0.05$. For every one-unit increase in warning labels, credibility decreases by 0.47 units. Therefore, posts with warning label 1 are found less credible compared to posts without warning labels.

The effect of participants' gender on message credibility was not statistically significant, $b= -.17$, $t(99)= -1.1$, $p=.25$. The effect of participants' age on message credibility was also not statistically significant, $b= -.04$, $t(99)= -.56$, $p=.57$. However, participants' education had a statistically significant effect on message credibility, $b= -.24$, $t(99)= -2.01$, $p=.04$.
H4b predicted that the more credible a post is perceived, the more it will be shared. H4b was supported. The effect of message credibility on willingness to share was statistically significant, $b= .54$, $t(98)= 6.77$, $p<0.001$. The total effect of warning labels on willingness to share was not statistically significant, $b= -.29$, $t(99)= -1.3$, $p=.19$.

The effect of covariates on the total effect was also not statistically significant. The indirect effect of warning labels on willingness to share was equal to -.26, SE=.13, 95% CI[-.53, -.01]. Since the confidence interval does not include 0, we can conclude that there is mediation when comparing group 1 with group 2. The standardized effect was equal to -.23, SE=.11, 95%

CI[-.46, -.01]. Since the confidence interval does not include 0, we can conclude that there is mediation when comparing group 1 with group 2.

Group 1 vs. Group 3

According to the generated Model 4 matrix, the entire model explained 43% of the variance in willingness to share, $R^2 = 0.43$, $F(5, 98) = 14.7$. Therefore, 43% of the variance can be attributed to all our predictors to Y. The model is also statistically significant, $p<0.001$. Thus, H4: The effect of warning labels on willingness to share is mediated by message credibility was supported. The effect of covariates was not statistically significant regarding the entire model. H4a predicted that posts without warning labels are perceived as more credible than posts with warning labels (in this case, compared to warning label 2 "This tweet may be misleading." Model 4 analysis showed that H4a was not supported, b= -.28, t(4, 99)= -1.1, p=.25. For every one-unit increase in warning labels, credibility decreases by 0.28 units. Therefore, posts with warning label 2 are found less credible compared to posts without warning labels. However, this decrease in credibility is not statistically significant.

H4b predicted that the more credible a post is perceived, the more it will be shared. H4b was supported. The effect of message credibility on willingness to share was statistically significant, b= .71, t(98)= 8.25, p=<0.001. Moreover, the total effect of warning labels on willingness to share was not statistically significant, b= .08, t(99)= .29, p=.77. The effect of covariates on the total effect was also not statistically significant. The indirect effect of warning labels on willingness to share was equal to -.19, SE=.17, 95% CI[-.55, .14]. Since the confidence interval includes 0, we can conclude that no mediation occurs when comparing groups 1 and 3. The standardized effect was equal to -.14, SE=.12, 95% CI[-.40, .10]. Since the confidence interval includes 0, we can conclude that no mediation occurs when comparing groups 1 and 3.

Group 1 vs. Group 4

According to the generated Model 4 matrix, the entire model explained 32% of the variance in willingness to share, $R^2 = 0.32$, $F(5, 98) = 9.16$. Therefore, 32% of the variance can be attributed to all our predictors to Y. The model is also statistically significant, $p<0.001$. Therefore, H4: The effect of warning labels on willingness to share is mediated by message credibility was supported. The effect of covariates was not statistically significant regarding the whole model.

H4a predicted that posts without warning labels are perceived as more credible compared to posts with warning labels (in this case, compared to warning label 3 "Find more information here." Model 4 analysis showed that H4a was not supported, b= .015, t(4, 99)= .07, p=.94. For every one-unit increase in warning labels, credibility increases by 0.015 units. Therefore, posts with warning label 3 are found more credible compared to posts without warning labels. However, this increase in credibility is not statistically significant. The effect of participants' gender on message credibility was statistically significant, b= -.37, t(99)= -1.99, p=.04. The effect of participants' age on message credibility was not statistically significant, b= .034, t(99)= .46, p=.64. Participants' education also had a non statistically significant effect on message credibility, b= -.07, t(99)= -.87, p=.38.

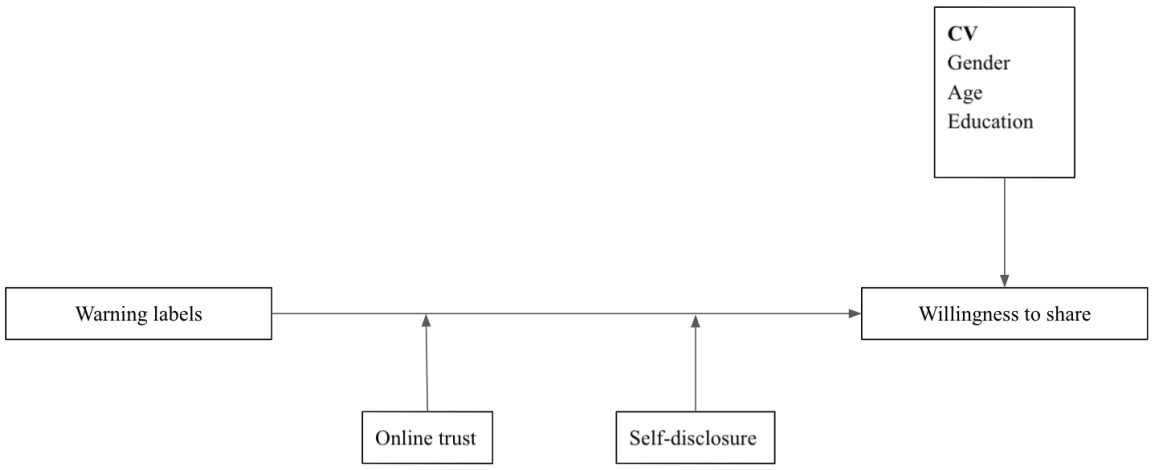
H4b predicted that the more credible a post is perceived, the more it will be shared. H4b was supported. The effect of message credibility on willingness to share was statistically significant, b= -.65, t(98)= 6.47, p<0.001. The total effect of warning labels on willingness to share was not statistically significant, b= -.02, t(99)= -.07, p=.93. The effect of covariates on the total effect was also not statistically significant. The indirect effect of warning labels on willingness to share was equal to .01, SE=.13, 95% CI[-.25, .29]. Since the confidence interval

includes 0, we can conclude that no mediation occurs when comparing group 1 with group 4. The standardized effect was equal to .01, SE=.11, 95% CI[-.19, .23]. Since the confidence interval includes 0, we can conclude that no mediation occurs when comparing group 1 with group 4.

6.4 Moderating Effect

Figure 6

Moderating effect



Model 2 in PROCESS macro (Hayes, 2021) was used to test the moderating effect of online trust and self-disclosure on willingness to share. Since Model 2 in Process is used to test moderation, it was the relevant one to test moderation effects. Dummy variables were created to compare Group 1 (without warning labels) with all the other groups (with warning labels). The control group was coded as 0, while all the other groups were coded as 1. Three covariates (gender, age, and education) were added to the model. For all comparisons, 95% bias-corrected bootstrap confidence intervals (CIs) with 10000 bootstrap samples were computed. According to the

matrix generated by Model 2 in Process, the entire model explained a significant proportion of variance in willingness to share, R^2=.16, F(8,199)=4.7515. Therefore 16% of the variance can be attributed to our predictors to Y, and this model is statistically significant, $p < 0.001$.

H5a predicted that the effect of warning labels on willingness to share is moderated by online trust in that participants with lower online trust will be less affected by warning labels. Results showed that the moderating effect of online trust on willingness to share was not statistically significant, b= .09, t(199) = .56, p=.57. Therefore, no support was found for H5a. H5b predicted that the effect of warning labels on willingness to share is moderated by Social Media Self Disclosure, in that participants with low Social Media Self-disclosure will be more affected by warning labels. Results showed that the moderating effect of self-disclosure was not significant, b= .12, t(199) = 1.42, p=.35, Therefore, no support was found for H5b.
Likewise, neither of the covariates (gender, age, and education) has a statistically significant effect on willingness to share.

7. Discussion

This study tested the effectiveness of Twitter warning labels in preventing the spread of misinformation and compared three types of warning labels to test which kind of warning label is more effective in reducing message credibility and people's willingness to share. Since the sample used to test the hypothesis of this study was relatively homogeneous, the difference between groups was not statistically significant, and the study found that most of the respondents, no matter in which group they belonged to, leaned towards the "not likely to share this tweet at all" alternative from the 7-point Likert Scale. Therefore, no support was found for H1 and H2, which tested the main effect of warning labels on willingness to share and whether

there was any difference in willingness to share when comparing negative warning labels with a positive or neutral warning label.

As the majority of the respondents (around 90%) had either a Bachelor's degree (35.6%), a Master's degree (48%), or a Doctorate (5.8%), the level of education played an important role when it comes to testing the indirect effect of warning labels on willingness to share through message credibility, as indicated in a study by Allcott & Gentzkow (2017). "Education, age, and total media consumption are strongly associated with more accurate beliefs about whether headlines are true or false" (Allcott & Gentzkow, 2017, p. 213). However, the level of education only had a significant effect ($p=.04$) on message credibility when comparing group 1 with no warning labels and group 2 with warning label 2 "This tweet is misleading."

This experimental study showed that unambiguous warning labels are more effective in reducing the credibility level of misleading posts than other warning labels (H3). As hypothesized, the most unambiguous warning label, "This tweet is misleading," significantly lowered the credibility level compared to the other two warning labels: "This tweet may be misleading" and "Find more information here." This is consistent with Ecker et al. (2010) and Lewandowsky et al. (2012), which found that the clearer a warning label, the more effective it will be.

Additionally, the study finds that message credibility acts as a mediator between warning labels and willingness to share (H4), adding to previous findings by Mena (2019) and Pennycook et al. (2020). Even though posts without warning labels attached to them were expected to be more credible than all othe other posts with one of the warning labels (H4a), based on the "implied truth" phenomenon indicated in previous studies (Pennycook et al., 2020), posts with a warning label "Find more information here" were found more credible compared to posts

without warning labels and posts with the other two warning labels. One reason for this may be the positive aspect of having the opportunity to read more on a specific topic or news that is being shared. While the addition of "negative" warning labels may increase credibility on posts that do not contain any labels (Pennycook et al., 2020), warning labels or tags that have a link where people can find out more information are found to increase the credibility level on such posts.

The current study also found support for (H4b) in that the more credible a post is perceived, the higher the chances it is going to be shared. Hypothesis H4b was found statistically significant, regardless of the type of warning label a post contains. Considering that the main effect of warning labels on willingness to share was found to be non-significant for all group comparisons, the impact of the two moderator variables 'Online trust' and 'Self-disclosure' on willingness to share was also found to be non-significant.

7.1 Limitations

Since the online survey was open to anyone 18 years or older and the study sample had participants from all over the world with different backgrounds, the findings of this study can be generalized in different cultural settings. However, the current research has three main limitations.

First, as mentioned earlier, the study sample was relatively homogeneous, with most respondents identifying as females (over 70%) and the majority of respondents having a university degree (around 90%). Even though a more extensive data sample was collected than expected, with a dataset consisting of 208 valid cases, the data sample is still small since this was a student study collecting data through an unpaid online survey. To better test the hypotheses of

this study, a larger and more diverse sample (in terms of demographics) would ensure results with high external validity.

Second, the majority of participants responded that they do not use Twitter. Less than 25% of the participants answered that they use Twitter, which can be another limitation since they may be unfamiliar with the format of Twitter and how Tweets and Re-tweets work. This may have influenced participants' willingness to share the posts shown to them, considering they do not use Twitter. Future studies testing warning labels based on the warning labels introduced by Twitter are recommended to collect a sample of users that already use Twitter and are familiar with this social media platform.

Third, testing willingness to share the posts shown to them through an online survey where people select how likely they are to share the tweets presented to them is not the same as testing how such posts are shared on Twitter and online in real-time. Moreover, people may not feel comfortable admitting which type of posts they are more willing to share, even in an anonymous setting such as this one. Future studies can tackle this limitation by conducting the experiment in a lab and observing which tweets are shared the most and which warning labels deter people from sharing the most. Scholars are recommended to address the above-mentioned limitations to ensure that these drawbacks will not be present in future investigations.

7.2 Suggestions for future studies

Despite increasing interest in testing the effect of warning labels on reducing misinformation, there has been a lack of qualitative research or studies that combine both a quantitative and qualitative approach. Therefore, future studies can combine a qualitative and quantitative approach by using an online survey experiment and guided interviews with a

considerable number of participants to provide a deeper understanding of the impact of warning labels introduced by social media platforms.

While the majority of previous studies have focused on testing Facebook warning labels (Mena, 2019; Pennycook et al., 2020; Buchanan & Benson, 2019) and a few recent studies have focused on testing Twitter warning labels (Starbird et al., 2014; Sanderson et al., 2021), new studies that compare the effectiveness of Facebook warning labels with Twitter warning labels and test which social media channel has a more significant impact on reducing false news and misinformation are needed, especially since warning labels and other labeling initiatives are expected to increase in the near future (Newman, 2019).

Since there was a slight indication in this study that people are more willing to share posts containing warning labels with their family and friends than sharing the posts on their social media feeds, supporting findings by Sundar et al. (2021), it is worth exploring this phenomenon more in-depth and finding out the reasons behind it. Moreover, past research has shown that while these new strategies implemented by social media channels may reduce the spread of fake news and misinformation, the news or the information people receive does not become harmless once it is debunked. Quite the contrary since misinformation effects are complex and they tend to persist even after correction" (Ardèvol-Abreu et al., 2020; Thorson, 2015). Therefore, it is essential to test people's willingness to still share and interact with Facebook and Twitter posts that contain warning labels and further explore the reasons behind this controversial phenomenon.

Other studies also suggest that "some people may intentionally share made up news on social media, although their motivations are not fully explained" (Ardèvol-Abreu et al., 2020). That is why it is also essential to shed light on why this happens and find ways how social media

platforms such as Facebook and Twitter, but not only, can address this issue and take additional measures to prevent it from happening.

7.3 Conclusion

This experimental study contributed to research on testing the effect of warning labels on reducing the spread of misinformation. It showed that not all warning labels have the same deterring effect when it comes to reducing the spread of misinformation. The study indicates that unambiguous warning labels are the most effective warning labels for reducing message credibility and preventing the spread of misinformation.

8. References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211-236. https://doi.org/10.1257/jep.31.2.211

An, J., Quercia, D., Cha, M., Gummadi, K., & Crowcroft, J. (2014). Sharing political news: The balancing act of intimacy and socialization in selective exposure. *EPJ Data Science, 3*(1). doi:10.1140/epjds/s13688-014-0012-2

Apuke, O. D., & Omar, B. (2020). Modelling the antecedent factors that affect online fake news sharing on COVID-19: The moderating role of fake news knowledge. *Health Education Research, 35*(5), 490-503. doi:10.1093/her/cyaa030

Ardèvol-Abreu, A., Delponti, P., & Rodríguez-Wangüemert, C. (2020). Intentional or inadvertent fake news sharing? Fact-checking warnings and users' interaction with social media content. *Profesional De La Información*, *29*(5). doi:10.3145/epi.2020.sep.07

Brehm, S. S., & Brehm, J. W. (1981). Introduction: Freedom, Control, and Reactance Theory. *Psychological Reactance,* 1-7. doi:10.1016/b978-0-12-129840-1.50005-3

Buchanan, T., & Benson, V. (2019). Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"? *Social Media Society, 5*(4), 205630511988865. doi:10.1177/2056305119888654

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web - WWW 11*. doi:10.1145/1963405.1963500

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., . . . Nyhan, B. (2019).

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. doi:10.1007/s11109-019-09533-0

Coronavirus disease (COVID-19). (n.d.). Retrieved June 21, 2021, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Courtney, I. (2020, May 26). Misinformation in the time of COVID-19 - A review of three recent infodemic themed webinars. Retrieved from https://www.medialiteracyireland.ie/news/misinformation-in-the-time-of-covid-19-a-review-of-three-recent-infodemic-themed-webinars

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition, 38*(8), 1087-1100. doi:10.3758/mc.38.8.1087

Egelhofer, J., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals Of The International Communication Association*, *43*(2), 97-116. https://doi.org/10.1080/23808985.2019.1602782

Gao, M., Xiao, Z., Karahalios, K., & Fu, W. (2018). To Label or Not to Label. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 1-16. doi:10.1145/3274324

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Guo, C., & Saxton, G. (2013). Tweeting Social Change. *Nonprofit And Voluntary Sector Quarterly*, *43*(1), 57-79. https://doi.org/10.1177/0899764012471585

Hamdi, T., Slimi, H., Bounhas, I., & Slimani, Y. (2019). A Hybrid Approach for Fake News

Detection in Twitter Based on User Features and Graph Embedding. *Distributed Computing and Internet Technology Lecture Notes in Computer Science,* 266-280. doi:10.1007/978-3-030-36987-3_17

Hayes, A. F. (2021). The PROCESS macro for SPSS, SAS, and R, Version 4.0. Retrieved from https://www.processmacro.org/index.html

Heath, A. (2016, December 15). Facebook is going to use Snopes and other fact-checkers to combat and bury 'fake news'. Retrieved from https://www.businessinsider.com/facebook-will-fact-check-label-fake-news-in-news-feed-2016-12?r=DE&IR=T

Huynh TLD. (2020) "The COVID-19 risk perception: A survey on socioeconomics and media attention", Economics Bulletin. (1st ed., Vol. 40, pp. 758-764).

IBM Corp. (2020). IBM SPSS Statistics for Macintosh, Version 27.0. Armonk, NY: IBM Corp

Kellogg, K. (2020, February 05). The 7 Biggest Social Media Sites in 2020. Retrieved from https://www.searchenginejournal.com/social-media/biggest-social-media-sites/#close

Koch, T., Frischlich, L., & Lermer, E. (2021). The Effects of Warning Labels and Social Endorsement Cues on Credibility Perceptions of and Engagement Intentions with Fake News. doi:10.31234/osf.io/fw3zq

Kraus, R. (2021, June 11). Facebook has put warning labels on 180 million posts since March. Retrieved from https://mashable.com/article/facebook-labels-180-million-posts-false

Kurtzleben, D. (2020). *NPR Cookie Consent and Choices*. Npr.org. Retrieved from https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election

Lee, J. (2020). The effect of web add-on correction and narrative correction on belief in misinformation depending on motivations for using social media. *Behaviour & Information Technology*, 1-15. https://doi.org/10.1080/0144929x.2020.1829708

Leiner, D. J. (2021). SoSci Survey (Version 3.2.38). Retrieved from https://www.soscisurvey.de

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction. Psychological Science in the Public Interest, 13(3), 106-131. doi:10.1177/1529100612451018

Li, R., & Suh, A. (2015). Factors Influencing Information credibility on Social Media Platforms: Evidence from Facebook Pages. *Procedia Computer Science, 72*, 314-328. doi:10.1016/j.procs.2015.12.146

Lyons, T. (2018, October 19). New Research Shows Facebook Making Strides Against False News. Retrieved from https://about.fb.com/news/2018/10/inside-feed-michigan-lemonde/

MacKinnon, R. (2012). Consent of the Networked: The Worldwide Struggle for Internet Freedom. Retrieved from https://cyber.harvard.edu/events/2012/02/mackinnon

MacKinnon, R. (2013). *Consent of the networked: The worldwide struggle for Internet freedom*. New York: Basic Books.

Mantzarlis, A. (2017). Moving fake news research out of the lab. Retrieved from https://www.niemanlab.org/2017/12/moving-fake-news-research-out-of-the-lab/

Mena, P. (2019). Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet, 12*(2), 165-183. doi:10.1002/poi3.214

Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign

disinformation. *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-45

Nelson, J., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720-3737. https://doi.org/10.1177/1461444818758715

Newman, N. (2019). Journalism, Media and Technology Trends and Predictions 2019. Retrieved from https://reutersinstitute.politics.ox.ac.uk/our-research/journalism-media-and-technology-trends-and-predictions-2019

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior, 32*(2), 303-330. doi:10.1007/s11109-010-9112-2

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective Messages in Vaccine Promotion: A Randomized Trial. *Pediatrics, 133*(4). doi:10.1542/peds.2013-2365

Oremus, W. (2019, January 25). These Startups Want to Protect You From Fake News. Can You Trust Them? Retrieved from https://slate.com/technology/2019/01/newsguard-nuzzelrank-media-ratings-fake-news.html

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science, 66*(11), 4944-4957. doi:10.1287/mnsc.2019.3478

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived

accuracy of fake news. *Journal of Experimental Psychology: General, 147*(12), 1865-1880. doi:10.1037/xge0000465

Pennycook, G., Mcphetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. doi:10.31234/osf.io/uhbk9

Persily, N., & Tucker, J. A. (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge: Cambridge University Press. doi:10.1017/9781108890960

Reddy, C. A., Gadiraju, S. N., & Nagaraj, D. S. (2021). Detecting Fake News Tweets from Twitter. *Journal of University of Shanghai for Science and Technology, 23*(08), 532-537. doi:10.51201/jusst/21/08428

Rosen, G., Harbath, K., Gleicher, N., & Leathern, R. (2019, October 21). Helping to Protect the 2020 US Elections. Retrieved from https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/

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

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*. doi:10.37016/mr-2020-77

Sardarizadeh, S. (2020, May 12). Coronavirus: Twitter will label Covid-19 fake news. Retrieved from https://www.bbc.com/news/technology-52632909

Schul, Y. (1993). When Warning Succeeds: The Effect of Warning on Success in Ignoring Invalid Information. *Journal of Experimental Social Psychology, 29*(1), 42-62. doi:10.1006/jesp.1993.1003

Scott, J., & Maryman, J. (2016). Using Social Media as a Tool to Complement Advocacy. *Global Journal Of Community Psychology Practice*, *7*(1). https://doi.org/10.7728/0701201603

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2021, April 01). Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes. Retrieved from https://arxiv.org/abs/2104.00779

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *IConference 2014 Proceedings,* 654-662. doi:10.9776/14308

Stelter, B. (2016, December 15). Facebook to start putting warning labels on 'fake news'. Retrieved from https://money.cnn.com/2016/12/15/media/facebook-fake-news-warning-labels/

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? *Journal of Computer-Mediated Communication*. doi:10.1093/jcmc/zmab010

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services, 51*, 72-82. doi:10.1016/j.jretconser.2019.05.026

The COVID-19 infodemic. (2020). *The Lancet Infectious Diseases, 20*(8), 875. doi:10.1016/s1473-3099(20)30565-x

Thorson, E. (2015). Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication*, *33*(3), 460-480. https://doi.org/10.1080/10584609.2015.1102187

Tweet Generator: Tweetgen. (2018). Retrieved from https://www.tweetgen.com/create/tweet.html

Wang, Y., Mckee, M., Torbica, A., & Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine, 240*, 112552. doi:10.1016/j.socscimed.2019.112552

Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media – The pilot quantitative study. *Health Policy and Technology, 7*(2), 115-118. doi:10.1016/j.hlpt.2018.03.002

Weller, K. (2016). Trying to understand social media users and usage. *Online Information Review, 40*(2), 256-264. doi:10.1108/oir-09-2015-0299

Wells, G. (2020, November 12). Twitter Says Labels and Warnings Slowed Spread of False Election Claims. Retrieved from https://www.wsj.com/articles/twitter-says-labels-and-warnings-slowed-spread-of-false-election-claims-11605214925

Wood, T., & Porter, E. (2018). The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior, 41*(1), 135-163. doi:10.1007/s11109-018-9443-y

Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2017). Fact-Checking Effectiveness as a Function of Format and Tone: Evaluating FactCheck.org and FlackCheck.org. *Journalism & Mass Communication Quarterly, 95*(1), 49-75. doi:10.1177/1077699017710453

9. Appendices

9.1 Appendix A

Scales

Table 3
Study measures and items factor loadings of measurement model.

Study Measure (Reference)	Measurement items	Factor loadings
Authenticating news before sharing[a]	I ask my friends to check the authenticity of any message before sharing	0.82
	I ask my family/relatives to check the authenticity of any message before sharing	0.81
	I rely on TV news channels to check the authenticity of any message before sharing it	0.63
Online trust (Fang et al., 2016)	I trust the information that is shared on WA	0.78
	I trust the news that is shared on WA	0.77
Self-disclosure (Krasnova et al., 2009)	I reveal a lot of information about me on WA	0.77
	My WA profile tells a lot about me	0.83
	I have a detailed profile on WA	0.71
Social comparison (Cramer et al., 2016)	I feel less motivated to use WA to avoid comparing myself to others	0.88
	I feel less motivated to use WA as I compare myself to others through WA use	0.81
Fear of missing out (FoMO) (Przybylski et al., 2013)	I fear others have more rewarding experiences than me	0.78
	I fear my friends have more rewarding experiences than me	0.71
Social media fatigue (SMF) (Karasek, 1979)	Due to WA use, I feel rather exhausted	0.76
	After using WA, it takes effort to concentrate in my spare time	0.83
	During WA use, I often feel too fatigued to perform other tasks well	0.83
Sharing fake news online[a]	I often share fake news because I don't have time to check its authenticity	0.87
	I share fake news because I don't have time to check facts through trusted sources	0.87

Note. WhatsApp = WA.
[a] Measures developed based on qualitative study.

9.2 Appendix B

Original tweets

The Third Pole @third_pole · Jul 17

In 1838, German biologist Johann Jakob Heckel documented the presence of 16 fish species in Kashmir.

Only five of them can be found now.

New riverbed mining permissions may be final straw for Kashmir's fish

Stone crushers are springing up along rivers in Jammu and Kashmir, threatening the habitat of both native and introduced fish species

thethirdpole.net

9.3 Appendix C

Survey questions

In the following, we will ask you some questions about your media use. Please choose one of the alternatives. If you do not know the answer, select the one that seems most plausible to you.

How often do you use Social Media?

- Every day
- Every two days
- Several times per week
- Several times per month
- I do not use Social Media

Which Social Media channel or channels do you use the most?

Note: You can select multiple options if needed.

- Facebook
- Twitter
- LinkedIn
- Instagram
- YouTube
- Other

How frequently do you post on your preferred Social Media channel(s)?

- Once per day
- Twice per day
- More than twice every day
- Several times in a week, but less than every day
- Several times a month
- Less than once a month
- Never

In the following, we will ask you some questions about yourself. Please select one of the alternatives for each question.

How would you describe your gender?

- Woman
- Man
- Non-Binary
- Other
- Prefer not to say

Please select your age range.

- 18 – 22
- 23 – 27
- 28 – 32
- 33 – 37
- 38-42
- 43-47
- 48 or above
- Prefer not to say

What is your highest level of education?

- High school or equivalent
- Technical or occupational certificate
- Associate degree
- Some college coursework completed
- Bachelor’s degree
- Master’s degree
- Doctorate
- Professional
- Other

Page 06

PHP code

```
if (value('RG01') == 1) {
  goToPage('G1 INTRO');
} else if (value('RG01') == 2) {
  goToPage('G2 INTRO');
} else if (value('RG01') == 3) {
  goToPage('G3 INTRO');
} else if (value('RG01') == 4) {
  goToPage('G4 INTRO');
}
```

Page 07

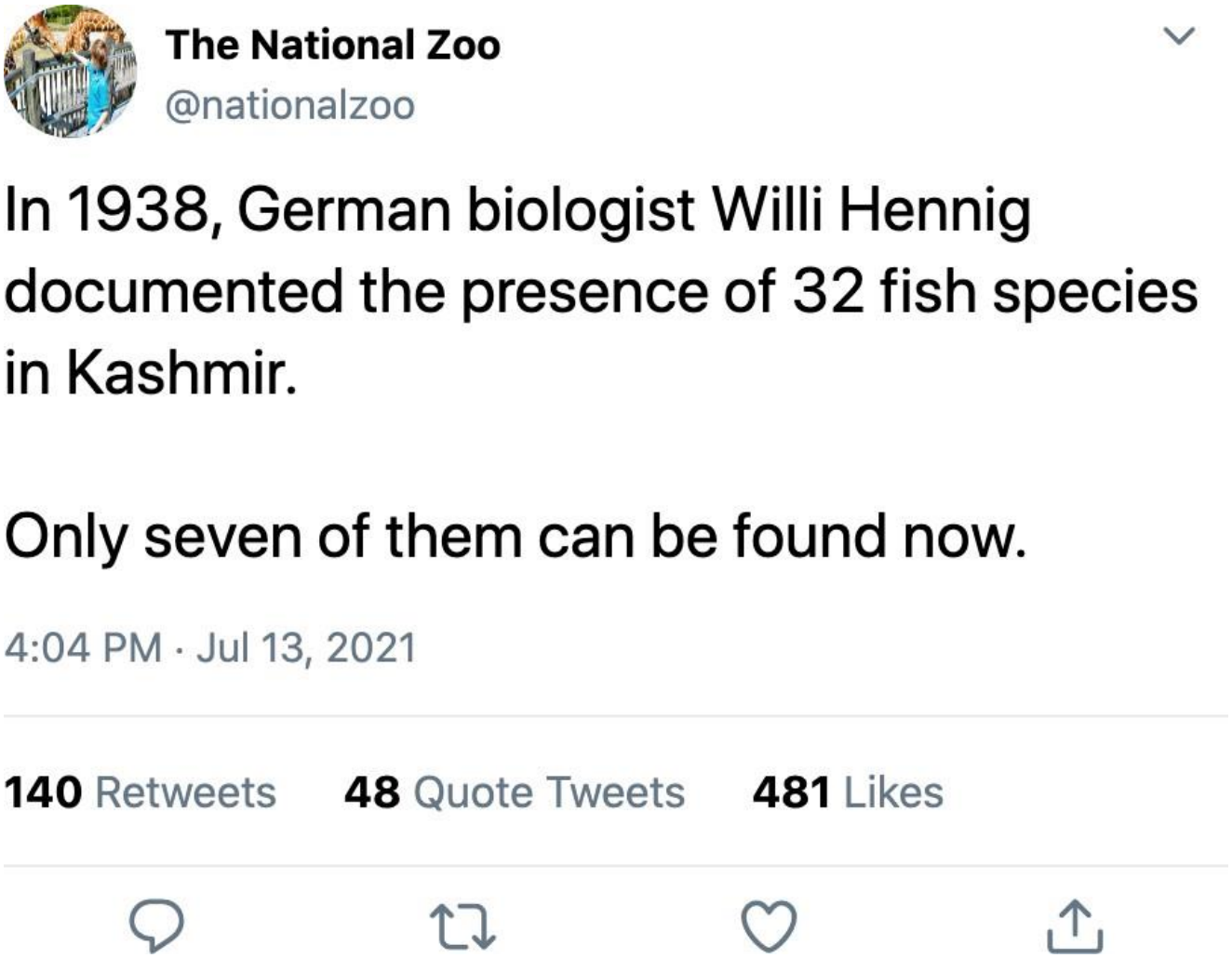
G1 INTRO

In the following section, you will find the screenshots of three tweets. After reading them, please answer the follow-up questions for each tweet.

PHP code

```
if (!isset($pages)) {
  $pages = array('G101', 'G102', 'G103');
  shuffle($pages);
  $pages[] = 'WL';
  registerVariable($pages);
}
setPageOrder($pages);
```

Please read the following tweet carefully, and then answer the three follow-up questions.



How would you evaluate the credibility of this tweet?

Not Credible at all — Very Credible

-3 -2 -1 0 +1 +2 +3

How likely are you to retweet this post or share it on your social media feed?

Not likely at all — Very likely

-3 -2 -1 0 +1 +2 +3

How likely are you to share this tweet with a colleague, a friend, or a family member?

Not likely at all — Very likely

-3 -2 -1 0 +1 +2 +3

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all -3 -2 -1 0 +1 +2 +3 Very Credible

How likely are you to retweet this post or share it on your social media feed?

Not likely at all -3 -2 -1 0 +1 +2 +3 Very likely

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all -3 -2 -1 0 +1 +2 +3 Very Credible

How likely are you to retweet this post or share it on your social media feed?

Not likely at all -3 -2 -1 0 +1 +2 +3 Very likely

How likely are you to share this tweet with a colleague, a friend, or a family member?

In the following section, you will find the screenshots of three tweets. After reading them, please answer the follow-up questions for each tweet.

PHP code

```
if (!isset($pages)) {
  $pages = array('G201', 'G202', 'G203');
  shuffle($pages);
  $pages[] = 'WL';
  registerVariable($pages);
}
setPageOrder($pages);
```

Please read the following tweet carefully, and then answer the three follow-up questions.

Not Credible at all — Very Credible

-3 -2 -1 0 +1 +2 +3

How would you evaluate the credibility of this tweet?

	Not likely at all						Very likely
	-3	-2	-1	0	+1	+2	+3

How likely are you to retweet this post or share it on your social media feed?

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all						Very Credible
-3	-2	-1	0	+1	+2	+3

How likely are you to retweet this post or share it on your social media feed?

Not likely at all						Very likely
-3	-2	-1	0	+1	+2	+3

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all -3 -2 -1 0 +1 +2 +3 Very Credible

How likely are you to retweet this post or share it on your social media feed?

Not likely at all -3 -2 -1 0 +1 +2 +3 Very likely

How likely are you to share this tweet with a colleague, a friend, or a family member?

In the following section, you will find the screenshots of three tweets. After reading them, please answer the follow-up questions for each tweet.

PHP code

```
if (!isset($pages)) {
  $pages = array('G301', 'G302', 'G303');
  shuffle($pages);
  $pages[] = 'WL';
  registerVariable($pages);
}
setPageOrder($pages);
```

Please read the following tweet carefully, and then answer the three follow-up questions.

Not Credible at all — Very Credible

-3 -2 -1 0 +1 +2 +3

How would you evaluate the credibility of this tweet?

	Not likely at all						Very likely
	-3	-2	-1	0	+1	+2	+3

How likely are you to retweet this post or share it on your social media feed?

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all -3 -2 -1 0 +1 +2 +3 Very Credible

How likely are you to retweet this post or share it on your social media feed?

Not likely at all -3 -2 -1 0 +1 +2 +3 Very likely

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all -3 -2 -1 0 +1 +2 +3 Very Credible

How likely are you to retweet this post or share it on your social media feed?

Not likely at all -3 -2 -1 0 +1 +2 +3 Very likely

How likely are you to share this tweet with a colleague, a friend, or a family member?

In the following section, you will find the screenshots of three tweets. After reading them, please answer the follow-up questions for each tweet.

PHP code

```
if (!isset($pages)) {
  $pages = array('G401', 'G402', 'G403');
  shuffle($pages);
  $pages[] = 'WL';
  registerVariable($pages);
}
setPageOrder($pages);
```

Please read the following tweet carefully, and then answer the three follow-up questions.

Not Credible at all — Very Credible

-3 -2 -1 0 +1 +2 +3

How would you evaluate the credibility of this tweet?

	Not likely at all						Very likely
	-3	-2	-1	0	+1	+2	+3

How likely are you to retweet this post or share it on your social media feed?

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all						Very Credible
-3	-2	-1	0	+1	+2	+3

How likely are you to retweet this post or share it on your social media feed?

Not likely at all						Very likely
-3	-2	-1	0	+1	+2	+3

How likely are you to share this tweet with a colleague, a friend, or a family member?

Please read the following tweet carefully, and then answer the three follow-up questions.

How would you evaluate the credibility of this tweet?

Not Credible at all -3 -2 -1 0 +1 +2 +3 Very Credible

How likely are you to retweet this post or share it on your social media feed?

Not likely at all -3 -2 -1 0 +1 +2 +3 Very likely

How likely are you to share this tweet with a colleague, a friend, or a family member?

Page 23

WL

Did you notice a warning label attached to any of the tweets we showed you?

Yes

No

I am not sure

Page 24

WL2

If applicable, please choose from -3 to +3.

Not helpful at all — Very helpful

-3 -2 -1 0 +1 +2 +3

I didn't see a warning label

In your opinion, how helpful was the warning label?

If applicable, please choose from -3 to +3.

Not informative at all — Very informative

-3 -2 -1 0 +1 +2 +3

I didn't see a warning label

In your opinion, how informative was the warning label?

Page 25

FND

Do you think any of the tweets we showed you contain misleading or false information?

Yes

No

I am not sure

Abstract

From threatening democracy and peace to risking people's health, the rapid spread of false news and misinformation has become one of the greatest challenges of the last decade. Strict measures from regulatory entities and, in particular, from social media giants such as Facebook and Twitter, which have been singled out as "incubators of fake news and propaganda" (Persily & Tucker, 2020), are necessary to stop or prevent the spread of false news and misinformation. The use of warning labels is one of the latest strategies introduced by social media platforms such as Facebook and Twitter to counter the spread of misinformation. However, research on testing which types of warning labels are more effective is still limited.

This study evaluates the effectiveness of Twitter warning labels on message credibility and willingness to share and tests which types of warning labels are more effective in preventing the spread of misinformation. Data from an online survey experiment (N = 208) indicate that unambiguous warning labels such as "This tweet is misleading" significantly lowered the credibility level of misleading posts, compared to other warning labels such as "This tweet may be misleading" or "Find more information here."

The study findings indicate that message credibility mediates the effect of warning labels on willingness to share and that posts without warning labels are perceived as more credible compared to posts with warning labels, but only when the comparison is with an unambiguous warning label. In addition, the study finds that the more credible a post is perceived (no matter if it has a warning label attached or not), the more it will be shared. Online trust and social media self-disclosure did not affect willingness to share. The findings of this experimental study contribute to research on finding effective strategies to prevent misleading posts from being shared or going viral.

Abstrakt

Von der Bedrohung von Demokratie und Frieden bis hin zur Gefährdung der Gesundheit der Menschen ist die rasche Verbreitung von Falschmeldungen und Fehlinformationen zu einer der größten Herausforderungen des letzten Jahrzehnts geworden. Strenge Maßnahmen von Regulierungsbehörden und insbesondere von Social-Media-Giganten wie Facebook und Twitter, die als „Inkubatoren von Fake News und Propaganda“ (Persily & Tucker, 2020) ausgezeichnet wurden, sind notwendig, um die Verbreitung von Falschmeldungen und Fehlinformationen zu stoppen oder zu verhindern. Warnhinweise sind eine der neuesten Strategien von Social-Media-Plattformen wie Facebook und Twitter, um der Verbreitung von Fehlinformationen entgegenzuwirken. Allerdings ist die Forschung bezüglich Tests, welche Arten von Warnhinweisen wirksamer sind, noch begrenzt.

Diese Studie bewertet die Wirksamkeit von Twitter-Warnhinweisen aufgrund der Glaubwürdigkeit und der Bereitschaft zum Teilen von Nachrichten und testet, welche Arten von Warnhinweisen wirksamer sind, um die Verbreitung von Fehlinformationen zu verhindern. Daten aus einem Online-Umfrageexperiment (N = 208) zeigen, dass eindeutige Warnhinweise wie „Dieser Tweet ist irreführend“ die Glaubwürdigkeit irreführender Beiträge im Vergleich zu anderen Warnhinweisen wie „Dieser Tweet ist möglicherweise irreführend“ oder „Finden Sie mehr Infos hier” deutlich senkten.

Die Studienergebnisse zeigen, dass die Glaubwürdigkeit von Botschaften die Wirkung von Warnhinweisen auf die Bereitschaft zum Teilen vermittelt und Beiträge ohne Warnhinweise im Vergleich zu Beiträgen mit Warnhinweisen als glaubwürdiger wahrgenommen werden, jedoch nur wenn sie im Vergleich zu einem eindeutigen Warnhinweis stehen. Darüber hinaus stellt die Studie fest, dass je glaubwürdiger ein Beitrag wahrgenommen wird (egal ob mit einem

Warnhinweis versehen oder nicht), desto mehr wird er geteilt. Online-Vertrauen und Selbstauskunft in den sozialen Medien hatten keinen Einfluss auf die Bereitschaft zum Teilen. Die Ergebnisse dieser experimentellen Studie tragen zur Erforschung effektiver Strategien bei, um zu verhindern, dass irreführende Beiträge geteilt oder viral werden.