Does Being Verified Make You More Credible? Account Verification's Effect on Tweet Credibility

Tavish Vaidya
Georgetown University

Daniel Votipka
University of Maryland

Michelle L. Mazurek
University of Maryland

Micah Sherr
Georgetown University

ABSTRACT

Many popular social networking and microblogging sites support *verified accounts*—user accounts that are deemed of public interest and whose owners have been authenticated by the site. Importantly, the *content* of messages contributed by verified account owners is not verified. Such messages may be factually correct, or not.

This paper investigates whether users confuse authenticity with credibility by posing the question: *Are users more likely to believe content from verified accounts than from non-verified accounts?* We conduct two online studies, a year apart, with 748 and 2041 participants respectively, to assess how the presence or absence of verified account indicators influences users' perceptions of tweets. Surprisingly, across both studies, we find that—in the context of unfamiliar accounts—most users can effectively distinguish between authenticity and credibility. The presence or absence of an authenticity indicator has no significant effect on willingness to share a tweet or take action based on its contents.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Trust**; • **Security and privacy** → *Authentication*; *Human and societal aspects of security and privacy*.

KEYWORDS

Social; Twitter; Credibility; Trust; Authentication

ACM Reference Format:
Tavish Vaidya, Daniel Votipka, Michelle L. Mazurek, and Micah Sherr. 2019. Does Being Verified Make You More Credible? Account Verification's Effect on Tweet Credibility. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK*. ACM, New York, NY, USA, 13 pages. **https://doi.org/10.1145/3290605.3300755**

1 INTRODUCTION

On many social networking and microblogging sites, it is no longer the case that nobody knows you're a dog [43]. Twitter, Instagram, Reddit, Facebook, and other popular sites support *verified accounts*—accounts whose owners have been authenticated by the site [48].[1] Verified accounts are often reserved for well-known organizations and individuals, and are indicated with a badge (e.g., ✓) that appears next to the account holder's name (see Figure 1a). Verified accounts are intended to allow users to easily distinguish between authentic accounts of public interest and those belonging to parody accounts or impostors.

Importantly, accounts earn verified status by being authenticated (i.e., the user is who she claims to be). Sites specifically do not assess an account owner's trustworthiness when assigning verified status, and most major online community sites do not require that verified account owners post only factual content. More plainly, posts from a verified account may be factual, or not [11].

This paper examines whether users conflate authenticity with credibility, in the context of microblogging sites such as Twitter. That is, are users more likely to believe, and therefore be willing to act upon, messages that originate from verified accounts? We define credibility as the "believability" of a piece of information—that is, the degree to which it influences whether an individual is persuaded to adopt the opinion offered and take action accordingly [24]. Indicators of authenticity (such as a verified account badge) signal that a message originated from its purported source. Credible information may of course stem from both verified and unverified accounts; conversely, the same is true for incorrect information. Equating credibility with authenticity is thus a logical fallacy.

CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300755

[1] The terminology can vary slightly among sites; for example, Facebook uses the term "verified profile." For consistency, we refer to accounts whose authenticity has been confirmed by site operators as *verified accounts*.

Three significant factors suggest that verified badges could influence people's perceptions of user-generated content. Literature from psychology and other social sciences shows that perceptions of credibility are more influenced by the source of information than by its credulity [6, 14, 17, 21, 45]. Notably, Hovland et al. found that the credibility of a message is influenced by the source, setting, and presentation of arguments of the message along with receivers' predisposition toward the message [24, pg. 10-13]. Berlo et al. found that the message's source affects message credibility along three axes: safety, or whether the recipient believes the source might have an agenda or a reason to mislead; qualification, or how qualified the source is to comment on the given topic; and dynamism, or how charismatic and persuasive the author is [3]. If information is widely endorsed (e.g., popularly "liked" on Twitter) and/or originates from a well-known source (e.g., a verified account), then the prior work suggests users will perceive it as more trustworthy (i.e., safe) and the author as more qualified, leading to higher perceived credibility.

Second, the computer security community has repeatedly shown that users have difficulty in understanding the concept of authenticity [1, 8, 13, 46] and are generally confused by security indicators [20, 40]. In particular, users do not easily distinguish between authenticity and integrity. For example, users have been shown to believe the content of websites so long as browser indicators (e.g., the golden lock) show that the website is secure [13]. This suggests that users might mistake verified account badges as an external validation that the posted content has been fact-checked. This effect could potentially be boosted by the iconography of verified account badges (e.g., Twitter's ✓ and Facebook's ✓).

Finally, the manner in which accounts become verified has led to confusion. In particular, Twitter verifies accounts *only when it deems the account owner to be of public interest* [48]. This has translated into a perception that authenticity indicators are markers of prestige and importance, not merely authenticity. Twitter has been criticized for granting verified account status to racists and other extremists [52]. (Here, the criticism is not that these extremists are not who they say they are, but rather that their accounts have earned "elite" status.) Twitter responded to this criticism by noting that "[v]erification was meant to authenticate identity & voice but is interpreted as an endorsement or an indicator of importance. We recognize that we have created this confusion and need to resolve it" [50]. However, we posit that Twitter's solution—to disable previous functionality that allowed anyone to apply for verified status and to enforce a code of conduct—could further exacerbate the misperception that the verified account owners (and their posted content) should be considered trustworthy.[2]

The implications of users conflating authenticity with credibility are worrisome. With the growing fraction of people who use online social media (with user-generated content) as their primary source of news [2] and the recent proliferation of fake news [54], a possible inability to judge the credibility of information from (un)verified accounts is especially disconcerting.

To understand the effect of account verification on the credibility of user-generated content, we conducted two studies on Amazon's Mechanical Turk comparing participants' perceived credibility of tweets from verified and unverified accounts. In both the studies, we varied factors such as the subject matter of tweets, the advertised credentials of the poster, the positivity/negativity of tweets towards their subject matter, and the indicators used to signify that an account was verified (e.g., a ✓ icon vs. a textual "Verified account" label). We measured perceived credibility both implicitly, by asking about actionability, and explicitly, by asking directly about the effect of the authenticity indicator.

Surprisingly, the results of Study 1 (n1=748, conducted in Aug-Sep 2017) showed no evidence of correlation between verified account status (i.e., verified or unverified) and the perceived credibility of tweets. To validate the null result, we conducted a second study (Study 2, n2=2041, Aug-Sept 2018) that confirmed our initial findings from Study 1. In this paper, for brevity, we primarily discuss the second study. (Results from the first study are provided as supplementary material.) As detailed in the next section, these findings directly contradict results from prior work [35, 51].

2 RELATED WORK

Researchers have long sought to understand the factors that influence users' perceptions of online content. Wineburg et al. [53] survey students to assess whether they can correctly judge the credibility of information from online sources. Fogg et al. [15] conduct a notable, large-scale study in which more than 2600 participants reviewed websites for credibility. They find that the "design look" (i.e., user interface) of a site is the most frequent (and dominating) factor in determining its credibility. However, neither Wineburg et al. or Fogg et al. study the impact of authenticity indicators (e.g., verified account badges) and Fogg et al. do not consider microblogging sites. In this paper, we focus on Twitter—a microblogging site in which all posts have fixed formatting and appearance—and investigate the effect of authenticity indicators on perceived credibility.

[2] We note that Twitter's statement that recognized the public's confusion over the meaning of verified accounts [50], as well as their disabling of verified account request functionality, occurred between our two studies.

Perhaps most similar to our work, Morris et al. [35] examine how features of tweets affect user perception. They find that the "verification seal" has a high impact in perceived credibility. However, their results are based on a questionnaire in which participants are directly asked which features they consider when deciding whether a tweet is credible; thus, they measure the *conscious* impact of verified account badges. Additionally, this question is posed generically (i.e., not related to a specific tweet). Morris et al. did not include the verification seal as a feature in their controlled experiment. Therefore, they only measured whether participants believe the authenticity indicator is important in the abstract, but do not test the behavioral changes we investigated. While their work serves as a motivation for our own, we examine both the conscious and subconscious effects of verified account badges. We additionally explore whether users understand the notion of authenticity and the degree to which users conflate authenticity and credibility.

Existing work has also looked at the related problem of identifying accurate (and conversely, inaccurate) information online. Castillo et al. attempt to automatically classify tweets as credible or not-credible based on their features [5]. They use Amazon Mechanical Turk to establish a ground truth for their machine learning classifier. Similarly, Qazvinian et al. use statistical models to identify tweets with false information [37]. Maddock et al. develop multi-dimensional signatures to detect false online rumors [32]. Kumar and Geethakumari apply techniques from cognitive psychology to develop a filtering algorithm that measures the credibility of online information [28]. These approaches could more accurately be described as automating the process of identifying which tweets are perceived by humans as being credible. In contrast, our work studies a different phenomenon—*how* users form their perceptions of credibility.

The Science of Perceived Credibility

The mechanisms by which people assess the credibility of online information have received considerable attention from social science researchers. A common finding is that users often form opinions about the veracity of information based not on its credulity, but rather on their familiarity with its source. In particular, Gigerenzer et al. find that, to reduce cognitive strain, people tend to believe information that stems from familiar sources [17]. Similarly, Hilligoss and Rieh find that the perception of whether the information provider is deemed "official" significantly affects users' beliefs in the information's legitimacy [21]. Metzger et al. synthesize these findings into a *reputation heuristic*, and further show that users assess the credibility of web sites based on the site's reputation or name recognition rather than on its content [34].

Credibility is not established in a vacuum. A number of studies [6, 14, 21, 45] have found that users are also more likely to perceive information as credible when they believe that *others* perceive it to be accurate. This effect is most commonly referred to as the *endorsement heuristic*.

Although the mechanisms for establishing perceptions about information's credibility are nuanced and complex (see Metzger et al.'s survey for a good overview [34]), the reputation and endorsement heuristics together provide a basis for reasoning about how users might (mis)understand verified accounts.

Typically, a verified account indicates that the site considers the account to be authentic [48]. This is an indicator that the account is official, and thus more likely to be trusted according to the reputation heuristic [21].

A verified marker also signals that the verified account has a positive reputation. Indeed, demonstrating that the account is of "public interest" is a requirement of earning the lauded verified account status on Twitter [48, 49]. Verified accounts are reported to have, on average, far more followers than non-verified accounts [25] and thus tweets from these accounts have a greater opportunity of being retweeted[3] or liked—both indicators that a tweet has been widely endorsed.

In summary, the reputation and endorsement heuristics suggest that users would assign greater validity to tweets from verified accounts, since they are both more official looking and have greater public following.

It should be emphasized that the pairing of credibility and authenticity is a psychological phenomenon, intended to inform human perception based on limited information and requiring little cognitive load. In actuality, however, credibility and authenticity are two independent concepts. Inaccurate information can stem from authentic, verified accounts. A primary goal of this paper is to determine whether users can differentiate authenticity from credibility, or whether (and why) users tend to conflate the two.

3 METHOD

We conducted two online studies to investigate how users interpret information from verified accounts on Twitter. Our goal is to understand how the presence or absence of authenticity indicators affect the perceived credibility of tweeted content, specifically the following hypotheses:

H1. Users notice the presence or absence of Twitter's authenticity indicators on posts.

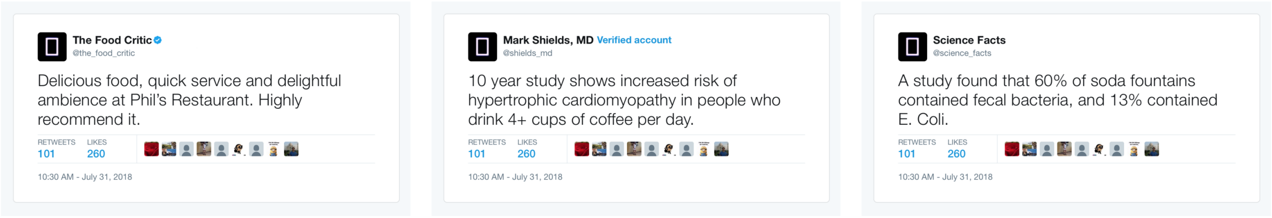
H2. Users are more likely to find tweets with authenticity indicators credible than those from unverified accounts, when measured implicitly.

H3. Users will explicitly *cite presence (or absence) of an authenticity indicator as an important influence on their decision about whether a tweet is credible.*

[3] Though, as many Twitter users note in their bios, not all retweets are endorsements.

Content Type	Tweet Text	Username	Display Name
Restaurant	Delicious food, quick service and delightful ambience at Phil's Restaurant. Highly recommend it. Average food, slow service and unpleasant ambience at Phil's Restaurant. Highly recommend staying away.	@the_food_critic	The Food Critic
Coffee	10 year study shows increased risk of hypertrophic cardiomyopathy in people who drink 4+ cups of coffee per day.	@shields_md @mark_shields	Mark Shields, MD Mark Shields
Grocery	Experts say this year's increase in cattle disease will cause 12% jump in average household's grocery bill.	@NAGrocers @the_grocery_couple	NorthAmericanGrocersAssociation The Grocery Couple
Soda	A study found that 60% of soda fountains contained fecal bacteria, and 13% contained E. Coli.	@science_facts	Science Facts

Table 1: Summary of conditions tested in Study 2. All 7 conditions were coupled with 3 types of authenticity indicators resulting in 21 total conditions, created from a full-factorial combination of authenticity indicators and content types.



The Food Critic ✓
@the_food_critic
Delicious food, quick service and delightful ambience at Phil's Restaurant. Highly recommend it.
RETWEETS 101 LIKES 260
10:30 AM - July 31, 2018

Mark Shields, MD Verified account
@shields_md
10 year study shows increased risk of hypertrophic cardiomyopathy in people who drink 4+ cups of coffee per day.
RETWEETS 101 LIKES 260
10:30 AM - July 31, 2018

Science Facts
@science_facts
A study found that 60% of soda fountains contained fecal bacteria, and 13% contained E. Coli.
RETWEETS 101 LIKES 260
10:30 AM - July 31, 2018

(a) (b) (c)

Figure 1: Example tweets from Study 2 in different conditions: (a) the *Restaurant* tweet with positive valence with *Check* indicator; (b) *Coffee* tweet with "MD" in the display name and *Text* authenticity indicator; (c) *Soda* tweet with no indicator.

H4. Users' understanding of Twitter verified accounts improves when presented with Twitter's description of verified accounts.

Hypothesis *H1* tests whether users notice the authenticity indicators that accompany tweets from verified accounts. This is a prerequisite for forming opinions about the credibility of a tweet based (at least in part) on the verified status of the tweet's author—the subject of hypothesis *H2*. *H2* is designed to test the *subconscious* effect of authenticity indicators on perceived credibility, and as described in the next section, is evaluated without priming study participants about the presence or absence of authenticity indicators.

H3 investigates the explicit or *conscious* effect of authenticity indicators. We evaluate *H3* based on self-reported factors including the participant's free-text explanation of their credibility decision (prior to mention of authenticity indicators), and their explicit Likert-scale rating of the importance of authenticity indicators to their decision.

Finally, *H4* tests whether Twitter's description of verified accounts is sufficiently clear and interpretable to increase users' understanding of what verified accounts are (i.e., authenticity indicators) and are not (i.e., credibility indicators).

Two studies. Study 1 (n=748, August 2017) revealed no significant relation between the presence or absence of authenticity indicators and perceived credibility, whether implicitly or explicitly measured. To increase confidence in this null result, we used the data from Study 1 to generate a power estimate and corresponding desired sample size (details below), then conducted a second, independent study at this sample size (n=2041, August 2018). We made minor changes in Study 2 intended to strengthen our study design. Because the results of both studies aligned closely, hereafter we report primarily on Study 2; we highlight differences between Study 1 and Study 2 where relevant.

Conditions

Participants were assigned round-robin to one of 21 conditions. Each participant was shown and asked questions about one tweet, which depending on their assigned condition, varied in content, as well as in presence or absence of an authenticity indicator.

The primary variable of interest in our study is the presence or absence of an authenticity indicator. We used three different settings: a checkmark badge (✓), referred to as *Check*, that mimics the real Twitter verified account icon; a *Text* version that displays the text "Verified account" next to the author's display name; and no indicator. We created the *Text* version to suggest meaning to participants who might not be familiar with the standard Twitter icon. We did not give a more detailed definition to avoid priming.

To control for other possible indicators of credibility, other elements of the tweet (author profile image, count of retweets and likes, and time since the tweet was published[4]) were held constant. Previous research has shown that these elements have a significant effect on how users decide whether the tweet content and author are trustworthy [35].

We used four types of tweet content for our study: a restaurant review (*Restaurant*), a warning about the bacterial content of soda fountains (*Soda*), a study describing the harm of drinking too much coffee (*Coffee*), and a tweet discussing the impact of cattle disease on grocery bills (*Grocery*). In each case, the author's display name and username were chosen to be relevant to the content: The Food Critic; Science Facts; Mark Shields, MD; and North American Grocers Association and The Grocery Couple (explained below) respectively. Tweet topics were designed to be familiar to the participants; conversely, account names were selected because they do not exist and should therefore be unfamiliar to participants. However, no participant's free-text responses indicated that they thought the tweets appeared to be fake. Additionally, the tweet content was specifically selected to not be polarizing or bombastic to avoid triggering motivated reasoning. Free-text responses suggest these goals were met. Figure 1 shows example tweets shown to participants.

For the *Restaurant* review, we used one version with a positive review and one with a negative review to examine whether the valence of the content affects credibility. For the *Coffee* tweet, we included versions with and without the MD signal in the display name and username, in order to examine how implied qualification (i.e., having a medical doctorate) interacts with authenticity to impact message credibility [3]. For the *Grocery* tweet, we used two different account names (North American Grocers Association, The Grocery Couple) to examine whether an organizational source (as compared to individuals) affects credibility.

Our full-factorial design across authenticity indicators and content types resulted in 21 conditions, summarized in Table 1. While we do not attempt to exhaustively examine all possible effects of tweet content or author signals (e.g., race, gender, qualification) on user perceptions, we attempted to select a broad range, based on the literature. We did vary the sources' authority; however, this was only included to examine whether any interaction effects might exist, and was not meant to be a comprehensive investigation of the source's effect.

Study 1 used a similar full-factorial design across 15, rather than 21, tweets. Study 1 did not include *Coffee* or *Grocery* tweets, but instead used a tweet about herbal medicine designed to address similar questions as the *Coffee* tweet. Because the topic of herbal medicine proved especially controversial in Study 1, we replaced it for Study 2.

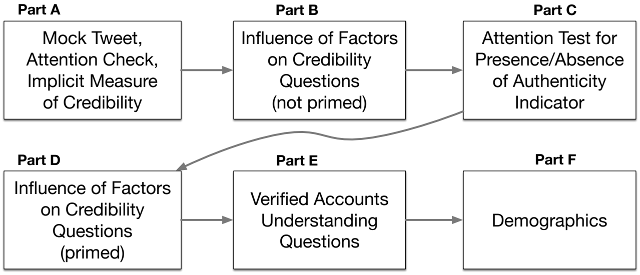

Part A
Mock Tweet, Attention Check, Implicit Measure of Credibility
Part B
Influence of Factors on Credibility Questions (not primed)
Part C
Attention Test for Presence/Absence of Authenticity Indicator
Part D
Influence of Factors on Credibility Questions (primed)
Part E
Verified Accounts Understanding Questions
Part F
Demographics

Figure 2: Sections and flow of the user study.

Study Procedure

Figure 2 illustrates the design of our online study. In Part A of the study, participants are first shown a tweet and asked to read its contents. They are then asked to answer three questions with the tweet remaining visible on the same page: First, they are asked about the content of the tweet as an attention check. They are then asked a content-specific question to assess their perceived credibility of the tweet. Third, they are asked to explain their credibility answer in free-text.

The content-specific questions asked whether the participant would be more or less likely to eat at the restaurant, purchase a fountain (rather than bottled) drink, change their daily coffee intake, or adjust their grocery-shopping or grocery-budgeting habits, with answers on a five-point scale from "much more likely" to "much less likely."

These content-specific questions were designed to evaluate hypothesis *H2* by asking about self-reported or expected behavior, rather than just an opinion change. Next, we showed the same tweet to the participants on a new page and asked how likely they were to share the information in the tweet (e.g. talking about it or retweeting), on a five-point scale from "very likely" to "very unlikely." This question (added in Study 2) serves as a second proxy for actionability; a participant might want to share, but not act directly on, information she found credible but not relevant to herself. Based on free-text actionability responses, few (8%) of our participants found the topics to not be relevant, evenly distributed among the authenticity indicator conditions (omnibus $\chi^2 = 0.09$, $p = 0.96$). The free-text follow-up allows us to examine *H3* by explicitly asking why they answered the behavior question.

We use behavioral questions rather than leading questions asking about credibility directly in order to avoid demand effects, in which participants provide what they expect to be "correct" answers [22]. Taking action in agreement with a message's content implies the user must first evaluate the message as credible [24], making this a reasonable proxy. Further, we deliberately asked about change in likelihood of taking an action instead of whether they would (or not) take the action. This phrasing appeared to be effective, as many

[4] In Study 2 we changed the tweet timestamp from "8 minutes ago" to "10:30 AM - July 31, 2018," to avoid considerations about how many likes or retweets would be reasonable within eight minutes.

participants indicated an increased likelihood to change behavior and stated in the free-text that they were more likely to take action, but wanted to seek out additional outside information first.

In Part B, we ask how each of four factors affected the participant's perception of the tweet's accuracy, on a five-point scale from "No effect" to "Major effect" (addressing hypothesis *H3*). These included the presence/absence of the authenticity indicator (the main feature of interest), the number of retweets and likes, the account username, and the account display name (all previously reported to have significant effects on social media credibility [5]). To avoid participants searching for clues to an expected "correct" answer [22], participants were not allowed to return to the tweet when answering these questions. Additionally, the four factors were presented in a randomized order, to minimize ordering effects [39]. We note that explicit mention of credibility (and of the authenticity indicator) in this section purposefully comes after the initial, unprimed reaction in Part A.

Part C of the study measures whether participants noticed the presence/absence of the authenticity indicator (*H1*). Participants were shown two versions of the original tweet side-by-side and asked to identify which they had seen previously. All participants saw their original tweet, plus either a version with no indicator (if they saw *Check* or *Text* originally), or a version with a verified icon (if they saw no indicator originally). To prevent random guessing, we asked participants to answer on a five-point scale from "Definitely the first image" to "Definitely the second image" with an "Unsure" option. We randomly ordered the two tweet images to avoid ordering effects.

Part D showed the participant their original tweet again, and asked them to answer the same four questions from Part B. Our goal is to understand how priming participants about the authenticity indicator in Part C changed their opinions.

In Part E, we measured participants' use and knowledge of Twitter. We first asked participants how much time per day they spend using Twitter on average. We then asked them to describe the meaning of Twitter's verified accounts in their own words and to rate their confidence in this description.

On a new page, we then presented participants with Twitter's definition of verified accounts and asked them to again describe verified accounts in their own words. This section allows us to examine how participants initially understood verified accounts (when rating credibility) and then whether they correctly interpret Twitter's definition.

Finally, we concluded (Part F) with demographic questions, including a seven-question scale measuring Internet Skill developed by Hargittai and Hsieh [18]. Participants took 9 minutes and 44 seconds on average to complete the study.

Recruitment

We used Amazon's Mechanical Turk (MTurk) crowdsourcing service to recruit participants.[5] To mitigate self-selection biases, account authenticity and verified accounts were not explicitly mentioned in the recruitment message.

We required participants to be at least 18 years old and located in the United States. To improve data quality, we also required participants to have completed at least 100 MTurk tasks with at least 95% approval [36]. Participants were paid $2.00 for completing the study, which was approved by the IRBs at Georgetown University and University of Maryland.

Data Analysis

To implicitly evaluate perceived credibility (*H2*), we use two ordinal logistic regressions, using the Likert scores for the action and sharing questions from Part A as outcome variables. We include as potential explanatory variables the assigned condition (i.e., content and authenticity indicator) as well as participant demographic characteristics (e.g., income, age, Twitter experience, etc.). To control for the effect of noticing the authenticity indicator (or lack thereof), we also include a three-level categorical covariate indicating whether the participant correctly identified which tweet they had seen with high confidence (correct answer and "definitely"); with low confidence (correct answer and "probably"); or got the answer wrong or marked unsure.

We constructed an initial regression model including the condition variables, the participant characteristics, and every two-way interaction between the authenticity indicator and the other variables, described in Table 2. We then used model selection on all possible combinations of these variables to select a parsimonious model without overfitting. We selected for minimum Bayesian Information Criterion (BIC), a standard metric [38]. Because authenticity indicator was the variable of key interest, we only considered candidate models that included the authenticity indicator variable.

We also examined the explicit influence of authenticity indicators on perceived credibility (*H3*). We compare participants' responses about the influence of four tweet features using non-parametric, repeated measures tests (appropriate for Likert-scale data with multiple answers per participant). To control for Type I error, we apply an omnibus Friedman test across all four features; if this result is significant, we apply the Wilcoxon signed-rank test to planned pairwise comparisons of authenticity indicator with every other feature. These comparisons were across tweet-content conditions.

For other comparisons among conditions, we use Pearson's χ^2 test, appropriate for categorical data [16]. Each comparison begins with an omnibus test over all three authenticity indicator options. If the result is significant, we

[5] For Study 1, we used TurkPrime [30] to manage requests.

Factor	Description	Baseline
Required factors		
Verified	The displayed authenticity indicator (or lack thereof)	None
Optional factors		
Content type	The displayed tweet content condition	Restaurant (negative)
Noticed	Whether the participant correctly indicated the presence or absence of the authenticity indicator in the original tweet with high or low confidence	Unsure/Incorrect
Internet skill	Participant's score on Hargittai and Hsieh's Internet skill scale [18]	–
Twitter experience	Time per day spent using Twitter (categorical)	None
Age	Age of participant	–
Gender	Gender of participant	Male
Education	Highest education level completed by participant	H.S. or below

Table 2: Factors used in regression models. Categorical variables are compared individually to the given baseline. Candidate models were defined using the required Verified factor plus all possible combinations of optional factors and interactions of the optional factors with the Verified factor. The final model was selected by minimum BIC.

apply χ^2 pairwise tests to two planned comparisons: the *Check* to *None* and *Text* to *None*.

Free response questions for the study were analyzed using open coding [44]. Two members of the research team reviewed each response individually in sets of 50, building the codebook incrementally. After each round of coding, the coders met to resolve differences in code assignment. After six rounds of pair coding (i.e., 300 responses), the coders achieved a Krippendorff's α [19] of 0.806 for the set of questions related to tweet credibility decisions and 0.849 for the questions related to the definition of verified accounts. Both values are within recommended bounds for agreement, so the remainder of the responses were divided evenly and coded by a single coder [19]. The resulting codebook appears in the supplemental material.

Finally, to test the effect of presenting Twitter's verified accounts definition on users' understanding of these accounts' meaning (*H4*), we compared participants' verified accounts definitions before and after priming, looking at whether they mentioned authenticity, public interest, and credibility. For these comparisons, we use McNemar's Chi-squared test, which is appropriate for within-subjects comparison with binary outcomes [33]. We apply a Holm-Bonferroni (H-B) correction to control for multiple comparisons [23].

Power Analysis. To calculate the necessary sample size, we applied Lyles et al.'s power simulation method with parameters estimated from Study 1 data [31]. We manually set the difference in distribution means to correspond to a "small" effect (Cohen's D = 0.2) and ran 50,000 simulations for each potential participant count. Power was measured as the percentage of simulations where a significant result was observed. To achieve 80% power, we estimated a required sample of 105 participants per condition. After recruiting 105 participants per condition in Study 2 and discarding invalid responses, we obtained 95 or more participants per condition. This equates to 78% power, which we consider sufficient [41, pg.296].

Ecological Validity and Limitations

We designed a controlled experiment with mock tweets. This allows us to reason about the effect of specific variables of interest on tweet credibility, but it does not capture potentially important real-world factors such as participants' history with the tweet's author or the reputation of someone who has "liked" or retweeted it. Further, there are many other types of content, and many metadata factors we did not test, that may (alone or in combination) influence the perceived credibility of tweets; Morris et al. describe the influence of several such features [35]. However, we believe our results provide useful insights for the particular questions we target.

Participants who are aware they are taking part in a study may have considered the tweet, and their response to it, more carefully than they would when casually browsing Twitter. We attempted to mitigate this by using a recruitment message that did not mention credibility, as well as by asking the primary credibility question prior to any priming about authenticity indicators.

To improve data quality, we kept the study brief and discarded responses that failed the attention check or gave non-meaningful answers to free-text questions. We used MTurk ID and browser cookies to prevent repeat participants. Prior research has generally found that MTurk provides high-quality data [4, 9, 27, 47], but we note that MTurkers located in the United States are typically somewhat younger, more tech-savvy and more privacy-sensitive than the general population, which may affect generalizability [26].

Although the populations of Twitter users and MTurkers are not identical, as we discuss in the next section, the majority of our study's participants used Twitter regularly (86%) and were familiar with Twitter's authenticity indicators (74.3%). We also control for prior experience with Twitter in our regression analysis.

Because these limitations apply across all 21 conditions, we rely primarily on comparisons between conditions.

Metric	Percentage	Metric	Percentage
Gender		**Age**	
Female	47.9%	18-29 years	33.0%
Male	51.8%	30-49 years	54.3%
Other	0.1%	50-64 years	9.9%
		65+ years	2.1%
Ethnicity			
Caucasian	79.6%	**Income**	
African American	9.1%	<$30k	22.1%
Asian	7.3%	$30k-$50k	25.4%
Hispanic	6.4%	$50k-$75k	23.7%
Other	1.8%	$75k-$100k	13.9%
		$100k-$150k	8.7%
Twitter Use		$150k+	3.4%
No Twitter use	13.9%	**Education**	
<1 hour per day	57.6%	H.S. or below	12.1%
1-2 hours per day	19.8%	Some college	36.4%
>2 hours per day	7.8%	B.S. or above	51.6%

Table 3: Participant demographics for Study 2.ercentages may not add to 100% due to non-response or selection of multiple options.

4 RESULTS

In this section, we present results related to each of our four hypotheses separately.

Participants

In total, 2238 people started our study, and 2211 completed it. We excluded 104 participants who failed the attention check and 66 participants who gave nonsensical or unresponsive answers to free-text questions. For the remainder of the paper we refer to the remaining 2041 participants. We obtained 94 to 100 valid responses per condition. Across conditions, 676 saw no authenticity indicator, 683 saw the *Check*, and 682 saw the *Text*.

Participant demographics are summarized in Table 3, and at least partially match Twitter's estimated user demographics [42]. As with many MTurk studies, our participants are somewhat whiter, younger, and more educated than the U.S. population. Our participants' average Internet skill was 31.4, slightly higher than the mean observed by Hargittai and Hsieh (30) in a more general population [18]. The majority of our participants (86%) reported using Twitter regularly. Twitter use did not vary significantly among the *Check*, *Text*, and *None* conditions ($\chi^2 = 13.79$, $p = 0.09$).

Noticing Verification (H1)

A little more than half of participants (consciously) noticed the verification status of the tweet. Overall, 1165 participants (57.1%) were able to correctly recall whether or not they saw a tweet with any authenticity indicator or not with at least some confidence; about half of these (681, 33.4% of all participants) were very confident about their choice. As a reminder, participants were offered a "not sure" option to prevent random guessing.

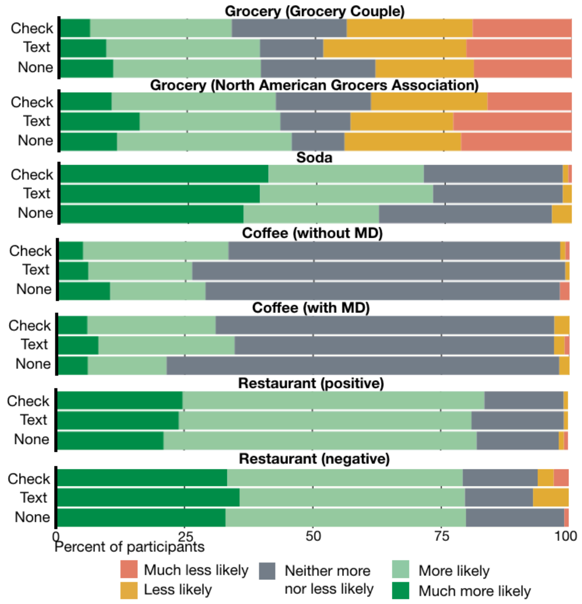


Figure 3: Likert-scale trust responses organized by assigned experiment condition (Study 2). Responses for the *Soda, Coffee*, and *Restaurant* (*Neg*) tweets were reversed to match polarity with the others.

Almost two-thirds of participants shown the *Text* indicator (63%) correctly recalled it with at least some confidence, compared to 52.9% for *Check* and 55.3% who saw no authenticity indicator. The corresponding omnibus test indicated significant differences among these conditions ($\chi^2 = 15.75$, $p < 0.001$). Planned pairwise tests indicate that while the *Text* indicator was more likely to be noticed than *None* ($\chi^2 = 8.07$, $p = 0.005$), there was no significant difference between *Check* and *None* ($\chi^2 = 0.74$, $p = 0.39$).

These results suggest that a majority of participants noticed the authenticity indicator (or lack thereof), regardless of which indicator they were assigned. We hypothesize that the *Text* indicator was most noticeable at least in part because it is not currently in use and was therefore novel.

Demonstrated Credibility (H2)

In examining H2, we were surprised to find no correlation between the presence of an authenticity indicator and whether the participant indicated they would act on or share the information in the tweet. For brevity, we primarily discuss participants' expressed willingness to act on the tweet's information (summarized as "credibility"), rather than expressed likelihood of sharing; the two questions provided similar results. We include in our supplemental materials graphs and tables for likelihood-to-share responses corresponding to those given in this section.

The authenticity indicator did not have a statistically significant effect on tweet credibility. In total, 53.4% of participants

Variable	Value	Odds Ratio	CI	p-value
Auth.	*None*	–	–	–
Indicator	*Check*	1.07	[0.88, 1.31]	0.473
	Text	1.07	[0.88, 1.30]	0.529
Content	restaurant (negative)	–	–	–
	restaurant (positive)	0.83	[0.62, 1.11]	0.217
	coffee (with MD)	**0.19**	**[0.14, 0.26]**	**< 0.001***
	coffee (without MD)	**0.2**	**[0.14, 0.26]**	**< 0.001***
	soda	0.89	[0.66, 1.22]	0.473
	grocery (org.)	**0.11**	**[0.08, 0.16]**	**< 0.001***
	grocery (person)	**0.09**	**[0.06, 0.12]**	**< 0.001***

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

Table 4: Summary of regression over participant trust in tweet credibility. Pseudo R^2 measures for this model were 0.08 (McFadden) and 0.22 (Nagelkerke).

who were shown an authenticity indicator (53.1% for *Check* and 53.8% for *Text*) found the tweet credible ("much more likely" or "more likely" to act). Participants in *None* conditions provided similar answers (51.2%).

A similar pattern is observed when considering each content type separately, as illustrated in Figure 3. Specifically, while rates of high and low credibility varied across content types, *within* each content condition results were fairly similar regardless of authenticity indicator. Further, credibility ordering among *Text*, *Check*, and *None* shows no consistent pattern across different content conditions.

In accordance with this lack of visible trends, our final selected regression model (Table 4) found no significant effect on credibility for either authenticity indicator compared to the baseline *None* condition (*Text*: $p = 0.529$, *Check*: $p = 0.473$).

Only the tweet content had a significant effect on credibility. Other than the authenticity indicator, which is the primary variable of interest and was therefore always included in the model, the only covariate selected for the final model was tweet content. Neither demographic covariates (including frequency of Twitter use) nor whether the participant had correctly noticed the authenticity indicator were retained; neither were any interactions. Figure 4, which shows participants' credibility responses grouped by the authenticity indicator and whether the participant correctly noticed it, demonstrates the lack of visible trend relating credibility results to noticing. We note that while the main content of the tweet mattered (*Restaurant* and *Soda* were most credible), within-content variations (authoritative title, individual or organization, and positive or negative valence) had no significant effects, as indicated by overlapping confidence intervals in the regression model.

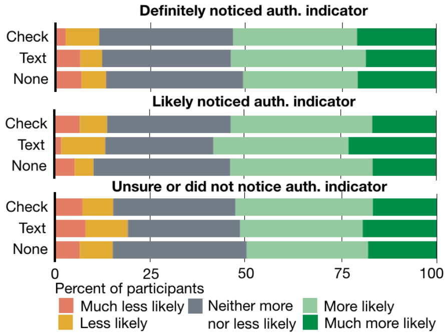


Figure 4: Likert-scale trust responses organized by assigned verified account status and whether or not they noticed the verified account identifier (or lack thereof).

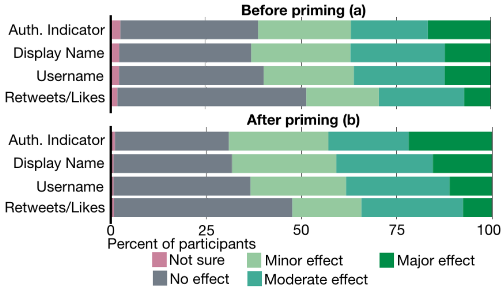


Figure 5: Likert-scale influence responses organized by influencing factor, (a) not primed (b) primed.

Average time using Twitter per day and Internet skill significantly correlate with likelihood to share. The likelihood of sharing exhibited similar results as the willingness to act for the authenticity indicator and tweet content. However, the final likelihood-to-share model retained two additional variables: participants who used Twitter more often and those who had lower Internet skill were more likely to share. We still observed no significant interaction between these variables and the authenticity indicator.

It is important to note that failing to find a significant effect is not affirmative proof that no effect exists; hypothesis tests do not bound false negatives. However, based on the power analysis describe above, we should have sufficient power to detect even a small effect. Further, this negative result held both for likelihood to act on the tweet and to share it. Additionally, visual inspection of our detailed results (including but not limited to Figures 3 and 4) does not indicate any trends or patterns related to the authenticity indicator that failed to reach a significance threshold. Overall, we are confident in the validity of this negative result.

Perception of Credibility Indicators (H3)

We explicitly examine perceptions of credibility (H3) in two ways: unprompted free-text responses about their answers to the actionable credibility question (Part A), and explicit prompting about the importance of the authenticity indicator (Parts B and D).

The authenticity indicator was rarely mentioned unprompted. Only 1.4% of participants explicitly mentioned the presence or absence of the authenticity indicator in their free-text response. In contrast, 70.7% of participants mentioned content as the most important factor. For example, one participant wrote, "It said the service was bad, so I don't want to risk getting bad service there."

Most participants indicated that the authenticity indicator had at most a minor effect on tweet credibility. Figure 5 summarizes the quantitative results before and after asking the participant explicitly whether they had seen an authenticity indicator. In both cases, a plurality of participants reported no effect, and a majority reported at most a minor effect.

Although reported effects were minor, in Part B the authenticity indicator did have a significantly stronger effect than other factors we asked about ($p < 0.05$ for the omnibus Friedman test and for planned comparisons of the authenticity indicator to each other factor). The effect sizes for these pairwise comparisons, however, were very small or very small (Cohen's D = 0.06 for display name, 0.09 for username, and 0.21 for retweets/likes) [7].

After calling further attention to the authenticity indicator (by asking the participant which tweet they saw), quantitative and qualitative responses in Part D show an increase in its reported importance. Almost half 45.4% of participants mentioned the authenticity indicator in free-text, and 36.9% of participants who originally reported the authenticity indicator had no effect in the Likert questions reported some level of effect afterward. These results may suggest calling attention to the authenticity indicator increases its salience as a proxy for credibility; however, this may also reflect so-called *demand effects*: participants who believed they understood the purpose of the study (to comment on authenticity indicators) and tried to provide "correct" responses.

Overall, we conclude that H3 was not well supported; participants were much more likely to focus on the tweet's content than the authenticity indicator, which (when unprimed) was largely reported as at most a minor influence on credibility decisions.

Beliefs About Verification (H4)

We examined participants' understanding of verified accounts on Twitter (H4) by qualitatively coding participants' free-text responses from Part E, both before (non-primed) and after (primed) showing them the official Twitter definition: "The blue verified badge ✓ on Twitter lets people know that an account of public interest is authentic" [48].

We regard participants who mention that verified accounts have been established, by Twitter, as actually belonging to the offline entity they claim to be as *correct.* For example, one participant wrote,"Twitter has verified the user is the person they claim to be."

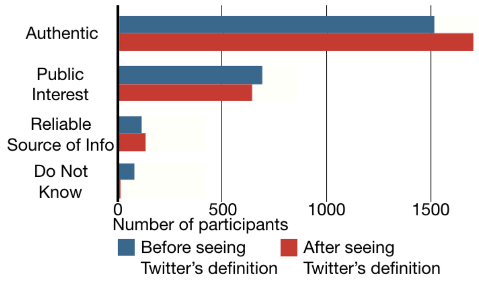


Figure 6: Number of participants who mentioned each theme regarding the definition of verified accounts before and after being primed with Twitter's official definition.

Most participants (74.3%) provided correct definitions even without priming. In addition, 34.2% of participants mentioned public interest. (Note that multiple codes could apply to the same response.) Only 5.9% indicated that verified accounts were more trustworthy sources of information. For example, one participant wrote: "Apparently, this means you can expect that the owner of the account has met some sort of scrutiny as a trusted source for information - such as a verified journalist for a media company."

Correct responses increased after priming. After priming with the definition, more participants (83.5%) mentioned authenticity, while slightly fewer (31.8%) mentioned public interest. The increase in mentions of authenticity was significant, but the decrease in mentions of public interest was not (McNemar's $\chi^2 = 73.77$, $p < 0.001$ and $\chi^2 = 3.90$, $p = 0.144$ respectively, both H-B corrected).

Somewhat surprisingly, the share of participants who incorrectly mentioned credibility and reliability after viewing the official definition rose slightly, to 6.9%, but this difference was not significant (McNemar $\chi^2 = 1.75$, $p = 0.557$, H-B corrected). In Study 1, however, we observed a similar increase—5.8% to 9.4%—that was statistically significant ($\chi^2 = 9.72$, $p = 0.002$, H-B corrected).

5 DISCUSSION

Our two studies found that authenticity indicators have little to no effect on users' perceptions of credibility. Users generally understand the meaning of verified accounts, and most notice the presence or absence of the authenticity indicators. However, users are not more likely to act on or share content that originates from verified accounts than from unverified accounts. Further, users rarely cited the authenticity indicator as an important influence on credibility decisions unprompted, or even with minimal prompting.

The apparent negative results for hypotheses $H2$ and $H3$ seem to contradict a large volume of existing work [6, 14, 21, 45] that portrays reputation and consensus as major influences on perceived credibility. In other work, study participants explicitly cited the presence of authenticity indicators as an important factor in forming credibility assessments of tweets [35]. Overall, our findings show the opposite: authenticity indicators themselves have little or no effect.

An optimistic view of our findings is that users simply have a more mature understanding of the differences between authenticity and credibility than we expected; that is, they understand that a tweet's author's authenticity does not imply that the contents of her tweets are accurate. This theory is bolstered by the high rate at which participants could accurately describe the meaning of verified accounts, even before being presented with the official definition.

However, there are other factors out of our study's scope whose interaction with the authenticity indicator could have an effect. The authenticity indicator's relationship with each of these factors should be studied further.

First, the effect of authenticity indicators could be more significant when users are familiar with the author. We used fictional accounts, and thus our participants lacked any familiarity, regardless of the account's verification status. The effect of celebrity endorsements on user behavior has been well-explored (see, for example, Erdogan's survey [12]); this literature suggests that users would be more likely to perceive tweets from familiar celebrities as more credible.

We purposefully chose neutral topics rather than politically or culturally polarizing ones. As a consequence, we avoid a potential motivated reasoning effect [29]. That is, our participants may have been less inclined to either strongly support or reject our tweets because they lacked motivation.

Notably, our studies were conducted during a period when dissemination of fake news [10, 54] has received considerable attention. Therefore, users may have developed a skepticism of online information, regardless of the source. In Study 2, which took place after a period of public upheaval about how Twitter's verification process works and whether verified status implies approval, a small number of participants (1.3%) stated that Twitter's account verification is politically biased. This sentiment was not seen at all in Study 1, possibly indicating the effect of the changing political environment.

Finally, like all MTurk studies, our participants were younger and more educated than the general population. In our studies, Twitter familiarity and Internet skill did not appear to interact with authenticity indicators in credibility decisions; however, future work is needed to validate this result with less knowledgeable populations.

REFERENCES

[1] Ruba Abu-Salma, M Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. 2017. Obstacles to the Adoption of Secure Communication Tools. In *IEEE Symposium on Security and Privacy (SP)*.

[2] Hunt Allcott and Matthew Gentzkow. 2017. *Social Media and Fake news in the 2016 Election.* Technical Report. National Bureau of Economic Research.

[3] David K Berlo, James B Lemert, and Robert J Mertz. 1969. Dimensions for Evaluating the Acceptability of Message Sources. *Public opinion quarterly* 33, 4 (1969), 563–576.

[4] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. arXiv:http://pps.sagepub.com/content/6/1/3.full.pdf+html

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 675–684.

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

[7] J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates.

[8] Rachna Dhamija, J. D. Tygar, and Marti Hearst. 2006. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 581–590.

[9] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System? Screening Mechanical Turk Workers. In *Conference on Human Factors in Computing Systems (CHI)*.

[10] Elizabeth Dwoskin. 2017. Twitter is Looking for Ways to Let Users Flag Fake News, Offensive content. *The Washington Post* June (2017).

[11] Emily Heil. 2017. Trump posts fake Abraham Lincoln quote. Available at **www.washingtonpost.com/news/reliable-source/wp/2017/02/13/trump-posts-fake-abraham-lincoln-quote/**.

[12] B. Zafer Erdogan. 1999. Celebrity Endorsement: A Literature Review. *Journal of Marketing Management* 15, 4 (1999), 291–314.

[13] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. 2016. Rethinking Connection Security Indicators. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS '16)*.

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

[15] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How Do Users Evaluate the Credibility of Web Sites? A Study with over 2,500 Participants. In *Proceedings of the 1st Conference on Designing for User Experiences (DUX '03)*.

[16] Karl Pearson F.R.S. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* 50, 302 (1900), 157–175.

[17] Gerd Gigerenzer, Peter M Todd, the ABC Research Group, et al. 1999. *Simple Heuristics that Make us Smart.* Oxford University Press.

[18] Eszter Hargittai and Yuli Patrick Hsieh. 2012. Succinct survey measures of web-use skills. *Social Science Computer Review* 30, 1 (2012), 95–107.

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

[20] Amir Herzberg and Ahmad Jbara. 2008. Security and Identification Indicators for Browsers Against Spoofing and Phishing Attacks. *ACM Transactions on Internet Technology (TOIT)* 8, 4 (Oct. 2008), 16:1–16:36.

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

[22] Allyson L Holbrook, Melanie C Green, and Jon A Krosnick. 2003. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public opinion quarterly* 67, 1 (2003), 79–125.

[23] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[24] Carl I Hovland, Irving L Janis, and Harold H Kelley. 1953. *Communication and Persuasion; Psychological Studies of Opinion Change.* Yale University Press.

[25] Haje Jan Kamps. 2015. Who Are Twitter's Verified Users? Medium post. Available at **https://medium.com/@Haje/who-are-twitter-s-verified-users-af976fc1b032**.

[26] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. 2014. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *USENIX Security Symposium.*

[27] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Conference on Human Factors in Computing Systems (CHI).*

[28] K. P. Krishna Kumar and G. Geethakumari. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences* 4, 1 (24 Sep 2014), 14.

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

[30] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for The Behavioral Sciences. *Behavior research methods* 49, 2 (2017), 433–442.

[31] Robert H. Lyles, Hung-Mo Lin, and John M. Williamson. [n. d.]. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine* 26, 7 ([n. d.]), 1632–1648. **https://doi.org/10.1002/sim.2617** arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2617

[32] Jim Maddock, Kate Starbird, Haneen J. Al-Hassani, Daniel E. Sandoval, Mania Orand, and Robert M. Mason. 2015. Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15).*

[33] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (01 Jun 1947), 153–157.

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

[35] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is Believing?: Understanding Microblog Credibility Perceptions. In *Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work (CSCW '12).* ACM, New York, NY, USA, 441–450.

[36] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (01 Dec 2014), 1023–1031.

[37] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor Has It: Identifying Misinformation in Microblogs. In *Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1589–1599.

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

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

[40] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. 2007. The Emperor's New Security Indicators. In *Proceedings of the 28th IEEE Symposium on Security and Privacy (S&P '07)*. 51–65.

[41] Howard J Seltman. 2012. Experimental design and analysis. *Online at: http://www. stat. cmu. edu/, hseltman/309/Book/Book. pdf* (2012).

[42] Aaron Smith and Monica Anderson. 2018. *Social Media Use in 2018*. Internet & Technology Report. Pew Research Center.

[43] Peter Steiner. 1993. On the Internet, Nobody Knows You're a Dog (Cartoon). *The New Yorker* (July 5 1993).

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

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

[46] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *Proceedings of the 18th USENIX Security Symposium (USENIX Security '09)*.

[47] Michael Toomim, Travis Kriplean, Claus Pörtner, and James Landay. 2011. Utility of Human-computer Interactions: Toward a Science of Preference Measurement. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA.

[48] Twitter. 2017. About Verified Accounts. **https://support.twitter.com/articles/119135**.

[49] Twitter. 2017. Request to Verify an Account. **https://support.twitter.com/articles/20174631**.

[50] Twitter Support. 2017. Statement Regarding Account Verification (Tweet). **https://twitter.com/TwitterSupport/status/930926124892168192**.

[51] Ward van Zoonen and Toni van der Meer. 2015. The Importance of Source and Credibility Perception in Times of Crisis: Crisis Communication in a Socially Mediated Era. *Journal of Public Relations Research* 27, 5 (2015), 371–388.

[52] Jonathan Vanian. 2017. Twitter Removes Verified Status From Users For Racist and Hateful Posts. Available at **http://fortune.com/2017/11/15/twitter-verification-racist-hateful/**.

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

[54] Nick Wingfield, Mike Isaac, and Katie Benner. 2016. Google and Facebook Take Aim at Fake News Sites. *The New York Times* 11 (2016).