https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Vol. 37, No. 1, March 2026, pp. 103–116
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

The Illusion of Authenticity in Online Reviews: Truth Bias and the Role of Valence

Dezhi Yin,[a] Samuel D. Bond,[b] Han Zhang[b,c,*]

[a] Muma College of Business, University of South Florida, Tampa, Florida 33620; [b] Scheller College of Business, Georgia Institute of Technology, Atlanta, Georgia 30308; [c] School of Business, Hong Kong Baptist University, Hong Kong, China

*Corresponding author

Contact: dezhiyin@usf.edu, https://orcid.org/0000-0003-1107-3232 (DY); sam.bond@scheller.gatech.edu, https://orcid.org/0000-0002-0457-3290 (SDB); han.zhang@scheller.gatech.edu, https://orcid.org/0000-0002-6258-2486 (HZ)

Received: June 7, 2023
Revised: March 16, 2024; August 10, 2024; November 27, 2024
Accepted: February 19, 2025
Published Online in Articles in Advance: May 20, 2025

https://doi.org/10.1287/isre.2023.0339

Abstract. Despite a growing stream of research documenting the prevalence of "fake" online reviews and improving their automated detection, little is known about how consumers make real or fake judgments of reviews with unknown veracity. Integrating literature on truth-default theory and deception motives, we propose that consumers have a general tendency to view reviews as real rather than fake (a truth bias) and to be more accurate at detecting real reviews than fake reviews (a veracity effect). Moreover, we argue that the truth bias is weaker for positive reviews than negative reviews (a valence effect) because of a largely automatic process in which consumers project deception motives onto reviewers. To test these proposals, we conducted five experiments in which participants classified sets of reviews as real or fake. Results provided broad support for our theorizing, and they have important implications for firms and platforms as they establish priorities for combating fake reviews.

History: Choon-Ling Sia, Senior Editor; David (Jingjun) Xu, Associate Editor.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2023.0339.

Keywords: online reviews • fake reviews • truth bias • veracity effect • valence effect • deception detection

Introduction

Online ratings and reviews play an increasingly important role in consumer purchase decisions, and mounting evidence confirms that they have a substantial impact on sales (Babić Rosario et al. 2016, Ismagilova et al. 2020). As a result, opportunistic businesses sometimes choose to artificially manipulate reviews of themselves or their competitors. The present research focuses on a common type of manipulation involving *fake reviews*, defined broadly as deceptive reviews intended to mislead consumers (Zhang et al. 2016, Wu et al. 2020). The proliferation of fake reviews has received increasing attention in business and mainstream media, with industry experts estimating that one quarter to one third of Amazon reviews are fake (Kapner 2019, Schoolov 2020). Meanwhile, an influential body of scholarly research has combined large, archival data sets with modern causal estimation techniques to document the substantial prevalence of fake reviews across numerous platforms (Mayzlin et al. 2014, Lappas et al. 2016, Luca and Zervas 2016, Song et al. 2023).

To combat the problem, platforms have increasingly taken steps to identify suspicious reviews and penalize known fake reviewers. Consistent with this trend, a growing body of research focuses on improving the automated detection of fake reviews (e.g., Zhang et al. 2016, Kumar et al. 2018, Ng et al. 2023). Typical approaches use machine learning to optimize models based on specific features of review content and other contextual variables. For example, Zhang et al. (2016) demonstrate that "nonverbal" features of reviewer profiles are useful for review detection, over and above features of the review text itself. Such advances have been instrumental for review platforms in mitigating the impact of potentially fake reviews. Nonetheless, platforms remain inherently limited in their ability to determine review veracity (i.e., to accurately classify reviews as real or fake) as strategic firms and malicious reviewers are constantly evolving their deception strategies in order to avoid detection (Anderson and Simester 2014).

In contrast to a technical perspective that focuses on algorithmic estimation of review veracity, we approach the issue from a *consumer perception* perspective by addressing two related questions. First, do consumers have a general tendency to suspect or trust reviews whose veracity is unknown? On the one hand, increased awareness and media coverage of fake reviews may have led consumers to doubt the credibility of reviews in general. If so, then such suspicion could negatively impact the bottom lines of platforms for which reviews are an important source of competitive advantage. On

the other hand, it may be reasonable for consumers to trust reviews in general given that the majority of reviews on popular platforms are real (Lappas et al. 2016, Luca and Zervas 2016). If so, then they could easily be misled at the expense of time, money, and effort, a concern that has reinforced the current emphasis on development of automatic detection tools.

Second, how does review valence influence the above-mentioned tendencies? As used here, review valence refers to the positivity or negativity of reviews reflected in both their ratings (e.g., five stars versus one star) and their textual content. Whereas valence has no analog in traditional settings of truth-lie detection (see below), it is an essential feature of online reviews. In particular, valence provides a salient, easily interpretable cue that may impact reader interpretation of both the reviewer and the review (e.g., Qiu et al. 2012). If so, then platforms may benefit from taking valence into account as they develop more nuanced interventions.

We address these questions both theoretically and empirically. Integrating "truth-default" theory (Gilbert 1991, Levine et al. 1999, Levine 2019) and the deception motives literature (Levine et al. 2016), we argue that consumers tend to perceive reviews as real rather than fake (i.e., a truth bias), and thus, they are more accurate at detecting real reviews than fake reviews (i.e., a veracity effect). In addition, we argue that truth bias should be weaker when consumers evaluate positive reviews than negative reviews (i.e., a valence effect) because of a largely automatic process by which positive reviews are more readily associated with deceptive motives. Across a series of five controlled experiments, we find converging evidence for these proposals.

Our research makes two primary contributions. First, it is among the first to explore judgments of review veracity from a consumer perception perspective by focusing on how consumers make veracity judgments in the absence of credible signals. Building on prior evidence (Plotkina et al. 2020), we utilize a highly conservative approach to detect evidence for truth bias in review settings. In doing so, we examine how truth bias relates to overall accuracy in discriminating real versus fake reviews, and we address the "rationality" of truth bias in real-world review settings. Second, we explore the effect of valence on perceptions of review credibility. Although there exists a substantial body of information systems research on review valence and perceived helpfulness (e.g., Yin et al. 2016, Lei et al. 2023), its association with credibility is largely and surprisingly unexplored. By demonstrating robust effects of valence and identifying plausible mechanisms underlying those effects, we inform a broader understanding of how review valence impacts consumer judgment.

Theory and Hypotheses Development

Deception Detection

Borrowing from Zhang et al. (2016, p. 456), we define fake reviews as "deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being evaluated," and we define real reviews as the opposite ("honest reviews provided without an intention to mislead … "). The valence of fake reviews may be either positive (using deception to promote a product or help a business) or negative (using deception to damage the reputation of competing products or businesses). Depending on the context, sources of fake reviews range from employees, family, or friends of the manipulating entity to paid freelancers, social media groups, or professional services (Gössling et al. 2018).

Although fake reviews may be generated by human authors or machines ("bots"), we restrict our focus to the former. This restriction is primarily for methodological reasons as the fact that bot-generated reviews are inherently fake makes it impossible to compare veracity judgments of real and fake subsets. For now, bot-generated reviews remain relatively rare (e.g., one recent investigation estimated that fewer than 10% of "unverified" reviews on the Yelp! platform were generated by bots or artificial intelligence; Gambetti and Han 2023). More importantly, the theoretical arguments that we develop below do not depend on the source of the review as long as that source is not disclosed. Regardless of whether a fake review is written by a human or a bot, therefore, our arguments and findings should logically apply.

Decades of research in applied psychology, communication, and related disciplines have examined deception—defined as "intentionally, knowingly, or purposefully misleading another person" (Levine 2019, p. 102)—and individuals' ability to recognize it. In the most common paradigm, participants are presented with one or more messages from a naturalistic setting (recorded courtroom testimonies, in-person interviews, etc.) and then asked to judge whether each message is truthful or deceptive. Within the paradigm are a vast number of "truth-lie detection accuracy" studies (Ekman and O'Sullivan 1991; Mann et al. 2004; Bond and DePaulo 2006, 2008), in which the truthfulness of each message is definitively known, and thus, judgment accuracy can be objectively determined. A consistent finding of these studies is that people are remarkably poor at distinguishing lies from the truth; in fact, average detection accuracy tends to be only slightly better than chance (54% in one prominent meta-analysis; Bond and DePaulo 2006). This finding has proven robust to different message topics, judges, situations, and attempted interventions. Moreover, judges tend to be poorly calibrated regarding their detection accuracy

such that their subjective confidence is both too high overall and only weakly correlated with their actual performance (DePaulo et al. 1997).

Various theoretical accounts of truth-lie detection have been proposed, among which prominent examples are Ekman's leakage theory, four-factor theory, and interpersonal deception theory (for a detailed summary, see Levine 2019, chapter 4). Although the accounts differ markedly, they all assume that (a) truths and lies are different in meaningful ways, (b) the differences will manifest through observable verbal or nonverbal cues, and (c) accurate detection of truths and lies depends on the ability of observers to identify those cues. Within the online review literature, similar assumptions underlie a stream of technical investigations that attempt to identify systematic differences between real and fake reviews (e.g., Mukherjee et al. 2013, Zhang et al. 2016). Contrary to these assumptions, however, a broad scholarly consensus has emerged that nonverbal and verbal cues to deception are generally faint and unreliable (Zuckerman et al. 1981; DePaulo et al. 2003; Sporer and Schwandt 2006, 2007).

Truth Bias and Its Consequences

An alternative lens for understanding truth-lie judgments builds on general cognitive tendencies to believe or disbelieve incoming information. Within this lens lies an important theoretical foundation for our research, known as "truth-default theory" (Gilbert 1991, Levine 2019). In contrast to theories assuming that people actively scrutinize messages in search of clues to their veracity, the basic premise of truth-default theory is that people tend to passively assume communication to be honest (Levine 2014). One consequence of this tendency is a "truth bias," defined as a general tendency to judge communications from others to be truthful (Zuckerman et al. 1981, McCornack and Parks 1986, Levine 2019). Within the truth-lie detection paradigm described above, truth bias is calculated by comparing the percentage of messages classified as "true" with the percentage that would logically be expected given the context (usually 50% in experimental investigations); thus, a positive difference indicates a truth bias. Although we acknowledge that the term "bias" has different connotations in other fields, we adopt the conventional term and definition in this research.

In everyday communication settings, the presence of truth bias does not need to be deleterious and may even be adaptive (Levine 2019). From a cognitive resource perspective, being constantly "on one's guard" against deception is taxing and unsustainable, whereas the presumption of honesty diminishes the need to engage in effortful scrutiny. Bolstering this perspective is the fact that "truths" are overwhelmingly more common than "lies." Contrary to the common belief that deception is pervasive, both survey research and experimental research consistently find that the vast majority of people are honest most of the time (Serota et al. 2010, Halevy et al. 2014). Hence, a general tendency to assume honesty can be efficient and functional for real-world interactions.

The benefits of the truth bias in broader communication settings logically extend to the setting of consumer reviews. As shopping increasingly moves online, the vast and growing number of reviews available only intensifies the need for consumers to process them efficiently. Moreover, research examining popular platforms consistently shows fake reviews to be heavily outnumbered by real reviews (Wu et al. 2020). Therefore, it is reasonable to expect that consumers will exhibit a truth bias when processing online reviews. Plotkina et al. (2020) provide initial evidence that this is the case: when presented with reviews of unknown veracity, participants judged a majority of the reviews to be real. However, such evidence is open to other interpretations (e.g., participants may have simply assumed that a majority of reviews were real and responded in line with that assumption). The most compelling and unambiguous evidence for truth bias occurs when recipients are *fully aware* of the actual distribution of "truths" and "lies" in a set of messages, but they nonetheless classify too many messages as "true" (Levine 2019). Applying this more conservative measure to our studies, we propose the following.

Hypothesis 1 (Truth Bias). *Even when consumers know that real reviews and fake reviews are equally likely, they tend to judge more reviews as real than fake.*

Our second hypothesis concerns the accuracy of consumers in detecting real and fake reviews. As noted above, a robust finding from the truth-lie detection literature is that individuals tend to be relatively poor in distinguishing truths from lies, yielding accuracy rates just above "chance" (Bond and DePaulo 2006, 2008). However, focusing on aggregate performance alone can mask an essential but often-ignored predictor of detection accuracy: the actual veracity of individual messages. Extending the logic of truth-default theory, the presence of truth bias has important implications for the relative classification accuracy of real and fake messages. As the tendency to classify messages as true rather than false increases, the likelihood that truths are correctly identified becomes higher, and the likelihood that lies are correctly identified becomes lower (Zuckerman et al. 1984). Hence, judges with an innate preference to classify messages as "true" will be more accurate at identifying truths than lies, a phenomenon known as a "veracity effect" (Levine et al. 1999).

Applying the preceding logic to online review settings, a tendency to judge individual reviews as real rather than fake implies that the accuracy rate for real

reviews will be higher than that for fake reviews. Given that most online reviews are in fact real, superior accuracy in detecting real reviews might be viewed as functional and beneficial. More generally, evidence of a veracity effect in this setting would have substantial implications for our understanding of review biases and the means of addressing them. Stated formally, we propose the following:

Hypothesis 2 (Veracity Effect). *Consumers are more accurate at identifying real reviews than fake reviews.*

Deception Motives and Review Valence

Notwithstanding the general pervasiveness of truth bias, its operation in specific contexts can be dramatically altered by situational cues that trigger suspicion (Levine et al. 2010, Levine 2019). Especially important are situational cues that signal the potential motives of message senders; stated simply, recipients are more likely to become suspicious when they believe that senders have a reason to lie.

Levine et al. (2016) present a comprehensive list of potential deception motives. Three of those motives are especially relevant in the online review context: economic advantage (i.e., the desire for monetary gain, which is applicable to both fake positive and fake negative reviews), impression management (the desire to appear more favorable to others, mainly applicable to fake positive reviews), and malice (the desire to harm others, mainly applicable to fake negative reviews). Thus, the valence of a review may serve as a salient cue by which review readers project deception motives onto authors.

Importantly, however, we propose that the projection of deceptive motives is much more likely in the case of positive reviews than negative reviews. Compared with fake negative reviews, it is easier for readers to envision both (1) the benefits that fake positive reviews can provide to parties involved and (2) the process by which they might come to exist. In fact, the projection of deception motives for positive reviews can be as simple as assuming that a firm "incentivized" its customers to provide favorable evaluations. In the case of negative reviews, however, the projection of deception motives requires readers to visualize an ambiguous, ill-intentioned competitor who has devised an effective process to manufacture and disseminate unfavorable evaluations; relative to the positive case, this visualization requires more cognitive effort and creativity.

If positive reviews are more likely than negative reviews to trigger suspicion of deception motives, then that suspicion should, in turn, weaken the tendency to classify positive reviews as real.[1] Stated formally, we propose the following:

Hypothesis 3 (Valence Effect). *Consumers exhibit a smaller truth bias for positive reviews than negative reviews.*

Our final hypothesis concerns the mechanism underlying the valence effect. The argument that deception motives are easier to envision for positive than negative reviews relates to the concept of cognitive accessibility (Feldman and Lynch 1988, Wyer 2018). Given that decision makers are "cognitive misers" (Simon 1956), they tend to automatically and unconsciously rely on highly accessible information. This tendency is pronounced when cognitive resources are limited, as is typically the case for consumers evaluating reviews in the midst of the shopping process. As described above, we propose that the valence effect is based on simple schema-based associations between review valence and deception motives that are stronger for positive than negative reviews. Hence, we propose that the underlying process is largely automatic.

Hypothesis 4. *The valence effect* (a) *is mediated by consumers' projection of deception motives, which* (b) *is a largely automatic process.*

Overview of Studies

We investigated our hypotheses with five experiments (see Table 1). To generate stimuli for the experiments, we obtained a pool of restaurant reviews whose veracity was objectively known (see below). We chose the restaurant category because it is relevant to most consumers and because consulting reviews is a common practice when choosing a restaurant. All of the experiments utilized a similar procedure, in which participants were informed that they would view an equal number of "real" and "fake" reviews that had been randomly selected from the review pool. After

Table 1. Primary Objectives of Studies

	Study focus	Hypothesis 1: Truth bias	Hypothesis 2: Veracity effect	Hypothesis 3: Valence effect	Hypothesis 4(a): Mediation of valence effect by deception motives	Hypothesis 4(b): Automaticity of valence effect
Study 1	Basic effects	×	×	×		
Study 2	Confounds	×	×	×		
Study 3	Confounds			×		
Study 4	Mechanisms				×	
Study 5	Mechanisms					×

reading each review, participants were asked to judge whether it was real or fake. We predicted that they would tend to classify the majority of reviews as real (a truth bias), that their accuracy at identifying real reviews would tend to be greater than their accuracy at identifying fake reviews (a veracity effect), and that both of these tendencies would be greater for negative reviews than positive reviews (a valence effect). Study 1 provided initial evidence for the truth bias, veracity effect, and valence effect. Studies 2 and 3 addressed potential confounds related to the procedure and review stimuli, respectively. The final two studies focused on mechanisms underlying the valence effect. Study 4 examined the mediating role of deception motives, and Study 5 considered alternative mechanisms involving cognitive deliberation.

Stimulus Materials

Prior to the main studies, we conducted a "review collection" pretest, which was designed to generate a pool of realistic restaurant reviews that differed in valence and veracity. This approach to stimulus creation is common in the lie detection literature (e.g., Forrest and Feldman 2000, Reinhard et al. 2013). We recruited 205 undergraduates from a U.S. university who received extra credit for participating. We deemed undergraduates appropriate because they are heavy users in the category; over half of U.S. undergraduates visit an off-campus restaurant at least once a week (Datassential 2019). Pretest instructions and materials are provided in Online Appendix A. The cover story introduced participants to Yelp!, a popular consumer review platform where users can learn about local businesses. Participants in the pretest were told that Yelp! and other review sites have become increasingly concerned about the problem of "fake" reviews and that the purpose of the research was to help identify and eliminate such reviews.

Participants were randomly assigned to one of four conditions based on review valence (*positive* or *negative*) and review veracity (*real* or *fake*). In the first step, participants were asked to think of a restaurant that they had recently visited at which their experience was positive or negative (depending on condition), and they were asked to provide the name of the restaurant. In the second step, participants were asked to write a review of the restaurant following specific instructions that varied by condition. Participants assigned to the *real* conditions were simply asked to write a review based on their own actual experience at the restaurant. Participants assigned to the *fake* conditions were asked to imagine that they worked for the "social media marketing team" of either the restaurant itself (*positive* condition) or one of its nearby competitors (*negative* condition) and that their boss had instructed them to write a "fake" review; they were told that "Your job is to convince prospective readers that your [positive/negative] review is based on your actual experience (even though in reality, it is not)" and to "Feel free to invent or imagine an experience as needed." Across all conditions, participants were instructed to take their time and ensure that their reviews were of acceptable quality (legible, reasonable in length, and not plagiarized). In addition to providing a review, participants rated their (real or imagined) restaurant experiences on a scale from one to five stars. At the end of the procedure, participants completed a veracity manipulation check that asked whether the review was based on their own actual experience ("yes" or "no").[2]

To select reviews for the final pool, we applied three criteria. First, we excluded reviews that were only one sentence or shorter in length. Second, we excluded *positive* reviews with a self-assigned rating less than four stars and *negative* reviews with a self-assigned rating greater than two stars. Third, we excluded reviews that failed the veracity manipulation check. The end result was a pool of 205 reviews, consisting of 47, 32, 32, and 36 reviews in the *real-positive*, *real-negative*, *fake-positive*, and *fake-negative conditions*, respectively. Table 2 provides a sample review for each condition.

Study 1

In the first experiment, participants were asked to read 20 reviews from the pool that differed in valence and veracity. The design incorporated a 50-50 base rate (i.e., an equal number of truthful and deceptive reviews); this base rate is common practice in the truth-lie detection literature and simplifies the calculation of accuracy (Bond and DePaulo 2006). Participants were informed in advance that half of the reviews were "real" and that the other half were "fake." Therefore, classifying more than half of the reviews as real or fake would represent evidence of a truth bias or lie bias, respectively. The reviews appeared one at a time on separate screens, and participants were asked to classify each review into one of the two categories.

Procedure

We recruited 113 undergraduate students (69 female) at a U.S. university who took part for course credit.[3] Study materials are provided in Online Appendix B. In the cover story, participants were introduced to a fictitious third-party review site "RestaurantReviews.com." The cover story explained that the site had become increasingly concerned about the increasing "fake" review problem, that the researchers were working with the site to help identify fake reviews, and that we had gathered a collection of "authentic" and fake reviews describing various restaurants.

Table 2. Sample Reviews

	Positive	Negative
Real	"I was very well pleased by my dining experience at [restaurant name]. Our server was very friendly and would always check to make sure we needed anything, she was very prompt and fast. The food was amazing, hot, and cooked properly. I have nothing negative to say about my recent dining experience at [Restaurant Name]. Would definitely come back and 10/10 would recommend to a friend."	"While at [restaurant name], my family and I had a really poor experience. To start the night, the staff was extremely slow at finding us a place to sit even though they were not at all busy, the waiter was extremely rude and almost never came to our table to ask how we were doing, or if we wanted a refill, ect. On top of that, my mom actually found an actual worm in her noodles. This experience was by far one of the worst I have had at a restaurant."
Fake	"This place was so cool! I walked into a white, clean, brand new thai place. I order ice cream off of a very detailed and expansive menu. For the low price of seven dollars I was able to watch the workers craft a gigantic cookie dough ice cream Sunday. This ice cream could be compared to cold stone but with the unique twist that they scrape the ice cream into rolls. I loved this place and cannot wait to go back."	"I would not recommend anyone to go to [restaurant name]. Their service was absolutely awful, it took 15 minutes before someone even noticed that we were even there and offered to help us. The person that helped us did not know anything about the menu and the products overall. The food itself was mediocre at best, nothing to rush back too. In addition, the restaurant itself was not clean and was littered with trash. Overall, a negative experience."

Notes. The table contains actual reviews written by pretest participants and utilized in the main experiments. Restaurant names were concealed, but the text was otherwise not altered.

Authentic reviews were defined as "based on the reviewer's own actual experience with a restaurant," and fake reviews were defined as "intended to be realistic, persuasive and believable—convincing readers that it is based on the reviewer's actual experience (even though in reality it is not)." Participants were told that a fake review can be "either positive (written to help the restaurant itself) or negative (written to help a competitor)."

Participants were then asked to read and evaluate 20 reviews, one at a time. Before beginning the task, they were given explicit base rate information; they learned that 10 of the reviews (50%) were known to be authentic and that 10 of the reviews (50%) were known to be fake. The 20 reviews consisted of 5 reviews drawn randomly from each of the four veracity × valence conditions, and the order of the reviews was randomized. After reading each review, participants responded to a simple single-item measure of veracity ("In your opinion, is this review authentic or fake?") with two response options ("authentic," "fake").

Results

On average, participants classified 12.35 (61.77%) reviews as authentic. This proportion was substantially and reliably greater than 50% (standard deviation (SD) = 0.11, $t(112) = 10.98$, $p < 0.001$), indicating the presence of a truth bias.[4] Thus, Hypothesis 1 was supported.

We next examined detection accuracy separately for real and fake reviews. On average, participants classified 10.53 (52.65%) reviews correctly, a proportion that was significantly but only slightly better than the chance level of 50% (SD = 0.10, $t(112) = 2.94$, $p = 0.004$). To test for the presence of a veracity effect, we conducted a repeated-measures analysis of variance (ANOVA) with classification accuracy as the outcome variable and with review veracity and valence as independent factors. Results revealed a significant main effect of veracity such that accuracy was substantially greater for *real* reviews than *fake* reviews (mean (M) = 64.4% versus 40.9%, $t(112) = 11.19$, $p < 0.001$).[5] Thus, Hypothesis 2 was supported.

To examine the difference in truth bias for *positive* versus *negative* reviews, we conducted a repeated-measures ANOVA with review valence as a within-subject factor and truth bias as the dependent variable. For a given participant and valence, we measured truth bias by calculating the percentage of reviews (of that valence) classified as real. Results revealed that the truth bias was stronger for *negative* reviews than *positive* reviews ($M = 67.1\%$ versus 56.5%, $t(112) = 3.79$, $p < 0.001$). Thus, Hypothesis 3 was supported.

Discussion

Study 1 provided initial evidence of systematic truth-lie detection tendencies in an online review setting. Despite being aware that only half of the reviews were real, participants tended to classify more than half of the reviews that way (a truth bias). As such, they tended to be more accurate in classifying real reviews than fake reviews (a veracity effect). Importantly, however, and consistent with our argument that consumers are less likely to project deception motives onto reviewers who are critical, the truth bias was weaker for positive reviews than negative reviews (a valence effect).

Two related concerns impact the ability to draw firm conclusions from these results. First, participants may have misremembered the base rate of real reviews

provided in the instructions (50%). If they incorrectly remembered a higher base rate, then it would be perfectly reasonable to classify the majority of reviews as real. Second, participants were not allowed to "go back" during the task. Thus, they may have misremembered how many reviews they had already classified as real and fake, or they may have wanted to change their earlier responses but been unable to do so. Each of these procedure-related confounds could produce results consistent with truth bias, and we designed the next experiment to address them.

Study 2

Procedure

We recruited 185 undergraduate students (96 female) who received course credit for participation. Study materials are provided in Online Appendix C. The cover story and procedure were similar to that of Study 1, with two major exceptions. First, all 20 reviews appeared on a single screen. As a result, participants could see (and make changes to) their prior real/fake classifications as they proceeded through the task, and they could easily tally the number of reviews that they had classified in each category. Second, the final step included an attention check to ensure that participants remembered the base rate: "According to the instructions, how many of the 20 reviews were actually authentic?"

Results

Participants spent approximately 10 minutes evaluating the reviews (29.20 seconds per review), suggesting that they expended a reasonable level of effort. Memory for the base rate was highly accurate. On average, participants remembered being told that 10.01 of the 20 reviews were authentic (SD = 1.55). Nonetheless, they chose to classify an average of 11.38 (56.92%) reviews as authentic. This proportion is significantly above 50% (SD = 0.10, $t(184) = 9.23$, $p < 0.001$) and represents a truth bias (Hypothesis 1) similar to that observed in Study 1.

Participants classified 52.92% of the reviews accurately on average, representing performance that was significantly but only slightly better than chance (SD = 0.11, $t(184) = 3.75$, $p < 0.001$). Results of a repeated-measures ANOVA revealed evidence of a veracity effect (Hypothesis 2) similar to that observed in Study 1. Classification accuracy was substantially greater for *real* reviews than *fake* reviews ($M = 59.8\%$ versus 46.0%, $t(184) = 9.20$, $p < 0.001$). Finally and supporting Hypothesis 3, results of a repeated-measures ANOVA with truth bias as the dependent variable revealed a main effect of valence such that the truth bias was stronger for *negative* reviews than *positive* reviews ($M = 63.5\%$ versus 50.4%, $t(184) = 6.24$, $p < 0.001$).

Discussion

Study 2 provided additional support for Hypothesis 1–3 while alleviating two procedure-related confounds. The observed truth bias is especially striking given that participants were allowed to adjust their classifications dynamically (and on the same screen) to conform to the provided base rate. Seemingly, participants were so unconfident in their ability to identify fake reviews that they knowingly classified "too many" reviews as real.

As before, the observed truth bias was substantial for negative reviews but negligible for positive reviews, suggesting that negative review content is less likely to trigger suspicion.[6] However, it is possible that the asymmetry was driven by differences in review content rather than valence. For example, negative reviews might tend to be more specific or concrete than positive reviews, resulting in greater credibility and (in turn) greater willingness to accept them as true. To address this possibility, we designed the next study so that valence would no longer be useful for judging veracity. Specifically, we divided the classification task into two "blocks," one containing only positive reviews and the other containing only negative reviews. If the valence effect in the previous studies occurred because content characteristics of the negative reviews aroused less suspicion, then it should still occur under a blocked design. If the effect was instead driven by valence itself (as we propose), then it should be eliminated under a blocked design (see Table 3).

Study 3

Procedure

We recruited 71 undergraduate students (38 female).[7] Study materials are provided in Online Appendix D. The cover story and procedure were similar to Study 1, with one major difference and one minor difference. First, the positive reviews and negative reviews were not mixed together but were presented in two separate blocks. Participants were informed that within each block, half of the reviews were real, and the other half

Table 3. Study 3 Design Logic

Source of the valence effect	Influence of blocked design (separating positive and negative reviews)	Valence effect expected?
Review valence	Review valence can no longer be used to distinguish real and fake reviews	No
Content characteristics correlated with valence	Review content can still be used to distinguish real and fake reviews	Yes

were fake. The order of the two blocks was counterbalanced, and each review appeared on a separate screen. Second, the total number of reviews was expanded to 24 (12 reviews per block). Given the simplified block format, we deemed it unlikely that the additional reviews would cause fatigue.

Results

On average, participants classified 13.59 (56.63%) reviews as authentic, indicating the presence of a truth bias (SD = 0.08, $t(70) = 7.05$, $p < 0.001$) whose magnitude was similar to that of Studies 1 and 2. As in the earlier studies, overall classification accuracy was only slightly above chance levels (53.46%, SD = 0.10, $t(70) = 3.05$, $p = 0.003$). Results of a repeated-measures analysis of covariance (ANCOVA) with veracity and valence as categorical predictors and block order as a covariate again revealed evidence of a veracity effect such that classification accuracy was substantially higher for real reviews than fake reviews ($M = 60.1\%$ versus 46.8%, $t(70) = 7.00$, $p < 0.001$).

In contrast to the earlier studies but consistent with expectations, analyses did not reveal evidence of a valence effect. The magnitude of truth bias was almost identical for the positive and negative blocks: $M = 56.4\%$ versus 56.8%. In a repeated-measures ANCOVA with block order as a covariate, the valence effect did not approach significance ($F(1, 69) = 0.04$, $p = 0.9$).

Discussion

Supporting Hypotheses 1 and 2, Study 3 provided additional evidence for truth bias and veracity effects in the perception of online reviews. However, the valence effect in the prior studies was absent in Study 3, presumably because of the blocked presentation format that prevented valence from serving as a cue to veracity. Hence, the findings indirectly support Hypothesis 3 by suggesting that the valence effect in those studies was not driven by differences in review content.

Our final two studies further explored the mechanism underlying the valence effect. Both studies incorporated a "moderation-of-process" design, which is often recommended when a process is hard to measure but easy to manipulate (Spencer et al. 2005). In Study 4, we tested our assertion that the valence effect occurs because readers of positive reviews are more likely to project deception motives onto the reviewer (Hypothesis 4(a)). To do so, we manipulated the salience of deception motives to participants before they encountered the reviews. Logically, our theorized mechanism will be constrained when deception motives are already highly salient for both positive and negative reviews. Hence, an interaction of valence with the salience manipulation would support Hypothesis 4(a), whereas the absence of an interaction would suggest that a different mechanism(s) underlies the valence effect (see Table 4).

Study 4

Procedure

We recruited 161 undergraduate students (88 female) for the study. The design and procedure were similar to Study 1 (see Online Appendix E). The major exception was the inclusion of an additional variable—deception motive reminder—that was manipulated between subjects. Participants assigned to the *reminder-present* condition saw the following messages before every *positive* review and *negative* review, respectively: "Remember, positive reviews of a restaurant may be fake—created by the restaurant to benefit itself," and "Remember, negative reviews of a restaurant may be fake—created by a competitor to benefit itself." Subjects assigned to the *reminder-absent* condition did not see any reminders. To verify the effectiveness of the reminder manipulation, participants were asked at the end of the study whether they recalled seeing the messages.

Results

Twenty-six of the 80 participants in the *reminder-absent* condition incorrectly recalled seeing the reminders, and 7 of the 81 participants in the *reminder-present* condition did not recall seeing the reminders. Excluding these participants from the analyses did not qualitatively change the results, and the analyses reported below include the entire sample.

Preliminary analyses yielded results consistent with the prior studies. Overall classification accuracy was 50.75%, a proportion not significantly different from chance (SD = 0.11, $t(160) = 0.85$, $p = 0.4$). Consistent with a truth bias, participants classified substantially more than half the reviews as authentic (59.94%; SD = 0.11, $t(160) = 11.20$, $p < 0.001$). Consistent with a veracity effect, analysis by mixed ANOVA revealed

Table 4. Study 4 Design Logic

Mechanism underlying the valence effect	Impact of exogenous increase in deception motive salience	Interaction of valence with motive salience expected?
Projection of deception motives	Increase in deception motives will be greater for negative reviews	Yes
Not projection of deception motives	Increase in deception motives will be similar for positive and negative reviews	No

that accuracy was higher for *real* reviews than *fake* reviews ($M = 59.4\%$ versus 39.5%, $t(160) = 7.11$, $p < 0.001$).

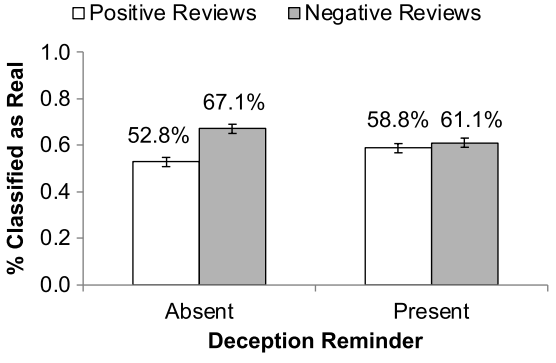
Consistent with a valence effect, analysis by mixed ANOVA with truth bias as the dependent variable revealed a main effect of valence such that the truth bias was greater for *negative* reviews than for *positive* reviews ($M = 64.1\%$ versus 55.8%, $t(160) = 3.65$, $p < 0.001$). Most importantly, however, the analysis also revealed a significant interaction of valence with deception reminder ($F(1, 159) = 7.11$, $p = 0.008$) (see Figure 1). Follow-up comparisons revealed that when the reminder was *absent*, the truth bias was significantly greater for *negative* reviews than *positive* reviews ($M = 67.1\%$ versus 52.8%, $t(160) = 4.50$, $p < 0.001$). When the reminder was *present*, however, the truth bias was similar for *negative* and *positive* reviews ($M = 61.1\%$ versus 58.8%, $t(160) = 0.72$, $p = 0.5$).

Discussion

Study 4 provided further evidence of a truth bias, veracity effect, and valence effect in judgments of review veracity, supporting Hypotheses 1–3. Moreover, the valence effect disappeared when participants were reminded that both positive and negative reviewers may have reasons to be deceptive. Supporting Hypothesis 4(a), this result suggests that the valence effect is driven by a tendency to project greater deception motives onto positive reviewers than negative reviewers.[8]

As stated in Hypothesis 4(b), our theorizing asserts that the projection of deception motives is a largely automatic process. However, this assertion was not tested in Studies 1–4, and we acknowledge the possibility of other mechanisms that involve substantial deliberation. For example, the truth bias may be stronger for negative reviews than positive reviews because the former contain fewer suspicion-activating cues in their textual content. If so, then readers who are sufficiently involved and capable of "picking up on" those cues would logically be more skeptical of positive reviews.

Figure 1. Interaction of Deception Reminder and Review Valence on Truth Bias (Study 4)



Note. Error bars represent standard errors.

Alternatively, the valence effect could be the result of a seemingly rational "cost-benefit" analysis. If readers determine that the potential cost of believing fake negative reviews (e.g., "missing out" on a good restaurant) is less than the potential cost of believing fake positive reviews (e.g., wasting time and money at an inferior restaurant), then they may conclude that it is reasonable to be less skeptical of negative reviews. In contrast to our accessibility-based mechanism, these alternative mechanisms entail a considerable degree of cognitive deliberation.

Given that it would be difficult to directly measure the degree of deliberation expended by review readers, our final study again employed a moderation-of-process design. Specifically, we manipulated the instructions given to participants so that in some conditions, they emphasized either "intuition" or "deliberation." Our main interest was the extent to which thinking style served to moderate the valence effect (see Table 5). Under the widely accepted "dual-process" view of human cognition (Sloman 1996, Kahneman 2011, Evans and Stanovich 2013), intuitive judgments represent automatic, heuristic-driven defaults that are overridden by more effortful, intentional deliberation only when there is sufficient motivation and ability to do so. According to our theorizing, review readers are instinctively more likely to associate positive reviews than negative reviews with deception motives; although capable of overriding this association if necessary (e.g., the "blocked format" of Study 3), they typically see no reason to do so. Therefore, neither "deliberation" nor "intuition" instructions should alter the process, and the differing instructions should have little effect on judgments. If, however, the valence effect is driven by a deliberate process, then it should be substantially disrupted by instructions emphasizing reliance on "intuition."

Study 5

Procedure

We recruited 168 undergraduate students (104 female). Study materials are provided in Online Appendix G. The design and procedure were similar to Study 1, with two important exceptions. First, the design included an additional variable, thinking style, which was manipulated between subjects at three levels: *deliberation*, *intuition*, and *control*. In the instructions preceding the classification task, participants in the *deliberation* condition were asked to "rely on your careful analysis and deliberation," "think carefully about reasons that the reviews may be authentic or fake," and "avoid relying on their intuition or first impressions." They were told that they "will be asked to explain how you arrived at your decisions" later in the study. Participants in the *intuition* condition were asked to "rely on your intuition and first impressions," "follow your

Table 5. Study 5 Design Logic

Mechanism underlying the valence effect	Influence of thinking-style manipulation	Interaction of valence with thinking style expected?
Automatic	The influence of thinking-style manipulation will be limited	No
Deliberative	The manipulation that emphasizes "intuition" but not "deliberation" will disrupt the deliberative mechanism	Yes

instincts about whether the reviews are authentic or fake," and "avoid careful analysis or deliberation." They were told that they "will not be asked to explain how you arrived at your decisions." Participants in the *control* condition did not receive any additional instructions. A thinking-style manipulation check was administered at the end of the procedure (see below). The second notable change was that the total number of reviews was reduced to 12 to keep the length manageable for all conditions.

Results

The thinking-style manipulation check asked participants to indicate how much they relied on deliberation or intuition as they evaluated the reviews using four items on seven-point Likert scales (e.g., "I thought carefully about reasons the review might be authentic or fake," and "I relied mainly on my first impressions and 'gut feelings'"). The items were scored so that a higher value indicated greater reliance on deliberation. Examination revealed that reliance on deliberation was higher in the *deliberation* condition than the *control* condition, which was, in turn, higher than the *intuition* condition ($M = 4.21$ versus 3.75 versus 3.18, $F(2, 165) = 28.17$, p-values < 0.001).

Preliminary analyses yielded results similar to those of the prior studies. Overall classification accuracy was 52.03%, representing performance that was not significantly better than chance (SD = 0.14, $t(167) = 1.87$, $p = 0.062$). Consistent with a truth bias, participants classified the majority of reviews as authentic (57.89%; SD = 0.10, $t(167) = 10.02$, $p < 0.001$). Consistent with a veracity effect, a mixed ANOVA with veracity and valence as within-subjects factors and thinking style as a between-subjects factor revealed that accuracy was higher for *real* reviews than *fake* reviews ($M = 59.9\%$ versus 44.1%, $t(167) = 9.88$, $p < 0.001$).

Consistent with a valence effect, results of a mixed ANOVA with truth bias as the dependent variable revealed a main effect of valence such that the truth bias was substantial for *negative* reviews but negligible for *positive* reviews ($M = 66.1\%$ versus 49.7%, $t(167) = 6.83$, $p < 0.001$). Most importantly, the valence effect was robust across thinking styles; neither the main effect of thinking style nor its interaction with valence approached significance (p-values = 0.7 and 0.5).

Discussion

Supporting Hypotheses 1–3 and replicating results from the prior studies, participants in Study 5 tended to classify reviews as real rather than fake, especially when the reviews were negative, and they were more accurate in classifying real reviews than fake reviews. Furthermore, the magnitude of truth bias across positive and negative reviews was similar whether participants based their judgments on intuition or on careful analysis and deliberation. Supporting Hypothesis 4(b), this result suggests that the valence effect is driven by a largely automatic process.

General Discussion

The research presented here represents one of the first attempts to investigate how consumers discriminate between "real" and "fake" online reviews (see also Plotkina et al. 2020). Building on truth-default theory and prior literature in deception motives (Gilbert 1991, Levine et al. 2016, Levine 2019), we propose that consumers have a general tendency to assume that reviews are true (a truth bias) and are more accurate in detecting real reviews than fake reviews (a veracity effect). Furthermore, we propose that the truth bias will be weaker for positive reviews than negative reviews (a valence effect) because the former are more likely to automatically activate suspicion of reviewer motives. To investigate these proposals, we generated a pool of reviews with known veracity and then conducted five experiments in which participants were asked to judge the veracity of reviews from the pool. Results of the experiments produced converging evidence for our proposals (see Table 6 for a summary).

Theoretical Implications

The problem of "fake" reviews has received growing and interdisciplinary scholarly attention. In our view, the present research makes two primary contributions to this important and emerging area. The first of these contributions is the adoption of a novel perspective that is *focused on consumer perceptions*. A number of prior investigations have documented the motivations of businesses to commit review fraud, the economic consequences of fake reviews, etc. (e.g., Mayzlin et al. 2014, Lappas et al. 2016, Luca and Zervas 2016), whereas other investigations have used technical methods to explore how various platforms can improve the

Table 6. Summary of Results

	Study focus	Hypothesis 1: Truth bias	Hypothesis 2: Veracity effect	Hypothesis 3: Valence effect	Hypothesis 4(a): Mediation of valence effect by deception motives	Hypothesis 4(b): Automaticity of valence effect
Study 1	Basic effects	Supported	Supported	Supported		
Study 2	Confounds	Supported	Supported	Supported		
Study 3	Confounds			Supported		
Study 4	Mechanisms				Supported	
Study 5	Mechanisms					Supported

automated detection of fake reviews (e.g., Kumar et al. 2018, Shan et al. 2021, Luo et al. 2023). However, less is known about how consumers themselves approach the problem, their lay theories regarding its magnitude and causes, or their aptitude for distinguishing real from fake reviews. A better understanding of the consumer perspective is critical, as it is consumers themselves who are the target of review manipulation. Their beliefs about review veracity play an important role not only in their immediate purchase decisions, but also in their evolving satisfaction and trust in e-commerce platforms.

Given that the proliferation of fake reviews has been widely reported, some researchers have speculated that consumers may simply discount the credibility of reviews in general (Mayzlin et al. 2014). In contrast to this reasonable speculation but consistent with "truth-default" theory, our experiments revealed a consistent and robust tendency for consumers to accept reviews as real (a truth bias). This tendency was observed even when the base rate of real versus fake reviews was provided (all studies) and even when participants were able to revise their prior judgments to align with those base rates (Study 2). Whereas truth-default theory has been applied extensively in high-stakes contexts that involve the evaluation of detailed information from specific individuals (job interviews, police interrogations, courtroom testimony, etc.; see Bond and DePaulo 2006), our findings demonstrate its applicability in the unique setting of online reviews—a (typically) low-stakes context that involves the evaluation of often-vague, asynchronous information provided by unknown reviewers.

Although we use the term "bias" to be consistent with prior literature, it is important to note that the truth bias may often be efficient and functional (see the earlier discussion). Nonetheless, being "fooled" by fake reviews can have damaging and long-lasting consequences; when consumers are highly risk averse, believing a single "fake" negative review might lead them to eliminate even the most promising option from consideration. Given that the prevalence of fake reviews varies considerably across platforms and product domains (Wu et al. 2020), a truth bias is clearly more problematic in platforms and domains where fake reviews are more common. To the extent that consumers are both overly trusting of reviews and poor at assessing their veracity, our findings highlight the potential value of outside assistance (see below).

The second primary contribution of our research is a deeper understanding of the role played by valence in consumer processing of online reviews. Valence is a fundamental characteristic of reviews that has no direct analog in the traditional study of truth-lie detection, and the current research represents (to the best of our knowledge) the first examination of its role in veracity judgments. By identifying valence as an important moderator of truth bias in the online review context, our findings extend understanding of the multiple ways that review valence can impact consumer perceptions and behavior (e.g., Yin et al. 2016; Liu et al. 2019; Lei et al. 2023, 2025). Moreover, our findings suggest that there exists a striking contrast between reality and perception; although real-world evidence indicates that the percentage of fake negative reviews is higher than that of fake positive reviews (Anderson and Simester 2014), participants in our studies were consistently and considerably more suspicious of positive reviews (in Studies 2 and 5, participants showed almost no truth bias at all for positive reviews). Beyond the review setting, valence is known to have a substantial impact on phenomena such as virality and the spread of fake news (Wang et al. 2022). Our findings suggest that its impact on veracity judgments may play a role in those phenomena.

Furthermore, our findings offer valuable insights into potential mechanisms underlying the valence effect. Relevant prior work has documented the role played by automatic, unconscious processes in helpfulness judgments of online reviews and the utilization of reviews for decision making (e.g., Yin et al. 2016, 2021; Lei et al. 2023, 2025). Our research (and Studies 4 and 5 in particular) complements this prior work by suggesting that judgments of review veracity are themselves influenced by an automatic process, in which positive valence by itself can trigger suspicion of reviewer motives. Although effortful deliberation undoubtedly plays an important role in veracity

judgments, our findings suggest that the judgments are more spontaneous than commonly assumed.

Practical Implications

As a result of mainstream media coverage, increased government scrutiny, and personal experience, consumers are increasingly aware that substantial portions of online reviews are fake. This increased awareness creates challenges for review platforms seeking to establish and maintain consumer trust. One important implication of our findings is that increased distrust of reviews and platforms in general may not translate to distrust of individual reviews; instead, the truth bias observed in our studies suggests that consumers are overly willing to accept the veracity of the reviews that they encounter. Compounded by the problem of low classification accuracy (especially for fake reviews), the presence of truth bias reinforces the need for platforms to identify and deal with fake reviews proactively. Among the myriad approaches currently employed, some approaches rely on readers themselves (e.g., Google Reviews allows users to "report" suspected fakes as either "spam" or "conflicts of interest"). Our findings strongly suggest that such reporting is unlikely to be effective and should be supplemented with other approaches.

Having identified probable fake reviews, many platforms opt to simply remove them. To the extent that readers are not generally capable of distinguishing review pools with "few" versus "many" fake reviews, however, our findings suggest that simple removal may do little to meaningfully enhance reader trust. To that end, a better approach might be to vividly signal the presence of protective mechanisms through warning labels, flags, "fact-checker" badges, etc. (for a similar recommendation based on different reasoning, see Ananthakrishnan et al. 2020).

Our findings regarding valence asymmetry have straightforward implications for the development and calibration of tools for detecting fake reviews. Over and above existing evidence that negative reviews are more sought after and more impactful than positive reviews (Chevalier and Mayzlin 2006, Lei et al. 2023), the fact that truth bias is more pronounced for negative reviews suggests that for malicious actors in the marketplace, fake negative reviews (of competitors) are particularly effective. Holding constant other considerations, therefore, it is reasonable to prioritize detecting fake negative reviews over detecting fake positive reviews. Along similar lines, results of our third study suggest that presentation formats that involve distinct "blocks" of positive and negative reviews can produce unintended effects. In particular, platforms that allow (or default to) valence-based review sorting should be aware that doing so may reduce user suspicion of positive versus negative reviews.

Future Research

The pervasiveness of truth bias in consumer review processing is likely to be impacted by a variety of factors related to review content and situational context. For example, future research might explore whether and when markers of poor communication (typos, grammatical mistakes, etc.), extremely short reviews, or overly "one-sided" reviews generate reader suspicion that reduces or eliminates the truth bias. Taking a more nuanced approach, scholars might compile the cues identified most frequently in the truth-lie detection literature (DePaulo et al. 2003) and then investigate the subset of cues most relevant to a review context.

In all of our studies, the base rate of real reviews was fixed at 50%. This base rate offers many advantages. It is consistent with common practice in the truth-lie detection literature (Bond and DePaulo 2006), is easy for participants to understand and remember, and produces the most conservative possible test of truth bias. However, real-world base rates tend to be substantially higher than 50%, and they also vary substantially across different platforms, products, etc. Thus, a straightforward opportunity exists to extend our investigation to a variety of different settings and a range of different base rates. Along similar lines, all of our studies utilized laboratory experiments and participants who were not actually engaged in the shopping process. We encourage future researchers to explore the ecological validity of our findings by use of incentive-compatible designs, field experiments, and observational methods.

When a delay exists between the processing of online reviews and downstream decisions, memory for review information becomes an important consideration (Lei et al. 2022, 2023). In many cases, the "recalled" rate of real versus fake reviews may be more important than the rate that was inferred at the time the reviews were initially encountered. Although memory and recall are out of the scope of the current investigation, they are worthy of exploration.

Future research should consider the broader consequences of truth-lie detection from a consumer perspective. Assuming, for example, that readers are not initially suspicious of the reviews that they consult, what happens if they become so? Will that suspicion lead them to disregard specific reviews, consult a larger set of reviews, or switch to an alternative review platform altogether? How will it impact their attitudes toward the product, the seller, and the review platform? Answers to such questions have important implications for all parties involved.

Conclusion

As consumers rely more and more on online reviews to inform their marketplace decisions, fake reviews present an increasingly insidious threat. Counteracting this threat requires a greater understanding of how readers perceive the veracity of online reviews, and our research provides initial steps in that direction. By demonstrating the prevalence of truth bias in the perception of online reviews and revealing the novel role of review valence, our findings deepen understanding of an overlooked area in the online review literature and lay the groundwork for additional exploration.

Acknowledgments

The authors sincerely appreciate the senior editor, Choon-Ling Sia, the associate editor, David (Jingjun) Xu, and three anonymous reviewers for their constructive guidance and insightful suggestions throughout the review process. The authors are also grateful to Marius Florin Niculescu, Lizhen Xu, Michael Smith, Adrian Gardiner, Hao Hu, and Katsiaryna Siamionava for their help in recruiting experiment participants.

Endnotes

[1] Because the veracity effect is intrinsically connected to truth bias, it should also be lessened for positive reviews. This was indeed the case across all of the experiments reported below. Given our main interest in truth bias rather than detection accuracy, however, we do not present a formal hypothesis or address this issue further.

[2] In a second stage of the pretest, participants were asked to write a second review that was "opposite" of the first review in valence and veracity (see Online Appendix A). However, examination of the second reviews suggested that they were highly contaminated by the act of generating the first reviews. As a conservative precaution, we retained only the first reviews for the pool.

[3] All studies were conducted at the same university. Participants in different studies were recruited from either different courses or different semesters, minimizing the likelihood of duplicates.

[4] Lie detection researchers sometimes estimate parameters based on signal detection theory. The parameter d' represents "sensitivity" (accuracy), and the parameter c represents "response bias" (tendency to favor one response over the other). When our data are analyzed using this approach, all results remain qualitatively the same. We focus on direct measures because they are easier to understand and correlate highly with the signal detection parameters (Bond and DePaulo 2006).

[5] Results also revealed a significant valence × veracity interaction effect ($F(1, 112) = 14.21$, $p < 0.001$). Follow-up pair-wise comparisons revealed that the veracity effect was stronger for *negative* reviews ($M = 70.3\%$ versus 36.1%, $t(112) = 10.06$, $p < 0.001$) than for *positive* reviews ($M = 58.6\%$ versus 45.7%, $t(112) = 3.58$, $p < 0.001$). Across all studies, the interaction of valence and the veracity effect was similar in significance and direction to the interaction of valence and truth bias. Because we did not hypothesize this interaction, we do not address it further.

[6] To explore this possibility, we included two questions near the end of the study: "Reviews that are more negative are more likely to be authentic," and "Reviews that are more unfavorable toward the restaurant are more likely to be authentic" (1 = *Strongly Disagree*, 7 = *Strongly Agree*). Mean responses were both significantly above the scale midpoint ($M = 4.37$ and 4.35, SD = 1.47 and 1.40, p-values < 0.01), suggesting that participants were indeed less suspicious of negative reviews.

[7] The sample was smaller than that of the prior studies because of course enrollment. However, the within-subjects design ensured adequate power. Analysis using G*Power (Faul et al. 2009) indicated that a sample size of 56 was sufficient for 95% power to detect small to medium effects ($f = 0.2$) with an alpha level of 0.05.

[8] A natural follow-up question is as follows. What kinds of deceptive motives did participants project? To address this question, we included two questions at the end of the study: "Consider the case of reviewers who write fake [positive or negative] reviews. In your opinion, how common are each of the following motives for writing fake [positive or negative] reviews?" (1 = *Not At All Common*, 7 = *Common*). Participants rated 10 motives adapted from Levine et al. (2016). The results are summarized in Tables F1 and F2 in Online Appendix F. The motives rated most common for fake positive reviews were "economic advantage" and "impression management," whereas the motives rated most common for fake negative reviews were "malice" and "economic advantage."

References

Ananthakrishnan UM, Li B, Smith MD (2020) A tangled web: Should online review portals display fraudulent reviews? *Inform. Systems Res.* 31(3):950–971.

Anderson ET, Simester DI (2014) Reviews without a purchase: Low ratings, loyal customers, and deception. *J. Marketing Res.* 51(3): 249–269.

Babić Rosario A, Sotgiu F, De Valck K, Bijmolt THA (2016) The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *J. Marketing Res.* 53(3):297–318.

Bond CF, DePaulo BM (2006) Accuracy of deception judgments. *Personality Soc. Psych. Rev.* 10(3):214–234.

Bond CF, DePaulo BM (2008) Individual differences in judging deception: Accuracy and bias. *Psych. Bull.* 134(4):477–492.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Datassential (2019) Campus dining isn't what it used to be. Accessed May 6, 2025, https://datassential.com/resource/campus-dining-isnt-what-it-used-to-be/.

DePaulo BM, Charlton K, Cooper H, Lindsay JJ, Muhlenbruck L (1997) The accuracy-confidence correlation in the detection of deception. *Personality Soc. Psych. Rev.* 1(4):346–357.

DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H (2003) Cues to deception. *Psych. Bull.* 129(1):74–118.

Ekman P, O'Sullivan M (1991) Who can catch a liar? *Amer. Psychologist* 46(9):913–920.

Evans JSB, Stanovich KE (2013) Dual-process theories of higher cognition: Advancing the debate. *Perspect. Psych. Sci.* 8(3):223–241.

Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* 41(4):1149–1160.

Feldman JM, Lynch JG (1988) Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psych.* 73(3):421–435.

Forrest JA, Feldman RS (2000) Detecting deception and judge's involvement: Lower task involvement leads to better lie detection. *Personality Soc. Psych. Bull.* 26(1):118–125.

Gambetti A, Han Q (2023) Dissecting AI-generated fake reviews: Detection and analysis of GPT-based restaurant reviews on social media. *Internat. Conf. Inform. Systems* (ICIS, Hyderabad).

Gilbert DT (1991) How mental systems believe. *Amer. Psychologist* 46(2):107–119.

Gössling S, Hall CM, Andersson A-C (2018) The manager's dilemma: A conceptualization of online review manipulation strategies. *Current Issues Tourism* 21(5):484–503.

Halevy R, Shalvi S, Verschuere B (2014) Being honest about dishonesty: Correlating self-reports and actual lying. *Human Comm. Res.* 40(1):54–72.

Ismagilova E, Slade EL, Rana NP, Dwivedi YK (2020) The effect of electronic word of mouth communications on intention to buy: A meta-analysis. *Inform. Systems Frontiers* 22(5):1203–1226.

Kahneman D (2011) *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).

Kapner S (2019) Black Friday shoppers: Beware of fake five-star reviews. *Wall Street J.* (November 28), https://www.wsj.com/articles/black-friday-shoppers-beware-of-fake-five-star-reviews-11574937001.

Kumar N, Venugopal D, Qiu L, Kumar S (2018) Detecting review manipulation on online platforms with hierarchical supervised learning. *J. Management Inform. Systems* 35(1):350–380.

Lappas T, Sabnis G, Valkanas G (2016) The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Inform. Systems Res.* 27(4):940–961.

Lei Z, Yin D, Zhang H (2023) Positive or negative reviews? Consumers' selective exposure in seeking and evaluating online reviews. *J. Assoc. Inform. Systems* 24(4):1162–1183.

Lei Z, Yin D, Zhang H (2025) Deliberative or automatic: Disentangling the dual processes behind the persuasive power of online word-of-mouth. *MIS Quart.* 49(1):331–346.

Lei Z, Yin D, Mitra S, Zhang H (2022) Swayed by the reviews: Disentangling the effects of average ratings and individual reviews in online word-of-mouth. *Production Oper. Management.* 31(6): 2393–2411.

Levine TR (2014) Truth-default theory (TDT): A theory of human deception and deception detection. *J. Language Soc. Psych.* 33(4): 378–392.

Levine TR (2019) *Duped: Truth-Default Theory and the Social Science of Lying and Deception* (University of Alabama Press, Tuscaloosa, AL).

Levine TR, Kim RK, Blair JP (2010) (In)accuracy at detecting true and false confessions and denials: An initial test of a projected motive model of veracity judgments. *Human Comm. Res.* 36(1): 82–102.

Levine TR, Park HS, McCornack SA (1999) Accuracy in detecting truths and lies: Documenting the "veracity effect." *Comm. Monographs* 66(2):125–144.

Levine TR, Ali MV, Dean M, Abdulla RA, Garcia-Ruano K (2016) Toward a pan-cultural typology of deception motives. *J. Intercultural Comm. Res.* 45(1):1–12.

Liu X, Lee D, Srinivasan K (2019) Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *J. Marketing Res.* 56(6):918–943.

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Sci.* 62(12):3412–3427.

Luo J, Luo J, Nan G, Li D (2023) Fake review detection system for online e-commerce platforms: A supervised general mixed probability approach. *Decision Support Systems.* 175:114045.

Mann S, Vrij A, Bull R (2004) Detecting true lies: Police officers' ability to detect suspects' lies. *J. Appl. Psych.* 89(1):137–149.

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *Amer. Econom. Rev.* 104(8):2421–2455.

McCornack SA, Parks MR (1986) Deception detection and relationship development: The other side of trust. *Ann. Internat. Comm. Assoc.* 9(1):377–389.

Mukherjee A, Venkataraman V, Liu B, Glance N (2013) Fake review detection: Classification and analysis of real and pseudo reviews. Working paper, University of Illinois at Chicago, Chicago.

Ng KC, Ke PF, So MK, Tam KY (2023) Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach. *Production Oper. Management* 32(7):2101–2122.

Plotkina D, Munzel A, Pallud J (2020) Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews. *J. Bus. Res.* 109:511–523.

Qiu L, Pang J, Lim KH (2012) Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence. *Decision Support Systems* 54(1):631–643.

Reinhard M-A, Greifeneder R, Scharmach M (2013) Unconscious processes improve lie detection. *J. Personality Soc. Psych.* 105(5): 721–739.

Schoolov K (2020) Amazon is filled with fake reviews and it's getting harder to spot them. *CNBC* (September 6), https://www.cnbc.com/2020/09/06/amazon-reviews-thousands-are-fake-heres-how-to-spot-them.html.

Serota KB, Levine TR, Boster FJ (2010) The prevalence of lying in America: Three studies of self-reported lies. *Human Comm. Res.* 36(1):2–25.

Shan G, Zhou L, Zhang D (2021) From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems* 144:113513.

Simon HA (1956) Rational choice and the structure of the environment. *Psych. Rev.* 63(2):129–138.

Sloman SA (1996) The empirical case for two systems of reasoning. *Psych. Bull.* 119(1):3–22.

Song Y, Wang L, Zhang Z, Hikkerova L (2023) Do fake reviews promote consumers' purchase intention? *J. Bus. Res.* 164:113971.

Spencer SJ, Zanna MP, Fong GT (2005) Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *J. Personality Soc. Psych.* 89(6):845–851.

Sporer SL, Schwandt B (2006) Paraverbal indicators of deception: A meta-analytic synthesis. *Appl. Cognitive Psych.* 20(4):421–446.

Sporer SL, Schwandt B (2007) Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psych. Public Policy Law* 13(1):1–34.

Wang SA, Pang M-S, Pavlou PA (2022) Seeing is believing? How including a video in fake news influences users' reporting of fake news to social media platforms. *MIS Quart.* 46(3):1323–1354.

Wu Y, Ngai EWT, Wu P, Wu C (2020) Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems* 132:113280.

Wyer RS Jr (2018) The role of knowledge accessibility in cognition and behavior: Implications for consumer information processing. Haugtvedt CP, Herr PM, Kardes FR, eds. *Handbook of Consumer Psychology* (Routledge, New York), 31–76.

Yin D, Bond SD, Zhang H (2021) Anger in consumer reviews: Unhelpful but persuasive? *MIS Quart.* 45(3):1059–1086.

Yin D, Mitra S, Zhang H (2016) When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. *Inform. Systems Res.* 27(1):131–144.

Zhang D, Zhou L, Kehoe JL, Kilic IY (2016) What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *J. Management Inform. Systems* 33(2):456–481.

Zuckerman M, DePaulo BM, Rosenthal R (1981) Verbal and nonverbal communication of deception. *Adv. Experiment. Soc. Psych.* 14:1–59.

Zuckerman M, Koestner R, Colella MJ, Alton AO (1984) Anchoring in the detection of deception and leakage. *J. Personality Soc. Psych.* 47(2):301–311.