Red Flag! The Consequences of Alerting Consumers to Fake Reviews

AS THIS PAPER IS CURRENTLY IN THE REVIEW PROCESS, PLEASE DO NOT DISTRIBUTE WITHOUT CONSENT FROM THE AUTHORS

1.14.22

Jared Watson, New York University

Amna Kirmani, University of Maryland

ABSTRACT

This paper investigates consumer responses to fake review alerts, which are disclosures on a platform that it has identified and removed inauthentic positive reviews for a brand. While these alerts are intended to signal high veracity of a platform's content, we find that the alerts can backfire due to the content of the activated persuasion knowledge. The alert leads consumers to generate cognitions about a brand's expected dishonesty. At the same time, it leads to cognitions about a perceived ratings bias, informed in part by expectations of the brand's dishonesty, even when the alert informs consumers that the fake reviews have been removed. These two factors result in lowered brand evaluations. Using web scraped data, along with four experiments, we show that a fake review alert decreases evaluations, intentions, and choice, and this is simultaneously mediated by their beliefs about the brand and the reviews. These effects can be moderated by updating their lay beliefs about the brand or the impact of fake reviews.

Keywords: online reviews, fake reviews, everyday persuasion knowledge, bias

TripAdvisor has reasonable cause to believe that individuals or entities associated with or having an interest in this property may have interfered with traveler reviews. – TripAdvisor, 2020

Fake reviews are a problem for many companies and platforms, including Amazon, Yelp, and TripAdvisor. In 2020, Yelp removed 25% of reviews that violated its proprietary algorithm "designed to find and weed out reviews that are conflicts of interest, fake, solicited, low quality, or otherwise less reliable" (Yelp 2021). Similarly, Amazon announced the removal of over 20,000 suspicious reviews from only seven reviewers (Schiffer 2020). A substantial proportion of these removed reviews are fake, defined as those written with the intent to mislead or deceive. For instance, Luca and Zervas (2016) estimated that 16% of all Yelp reviews were fake. These fake reviews cause problems for all stakeholders. Honest brands are at a disadvantage when competitors solicit fake reviews to increase their own ratings. Platforms are at-risk for declining credibility and customer attrition. And consumers risk acting on unreliable or misleading information. Therefore, many platforms utilize proprietary detection software to identify and remove fake reviews.

As shown in the opening example, some platforms go a step further by informing consumers of businesses whom they suspect of engaging in egregious review manipulation. We term this notice a "fake review alert". For example, Yelp publishes fake review alerts on the page of specific businesses "to warn users when [they] find evidence of extreme attempts to manipulate a business's ratings and reviews" (Yelp 2019). In 2020 alone, over 1100 businesses received an alert for either incentivizing consumers to review their business or "suspicious" review activity (Yelp 2021). Although these fake review alerts are temporary (e.g., 90 days on Yelp), consumer response to the alerts is unclear. On the one hand, the alert should signal that the brand's reviews are being actively monitored, giving consumers confidence in the veracity of

the remaining reviews. On the other hand, the alert makes salient that someone was trying (albeit unsuccessfully) to manipulate consumers, leading consumers to question the brand and its reviews. These conflicting possible outcomes suggest the need to examine the downstream consequences of fake review alerts.

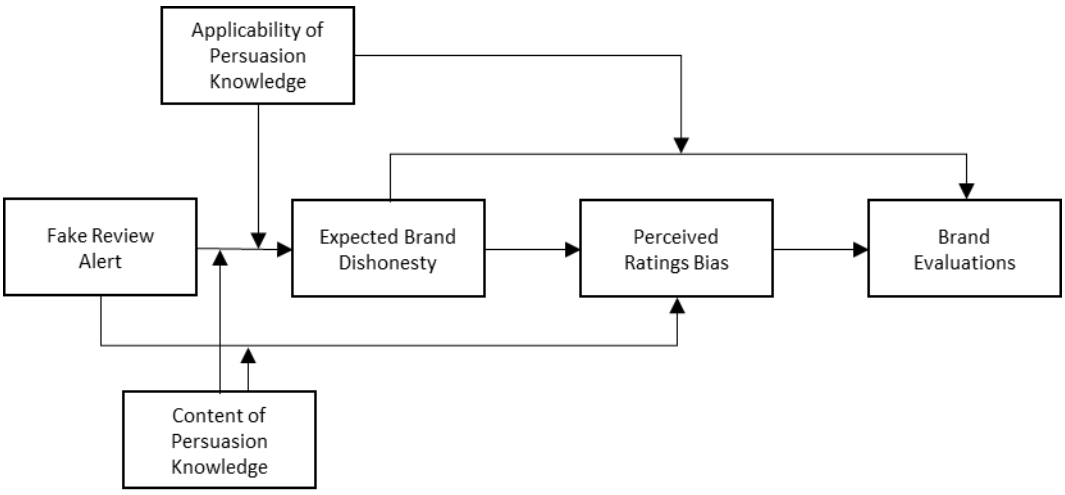
While fake negative reviews are sometimes used to hurt brands, most fake reviews have positive valence and are intended to boost the brand's rating (Lappas, Sabnis and Valkanas 2016; Luca and Zervas 2016). Although there have been some cases of alerts used to combat politically motivated reviews based on a business's action (often resulting in an influx of positive *and* negative fake reviews based on ideological beliefs of platform users), we focus on the more prevalent case where alerts inform consumers about fraudulent positive reviews.

We examine the effect of fake review alerts on brand evaluations. We propose that because a fake review alert informs consumers that someone was engaged in deception, it will activate persuasion knowledge (Friestad and Wright 1994), beliefs about marketers' motives and tactics as well as strategies for coping with such reviews. We investigate the nature and content of this persuasion knowledge as well as its effects on brand evaluations. Although there is a robust literature on the use of persuasion knowledge in advertising (Eisend and Tarrahi 2021), pricing (Hardesty, Bearden, and Carlson 2007), and sales (Campbell and Kirmani 2000), there is little work on persuasion knowledge in the context of product reviews. Moreover, prior literature largely examines the activation of persuasion knowledge through inferences of marketers' ulterior motives (Campbell and Kirmani 2000) or manipulative intent (e.g., Jain and Posavac 2004). In contrast, we examine consumers' lay beliefs about persuasion, which Friestad and Wright (1995) refer to as *everyday persuasion knowledge*. Everyday persuasion knowledge consists of consumers' beliefs (accurate or inaccurate) about the persuasion process (i.e., causal

effects, relationships, and effectiveness of tactics). In our context, it refers to beliefs about the brands that solicit fake reviews as well as how fake reviews may affect the brand. These beliefs help consumers cope with persuasion.

Figure 1 shows our conceptual model. We propose that the presence (vs. absence) of a fake review alert activates persuasion knowledge in the form of suspicion about the veracity of reviews in general as well as lay beliefs about how the reviews affect brand evaluations. We examine two types of lay beliefs: 1) beliefs about how deceptive brands will act in the future (i.e., that "cheaters will continue to cheat"); and 2) beliefs about how fake reviews affect the average product rating (i.e., "fake reviews distort product ratings"). Both these lay beliefs negatively affect brand evaluations (i.e., perceptions of the true product rating, intentions, and choice). In addition, we propose two theoretically derived moderators of this process that change either the content or the applicability of persuasion knowledge. First, we suggest that changing the content of everyday persuasion knowledge will attenuate the effect of an alert on brand evaluations. Specifically, priming consumers with information that suggests that fake reviews do not distort product ratings is likely to change their lay beliefs about the impact of fake reviews, leading to lower perceptions of ratings bias and more favorable brand evaluations. Second, changing the applicability of everyday persuasion knowledge will also increase brand evaluations. In particular, when the brand gets a new owner, fake reviews posted by the previous owner cannot be attributed to the new owner, making the alert information less relevant to expected future dishonesty. In this case, consumers are less likely to infer that the brand is likely to less in the future, leading to higher brand evaluations.

Figure 1: The Effect of a Fake Review Alert on Brand Evaluations



We test these predictions using both primary data web scraped from Yelp as well as four experiments that examine different aspects of the conceptual model. The paper makes both theoretical and substantive contributions. From a theoretical perspective, we enrich our understanding of everyday persuasion knowledge by investigating the content and applicability of beliefs about the effects of fake reviews. The context of fake reviews is somewhat different from the typical contexts of advertising (e.g., flattery, incomplete comparisons, sponsored content) that have been studied in the persuasion knowledge literature (Wilson, Darke and Sengupta 2021). Most of prior work on persuasion knowledge concerns potentially manipulative, yet legal, persuasion tactics, such as flattery, incomplete comparisons, and the use of celebrity endorsers. In contrast, fake reviews are illegal, so the attributions to the brand should be extremely negative and may lead to different outcomes than those in prior work. From a substantive perspective, we outline the effect of fake review alerts on consumer responses and suggest how to attenuate these negative effects.

CONCEPTUAL FRAMEWORK

Extant literature on fake reviews investigates their detection and effects. In computer science, much of the literature has concerned the development of algorithms to identify fake reviews based on syntactical cues (Akoglu, Chandy, and Faloutsos 2013; Feng, Banerjee, and Choi 2012; Mukherjee et al. 2013; Ott, Cardie, and Hancock 2013). This literature identifies the syntactical (e.g., excessive capitalization or repetitive punctuation) and contextual markers (e.g., IP address, review frequency, etc.) that indicate the likelihood of a review's being fake. In the field of marketing, research has investigated the brand characteristics that predict a higher presence of fake reviews. For instance, independent (vs. chain) brands are more likely to leave fake positive reviews for themselves and fake negative reviews for their competitors when the platform does not require consumer verification (Mayzlin, Dover, and Chevalier 2014). Moreover, these effects are often exacerbated when the independent brand has a weak reputation (Luca and Zervas 2016). Finally, Lappas et al. (2016) show that positive fake reviews increase the visibility of the brand, making it more likely to reach consumers' consideration sets. While these streams of literature have focused on the identification of fake reviews and the brands that solicit them, there is little work on how consumers respond to fake review alerts. Next, we develop theory about consumer responses to alerts as a function of persuasion knowledge.

Persuasion Knowledge

Persuasion knowledge (Friestad and Wright 1994) is a multi-dimensional construct consisting of beliefs about different variables and relationships in the persuasion process. Prior research has extensively examined certain aspects of persuasion knowledge, such as consumers' beliefs about marketers' ulterior motives (Campbell and Kirmani 2000), the manipulative intent

of the agent or tactic (e.g., Campbell 1995; Jain and Posavac 2004), or strategies for coping with persuasion (Kirmani and Campbell 2004). Less attention has been paid to the content of everyday persuasion knowledge (Friestad and Wright 1995). Everyday persuasion knowledge could include beliefs that celebrity endorsers increase attention (Boush, Friestad and Rose 1994) or that eliciting emotions is an effective advertising tactic (Friestad and Wright 1995).

A fake review alert informs consumers that a brand has been caught in an illegal action; and that the platform has protected consumers from this intended deception by removing the problematic reviews. In general, consumers see reviews as a source of information for gathering purchase-related details, learning how to use a product, or confirming one's impressions (Hennig-Thurau, Walsh, and Walsh 2003). By making salient that fake reviews exist, a fake review alert activates cognitions that the brand might be deceptive, and their reviews might be manipulative rather than informative. Darke and Ritchie (2007) found that general distrust about advertising can result in biased processing of ads from the same or different sources. Importantly, this defensive processing occurs only when consumers feel personally fooled by the advertisement; judgments remain unbiased when consumers are not personally deceived. In our context, we find that knowledge of the mere *attempt* to deceive consumers activates persuasion knowledge. This vigilance towards the brand, who is perceived to be the source of the manipulation attempt, spreads to other areas, like the product reviews for the brand. Even though the platform ascertains that the remaining reviews are credible, consumers' persuasion knowledge generates skepticism about the veracity of these reviews. Thus, in evaluating authentic content as deceptive, their persuasion knowledge may be misapplied. These lay beliefs about the brand can be considered the content of persuasion knowledge, otherwise known as everyday persuasion knowledge (Friestad and Wright 1995)

Everyday persuasion knowledge

When consumers encounter a fake review alert, they must determine how they will cope with this attempt at persuasion, as deception is a persuasion tactic (Rule, Bisanze and Kohn 1985). Relevant persuasion knowledge in this context will consist of their knowledge about how persuasion in general works as well as their knowledge about how fake reviews may affect the information at-hand. We propose that two types of relevant beliefs comprise everyday persuasion knowledge: beliefs about the brand and beliefs about fake reviews.

Beliefs about the brand have to do with the fact that the brand has behaved deceptively and, thus, is likely to behave dishonestly in the future. According to attribution theory (Kelley and Michela 1980), consumers are likely to blame the brand for the fraudulent reviews, as it is the entity that stands to benefit the most from fake positive reviews. If the brand has cheated by posting fake reviews, consumers will infer that the brand is a cheater. According to Skowronski and Carlston (1987), individuals make negative dispositional inferences from behavior in the realm of morality judgments. A single dishonest act leads to attributions that the brand is fundamentally dishonest (Reeder and Brewer 1979). This means that consumers are likely to apply their lay belief that "cheaters will continue to cheat" to form expectations about whether the brand will act deceptively in the future. Thus, the presence (versus absence) of a fake review alert increases expectations of brand dishonesty, leading to lower brand evaluations. In figure 1, this represents the direct link from the alert to expected brand dishonesty as well as from expected dishonesty to brand evaluations.

The second type of everyday persuasion knowledge that consumers will use to cope with a fake review alert has to do with their lay beliefs about how fake reviews affect product ratings. Prior research shows that consumers rely on the average product rating to form brand evaluations (Watson, Ghosh, and Trusov 2018), and these ratings are a function of the available reviews. In fact, consumers cite the average product rating as one of the most important cues in online reviews (BrightLocal 2020), and these ratings are highly predictive of purchase behaviors (Chevalier and Mayzlin 2006). Since the alert discloses that fake reviews have been detected and removed, consumers should trust the average product rating. However, consumers may be uncertain about whether all fake reviews were removed, leading to the perception that the average product rating is biased. This belief that "fake reviews distort the average product rating" stems from simple arithmetic: adding more positive numbers to an average increases the average. Moreover, it is consistent with assumptions made in prior research on fake reviews. For instance, Zhao et al. (2013) argue that the effect of fake reviews is to increase uncertainty about product quality (i.e., the average product rating), while Zappas et al. (2016) suggest that fake reviews increase the brand's rating, making it more likely to enter the consumer's consideration set. We are proposing that these beliefs are a viable element of everyday persuasion knowledge. Perceptions that the product rating is biased would lead consumers to adjust the rating downwards. In other words, the presence (vs. absence) of a fake review alert would increase perceived ratings bias, leading to lower brand evaluations. In figure 1, this is the direct link from the alert to perceived ratings bias as well as from perceived ratings bias to brand evaluations.

Finally, figure 1 shows sequential mediation from the alert to brand evaluations through expected brand dishonesty and perceived ratings bias. When consumers perceive a brand to be dishonest, they are less likely to trust the brand; this will dampen their impressions of other

brand attributes (Eisend and Tarrahi 2021). In this case, we posit that a dishonest brand will be perceived as willing to manipulate reviews again, affecting perceived ratings bias and ultimately, brand evaluations. Thus, the alert yields both direct and serial effects on brand evaluations through perceived ratings bias.

In summary, the alert affects different types of persuasion knowledge that lead to less favorable perceptions of brand evaluations. We propose that a fake review alert increases expectations of the brand's future dishonesty and perceptions of ratings bias. At the same time, expected brand dishonesty and perceived ratings bias affect brand evaluations both directly and sequentially. This leads to the following hypotheses:

H1: Relative to when an alert is absent, a fake review alert will:

a. increase expectations of the brand's future dishonesty;
b. increase perceptions that the average product rating is biased;
c. decrease brand evaluations.

H2: The effect on brand evaluations will be sequentially mediated through expected dishonesty and perceived ratings bias.

Moderators

Our theorizing asserts that a fake review alert activates lay beliefs that a dishonest brand will continue to be dishonest, and that fake reviews will result in biased product ratings. This suggests that to affect brand evaluations, we could either change these beliefs (i.e., alter the content of everyday persuasion knowledge) or change the applicability of these beliefs (i.e., alter the extent to which these beliefs are relevant to brand evaluations).

In terms of altering the content of persuasion knowledge, we propose that disconfirming lay beliefs would make consumers less likely to react negatively to the brand. Disconfirming consumers' beliefs about the impact of fake reviews on the average product rating is likely to lower perceived ratings bias. Disconfirmation could be in the form of telling consumers that fake reviews have a negligible impact on average product ratings (e.g., a vast majority are identified and deleted, so any remaining fakes should have a small statistical impact relative to the number of authentic reviews). If consumers learn that fake reviews do not have an impact on the average product rating, this means that the avenues by which the brand might mislead consumers is reduced, lowering the expected brand dishonesty. In short, in the presence of this lay belief disconfirmation, we expect attenuation of the effect of a fake review alert on the mediators, ultimately resulting in higher brand evaluations relative to when the disconfirmation is absent. The prediction of attenuation rather than elimination follows from the results of a recent meta-analysis by Eisend and Tarrahi (2021). They find that persuasion knowledge partially reduces the intended effects of advertising but does not eliminate them. This leads to the following hypothesis:

H3: Changing the content of everyday persuasion knowledge will attenuate the negative effects of H1.

The second moderator has to do with the extent to which everyday persuasion knowledge is applicable to judgments. Consumers expect that the brand will behave dishonestly in the future because "cheaters will continue to cheat." This belief is relevant because the source of the fake reviews is perceived to be the brand (i.e., brand owners and management), which leads to expectations of subsequent dishonesty. However, this belief becomes less relevant when the source of the fake reviews is not the brand (e.g., when brand ownership changes or when the fake

reviews are attributed to a third party). From a substantive point of view, failing brands sometimes use fake reviews to enhance their status and attract more customers (e.g., Lappas et al 2016). If fake reviews do not help (or the brand gets caught), the owners might sell the company to cut their losses. In this situation, consumers cannot attribute the fake reviews to the new owners, though there is still some uncertainty about the ethics of the new owners. As a result, expectations of the brand's future dishonesty will be attenuated though not eliminated. This leads to the following hypothesis:

H4: Altering the applicability of everyday persuasion knowledge will attenuate the negative effects of H1.

Next, we test these predictions in a series of studies.

OVERVIEW OF STUDIES

We present five studies to test the hypotheses. Study 1 uses web scraped data collected from Yelp.com to provide initial evidence for H1c. The next four studies are experiments. Study 2 demonstrates the effect of a fake review alert on expected brand dishonesty (H1a), perceived ratings bias (H1b), and brand evaluations (H1c), as well as serial mediation (H2). Studies 3 and 4 examine this process further via moderation by updating consumers' everyday persuasion knowledge (H3). In doing so, the studies provide evidence that consumers have a lay belief that "fake reviews distort product ratings." Finally, study 5 examines moderation by altering the applicability of persuasion knowledge (H4). The results of this study provide evidence that consumers have a lay belief that "cheaters will continue cheating" and that they attribute the fake reviews to the brand.

Study 1: Yelp Field Data

The objective of study 1 was to examine the effect of fake review alerts in the marketplace. In 2012, Yelp began their "Consumer Alert" program in which brands caught attempting to manipulate their own ratings receive an alert displayed on their Yelp page for 90 days. Depending on the specific evidence, the alert specified compensated review activity with a message that read "We caught someone red-handed trying to pay someone to write, change, or remove a review for this business. We weren't fooled but wanted you to know because these actions not only hurt consumers, but also honest businesses who play by the rules" or suspicious review activity with an alert that read "A number of positive reviews for this business originated from the same IP address, which may mean that someone was trying to artificially inflate this business's rating. Our review filter wasn't fooled but this apparent effort to mislead consumers was serious enough that we wanted to call it out". Thus, the alerts inform consumers that there had been an attempt to manipulate reviews to inflate the rating of the business, but that Yelp thwarted this attempt. After being displayed for 90 days, the alert is removed.

Procedure

To create the dataset, we analyzed news articles and press releases that mentioned the "Consumer Alert" program over a four-year period (2012-2016), which resulted in a list of 34 brands across 11 product categories (e.g., medical, restaurants, entertainment, etc.) that were known to have received a fake review alert. This is likely a small fraction of the fake review alerts deployed at the over this time period[1], though it could represent a large fraction given the

[1] At the time, the total number of fake review alerts was not disclosed to the public. Of late, Yelp has begun to report *some* of this information. In 2020, there were 457 compensated activity alerts and 699 suspicious review activity alerts (combined 1156 alerts). In 2019, they reported 319 compensated activity alerts but did not disclose the suspicious review activity alerts. This represents a 43% year-over-year increase in this alert type (Yelp 2021).

program began in 2012. We then web scraped all Yelp reviews for these brands. This resulted in an initial dataset of 34 brands and over 9000 reviews.

To avoid concerns of a brand's time on the market, we trimmed the data for each brand to only include a rolling 270-day timeframe for each brand split into three periods: the 90-day period *before* the alert was active, the 90-day period in which the alert was *active*, and the 90-day period *after* the alert. We did this because an alert is active for only 90 days, so we look at the same period length immediately preceding and following the display of a fake review alert, to hold the duration of the period constant while also minimizing the likelihood of the observed effects being caused by an endogenous factor (e.g., management changes at a restaurant). This resulted in 1998 reviews for the 34 brands; however, two reviews were missing average product ratings in the data, resulting in 1996 remaining reviews. Lastly, Yelp's proprietary algorithm classifies reviews as "recommended" or "not recommended". Whereas recommended reviews are displayed to the public and used to calculate average product ratings, not recommended reviews are hidden to the public and not used in calculating the displayed average product ratings. Of these 1996 remaining reviews, 905 reviews were recommended. We will use the recommended reviews for the reported analysis. For additional analyses including the not recommended reviews, see Web Appendix A (note: including the not recommended reviews as an additional factor yielded a null interaction, so the effects of the alert across recommendation status are conceptually similar). The dependent variable was the assigned rating in each review, so we extracted the rating from each review, and pooled the data across brands.

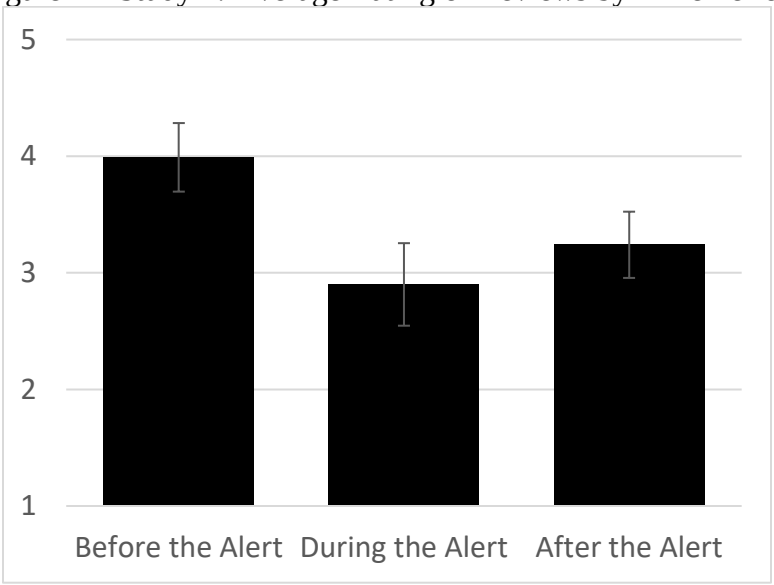
Our hypotheses are concerned with the presence versus absence of a fake review alert, so we compare the average product rating before and during the active alert. Although we make no

longitudinal claims, we also compare this to the period following an alert's removal to explore the persistent effects of alerts.

Results and Discussion

Average rating. H1c predicts that the presence (vs. absence) of a fake review alert leads to lower brand evaluations. With the assumption that each review comes from a unique individual, we conducted a one-way ANOVA of time period on average product ratings. The ANOVA yielded a significant main effect of time period ($F(2,902) = 33.58$; $p < .001$; $\eta^2_{partial} = .069$). See figure 2. Consistent with H1, planned contrasts demonstrate that the presence of a fake review alert decreased the average rating ($M_{during} = 2.90$) relative to before the alert was displayed ($M_{before} = 3.99$; $F(1,902) = 54.539$; $p < .001$; $\eta^2_{partial} = .057$). It also yielded a significantly lower rating when compared to after the alert was removed ($M_{after} = 3.24$; $F(1,902) = 28.103$; $p < .001$; $\eta^2_{partial} = .03$). Interestingly, while the average rating increased after the alert was removed, it remained lower than before the alert was displayed ($F(1,902) = 3.551$; $p = .06$; $\eta^2_{partial} = .004$), suggesting some persistent effects of a fake review alert.

Figure 2 – Study 1: Average Rating of Reviews by Time Period



Furthermore, while each of the brands had received fake review alerts at least once, it is unknown if these brands received more than one alert. As such, the results of this study are intended merely to provide motivation that the effects of fake review alerts should be further investigated, rather than to demonstrate the causal effect of a fake review alert. In the remainder of the paper, we use experiments to establish causality and allow us to identify the process by which fake review alerts influence consumers' judgments.

Study 2: Consequential Choice and Mediation

The objectives of this study were to examine the effect of an alert in a controlled setting and to test the proposed underlying process. Thus, this study was designed to test H1 and H2.

Two-hundred and thirty-four participants (M_{age} = 19.79; 49% female) completed the study in exchange for course credit. Participants were randomly assigned to one of two conditions (fake review alert: absent, present) between-subjects.

Participants were told that they would be given the product information for two fruit snack brands featured on www.shop.com, and at the end of the study they would actually receive their choice. They were then shown two products in randomized order in which most of the information was pixelated out except for the average product rating, number of reviews, and two brief reviews. The focal option (A) featured a rating of 4.3 out of 5.0 with 38 reviews. The other option (B) featured slightly inferior attributes (4.2 rating with 33 reviews). Both options displayed the text of a 5.0 rated review and 4.0 rated review, which were conceptually identical (e.g., "These fly out of the pantry at-home. Simply delicious." vs. "Great taste. This is the first thing taken when our kid's friends are over."). The fake review alert was manipulated by the presence vs. absence of a disclaimer above *only* Option A's reviews that read "Consumer Alert:

Suspicious Review Activity. We have identified and deleted some positive reviews about this product that we have determined are fake. This means that someone tried to artificially increase the rating but was unable to do so". Thus, only one option featured the alert. For the stimuli, see Web Appendix B. After viewing both options, participants responded to a series of measures.

Measures

The main dependent variable was choice. Participants chose "which brand of fruit snacks would you like to receive at the end of today's study?" option A (1) or option B (0). After this, we assessed the *alert manipulation check* (3 items; $\alpha = .694$), the degree to which the alert triggered a change-of-meaning of reviews. Participants indicated whether "most online reviews are ___" were manipulative/ corrupt/ misleading (1 = strongly disagree, 5 = strongly agree). Thus, the manipulation check was focused on suspicion of reviews in general.

Next, participants responded to questions about option A. The second measure of brand evaluations was the *perceived product rating*: "…what do you think is the true average rating for these fruit snacks?" on a 0-5 continuous scale. *Perceived ratings bias* (3 items; $\alpha = .894$) was assessed by the question, "if you were to read some of the reviews for this product (option A), to what extent do you think the current reviews would be ___" accurate/believable/authentic on a reverse-coded 7-point scale (1 = not at all, 7 = extremely). Higher numbers indicate greater bias. Finally, *expected brand dishonesty* (3 items; $\alpha = .817$) was assessed by future expectations of the brand's dishonesty in response to the statement "the company that makes these fruit snacks (option A) might: ___" followed by: lie to their investors when possible/try to cheat their customers/lie about the ingredients of their product on a 7-point scale (1 = strongly disagree, 7 = strongly agree), followed by demographics (i.e., age and gender).

Results

Alert manipulation check. A one-way ANOVA of the alert on the manipulation check resulted in a significant main effect of the alert ($F(1,232) = 5.947$; $p = .015$; $\eta^2_{partial} = .025$). A positive alert significantly increased suspicion of reviews in general ($M_{absent} = 2.59$; $M_{positive} = 2.80$).

Brand Evaluations. A binary logistic regression of choice on the alert yielded the alert as a significant predictor of choice (B = -2.194; Wald $\chi^2(1) = 52.122$; $p < .001$; odds ratio = .111). A fake review alert significantly lowered the choice share of the focal option ($M_{absent} = 79\%$, $M_{present} = 29\%$). Similarly, a one-way ANOVA of the alert on average product rating resulted in a significant main effect of the alert ($F(1,232) = 25.558$; $p < .001$; $\eta^2_{partial} = .099$). A fake review alert significantly lowered the perceived rating of the focal option ($M_{absent} = 4.04$; $M_{present} = 3.68$). This provides support for H1c.

Mediators. Supporting H1a, a one-way ANOVA of the alert on expected brand dishonesty resulted in a significant main effect of the alert ($F(1,232) = 26.285$; $p < .001$; $\eta^2_{partial} = .102$). As expected, the alert significantly increased the perceived dishonesty of the focal option's brand ($M_{absent} = 3.20$; $M_{present} = 3.93$). Similarly, supporting H1b, a one-way ANOVA of the alert on perceived ratings bias resulted in a significant main effect of the alert ($F(1,232) = 20.12$; $p < .001$; $\eta^2_{partial} = .08$). A fake review alert significantly increased the perceived ratings bias of the focal option ($M_{absent} = 3.36$; $M_{present} = 3.98$).

Serial mediation. To test H2, we used the PROCESS macro (model 6; Hayes 2017) to assess mediation of choice. We submitted alert condition as the independent variable (0 = alert absent, 1 = alert present), choice as the dependent variable, and expected brand dishonesty and perceived ratings bias as the serial mediators, respectively.

Supporting H2, the serial mediation effect via expected brand dishonesty and ratings bias was significant (B = -.1612; $CI_{95\%}$: [-.3198, -.0591]). Moreover, the indirect effect of the alert on choice via expected brand dishonesty was not significant (B = -.0772; $CI_{95\%}$: [-.3469, .1804]), suggesting that the effect of the expected brand dishonesty is fully mediated in a serial fashion. Meanwhile the indirect effect via perceived ratings bias was significant (B = -.1658; $CI_{95\%}$: [-.4014, -.0132]), suggesting that there is additional explanatory power directly through the latter mediator. Thus, we find evidence for our serial process, as well as directly via perceived ratings bias.

Using the same model with the average product rating as the dependent variable provided additional support for H2. The serial effect was again significant (B = -.0505; $CI_{95\%}$: [-.0938, -.0204]); while the indirect effect of the alert on product rating via expected brand dishonesty was not significant (B = -.0466; $CI_{95\%}$: [-.1058, .0066]); and the indirect effect via perceived ratings bias was significant (B = -.052; $CI_{95\%}$: [-.1057, -.005]). Again, we find evidence for our serial process on perceived product rating and additional mediation through perceived ratings bias directly. Importantly, reversing the order of mediators mitigates the serial mediation effect, giving us confidence that our proposed process is the appropriate one (i.e., perceptions of a brand's dishonesty inform their perceptions of ratings bias rather than the inverse), justifying our model.

Discussion

This study demonstrated the effect of an alert on consumer evaluations in a controlled setting. In doing so, we also examined possible mechanisms for the effect of an alert on evaluations and provide support for H1 and H2. We found that the effect of a fake review alert on brand evaluations (i.e., product choice and perceived product rating) operated by the lay

beliefs made salient from the activation of persuasion knowledge. Specifically, persuasion knowledge generated cognitions about the brand's willingness to deceive customers in the future (i.e., brand dishonesty) and about the potential for remaining reviews to be fraudulent (i.e., perceived ratings bias). In the next study, we further explore this mechanism by examining H3, that changing the content of everyday persuasion knowledge can attenuate the effect of the alert on brand evaluations.

Study 3: Moderation via Third-Party PK Disconfirmation

The objective of the study was to test whether consumer lay beliefs about fake reviews can be updated to attenuate their response to the alert. The subsequent study (4) was designed to test whether platforms can administer this lay belief updating directly. Thus, the next two studies test H3 and explore the efficacy of this moderating effect. If so, this would throw light on the content of the lay belief (i.e., that "fake reviews distort product ratings").

Procedure

Four-hundred and two participants (M_{age} = 37.37; 52% female) from Amazon mTurk completed the study in exchange for a $0.50 payment. Participants were randomly assigned to a 2 (PK prime: control, disconfirmation) x 2 (fake review alert: absent, present) between-subjects design.

A pretest conducted with 114 undergraduate students (M_{age} = 20.06; 52% female) confirmed that the disconfirmation prime decreased the perceived prevalence of ($M_{control}$ = 4.94, $M_{disconfirmation}$ = 3.90; $F(1,112) = 18.639$; $p < .001$; $\eta^2_{partial} = .143$), and harm from ($M_{control}$ = 4.29, $M_{disconfirmation}$ = 3.52; $F(1,112) = 8.908$; $p = .003$; $\eta^2_{partial} = .074$), fake reviews.

Participants learned that this was a two-part study. In part 1, they read an article from CNN. In the control prime, the article discussed the Disney+ upcoming releases. In the disconfirmation prime, the article stated that review platforms catch 95% of fake reviews automatically, and even if they miss one, fake reviews have very little impact (for the full articles, see Web Appendix B). This directly contradicted the lay belief that "fake reviews distort product ratings." Participants then responded to two comprehension questions about the article they read and were removed from the survey if they did not answer both correctly, before being assigned to a randomized fake review alert condition. The rate of failure did not significantly differ across conditions ($z = -.76$; $p > .40$).

Next, participants moved onto part 2 in which they imagined a scenario in which "You are on a work trip in Washington DC. You need to grab a light breakfast near your hotel, so you search for local places on a review site. This website simply hosts reviews for various businesses. While browsing the website, you find a place that seems to fit your needs, so you decide to click on the "West End Café" to learn more about it". Next, they viewed a screenshot of the landing page for the business on Yelp.com. In both conditions, they saw that the business had a 3.9 out of 5.0 rating with 96 reviews. Unlike the previous study, they did not view any individual reviews in order to maintain internal validity. The alert was manipulated as in the prior study (for the stimuli, see Web Appendix B).

Measures

Participants responded to the same items as in study 2, except for substituting in *brand intentions* for choice, with slight wording changes to account for different product category. Unlike study 2, the *alert manipulation check* (3 items; $\alpha = .856$) preceded both brand evaluation measures. The *perceived product rating* was measured as before. In addition, *brand intentions* (3

items; α = .854) were measured by three items: "how likely would you be to try this restaurant?", "How much would you be willing to pay for breakfast at this place?", and "How likely would you be to suggest this place to a friend?" on 7-point scales. Another difference from study 2 was that we reversed the administration of our two mediators: *expected brand dishonesty* (3 items; α = .919) and *perceived ratings bias* (3 items; α = .958).

Results

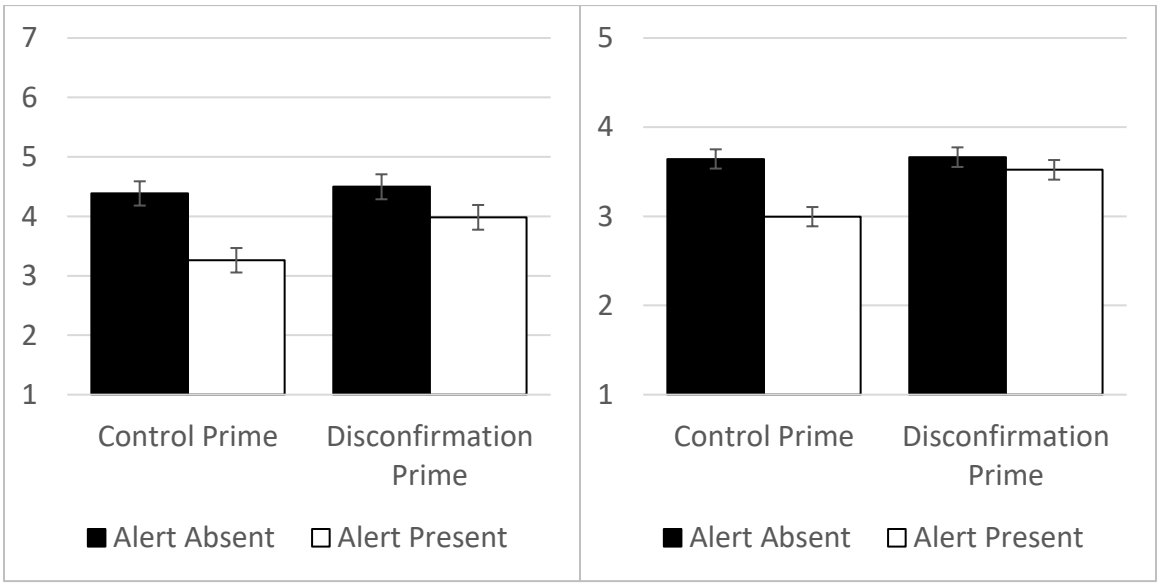
Alert manipulation check. A 2 (prime) x 2 (alert) ANOVA on the manipulation check resulted in a significant main effect of the alert ($F(1,398) = 12.787$; $p < .001$; $\eta^2_{partial} = .031$). Neither the main effect of the prime ($F(1,398) = 1.519$; $p = .219$; $\eta^2_{partial} = .004$) nor the interaction ($F(1,398) = .84$; $p = .36$; $\eta^2_{partial} = .002$) were significant. As expected, participants were more suspicious about reviews in general in the presence vs. absence of an alert ($M_{absent} = 2.37$, $M_{present} = 2.65$).

Brand evaluations. The results for brand intentions and perceived product ratings were similar and consistent with H3. A 2x2 ANOVA revealed significant main effects of prime [BI: ($F(1,398) = 16.241$; $p < .001$; $\eta^2_{partial} = .039$; Perceived product rating: $F(1,398) = 25.137$; $p < .001$; $\eta^2_{partial} = .059$)] and alert [BI: ($F(1,398) = 62.608$; $p < .001$; $\eta^2_{partial} = .136$; Perceived product rating: $F(1,398) = 52.624$; $p < .001$; $\eta^2_{partial} = .117$), qualified by a significant interaction ($F(1,398) = 8.718$; $p = .003$; $\eta^2_{partial} = .021$; Perceived product rating: $F(1,398) = 21.539$; $p < .001$; $\eta^2_{partial} = .051$). Planned contrasts demonstrated that in the presence of the control prime, a fake review alert significantly decreased the participants' brand intentions ($M_{absent} = 4.39$, $M_{present} = 3.26$; $F(1,398) = 59.921$; $p < .001$; $\eta^2_{partial} = .131$) and perceived product rating ($M_{absent} = 3.64$, $M_{present} = 3.00$; $F(1,398) = 71.821$; $p < .001$; $\eta^2_{partial} = .153$). This replicated study 2. However, this effect was attenuated, but not eliminated, with the treatment prime (BI: $M_{absent} = 4.50$,

$M_{present}$ = 3.98; F(1,398) = 12.119; p = .001; $\eta^2_{partial}$ = .03; Perceived product rating: M_{absent} = 3.66, $M_{present}$ = 3.52; F(1,398) = 3.364; p = .067; $\eta^2_{partial}$ = .008).

Comparing the simple effects of the prime within alert conditions, when the alert was absent the disconfirmation prime did not significantly impact brand intentions (F(1,398) = .58; p = .447; $\eta^2_{partial}$ = .001) or perceived product rating (F(1,398) = .069; p = .792; $\eta^2_{partial}$ = 0). However, when the fake review alert was present, the disconfirmation prime led to higher brand intentions (F(1,398) = 24.386; p < .001; $\eta^2_{partial}$ = .058) and perceived product rating (F(1,398) = 46.62; p < .001; $\eta^2_{partial}$ = .105). See figure 3.

Figure 3: a) Brand intentions and b) Perceived product rating by article prime and alert



Expected brand dishonesty. A 2x2 ANOVA yielded significant main effects of prime (F(1,398) = 17.624; p < .001; $\eta^2_{partial}$ = .042) and alert (F(1,398) = 50.335; p < .001; $\eta^2_{partial}$ =

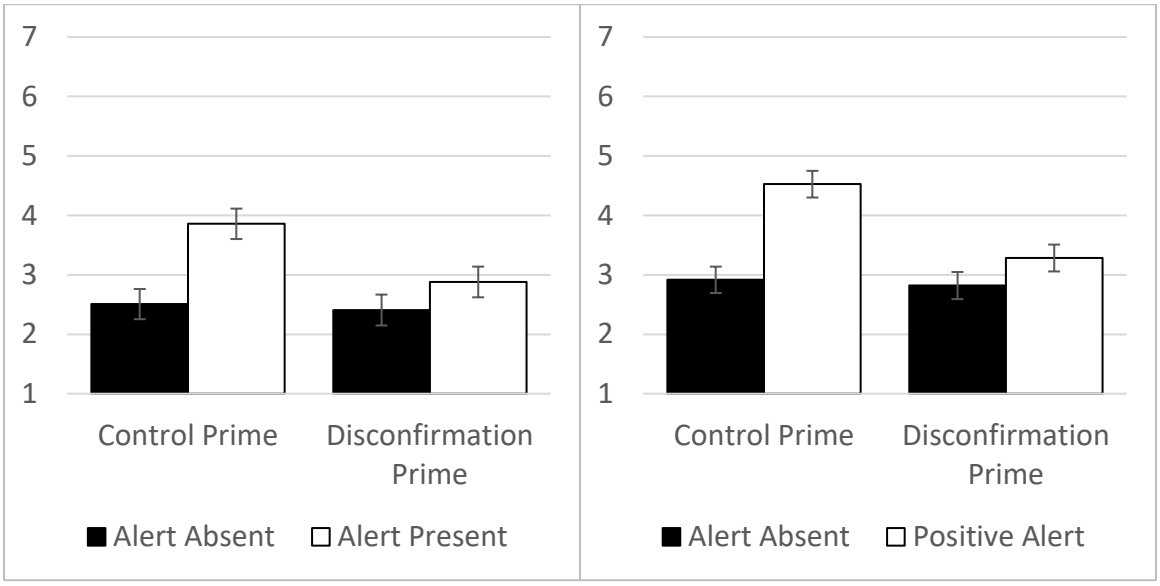
.112), qualified by a significant interaction ($F(1,398) = 11.695$; $p = .001$; $\eta^2_{partial} = .029$). Planned contrasts demonstrated that in the presence of the control prime, a fake review alert significantly increased expected subsequent dishonesty of the brand ($M_{absent} = 2.51$, $M_{prsent} = 3.86$; $F(1,398) = 56.115$; $p < .001$; $\eta^2_{partial} = .124$). Consistent with H3, however, with the disconfirmation prime, this effect was attenuated ($M_{absent} = 2.41$, $M_{present} = 2.88$; $F(1,398) = 6.653$; $p = .01$; $\eta^2_{partial} = .016$).

Comparing simple effects of the prime within alert conditions, when the alert was absent the prime did not significantly impact perceptions of expected brand dishonesty ($F(1,398) = .303$; $p = .582$; $\eta^2_{partial} = .001$). However, when the alert was present, the disconfirmation prime lowered brand manipulative intent ($F(1,398) = 29.025$; $p < .001$; $\eta^2_{partial} = .068$). See figure 4.

Perceived ratings bias. A 2x2 ANOVA on the perceived ratings bias resulted in significant main effects of prime ($F(1,398) = 35.211$; $p < .001$; $\eta^2_{partial} = .081$) and alert ($F(1,398) = 84.537$; $p < .001$; $\eta^2_{partial} = .175$), qualified by the interaction ($F(1,398) = 25.816$; $p < .001$; $\eta^2_{partial} = .061$). Planned contrasts demonstrated that in the presence of the control prime, a fake review alert significantly increased perceived ratings bias ($M_{absent} = 2.92$, $M_{present} = 4.53$; $F(1,398) = 103.437$; $p < .001$; $\eta^2_{partial} = .206$). Consistent with H3, however, with the disconfirmation prime, the effect of the alert was attenuated ($M_{absent} = 2.82$, $M_{present} = 3.28$; $F(1,398) = 8.336$; $p = .004$; $\eta^2_{partial} = .021$).

Comparing the simple effects of the prime within alert conditions, when the alert was absent the prime did not significantly impact the perceived ratings bias ($F(1,398) = .364$; $p = .547$; $\eta^2_{partial} = .001$). However, when the alert was present, the disconfirmation prime decreased the perceived ratings bias ($F(1,398) = 60.681$; $p < .001$; $\eta^2_{partial} = .132$). See figure 4.

Figure 4: a) Expected brand dishonesty and b) Perceived ratings bias by article prime and alert



Moderated serial mediation of brand intentions via expected brand dishonesty and ratings bias. We used the PROCESS macro (model 85; Hayes 2017) to assess moderated serial mediation of brand intentions[2]. We submitted alert condition as the independent variable (0 = alert absent, 1 = positive alert), the PK disconfirmation prime as the moderator (0 = control, 1 = PK prime), perceived product rating as the dependent variable, and expected brand dishonesty and ratings bias as serial mediators, respectively. As expected, the index of moderated serial mediation via expected brand dishonesty and ratings bias was significant (B = .1453; $CI_{95\%}$: [.0556, .2529]). Under the control prime, the serial effect was significant (B = -.2234; $CI_{95\%}$: [-.3293, -.1337]). This effect persisted under the PK disconfirmation prime (B = -.0781; $CI_{95\%}$: [-.1487, -.0209]) but was attenuated.

[2] The mediation results for perceived product rating are virtually identical. To save space and avoid redundancy, we report these results in Web Appendix C.

Again, the index of moderated mediation via brand dishonesty was also significant (B = .1336; $CI_{95\%}$: [.0292, .2892]). Under the control prime, brand dishonesty partially mediated the effect of the alert on brand intentions (B = -.2054; $CI_{95\%}$: [-.402, -.0551]). This remained in the presence of the PK disconfirmation prime (B = -.0718; $CI_{95\%}$: [-.1698, -.0101]) but was greatly attenuated.

As expected, the index of moderated mediation via perceived ratings bias was also significant (B =.3068; $CI_{95\%}$: [.1321, .5185]). Under the control prime, the effect of the alert on brand intentions was partially mediated via perceived ratings bias (B = -.4118; $CI_{95\%}$: [-.6293, -.2319]). This effect remained in the presence of the PK disconfirmation prime (B = -.105; $CI_{95\%}$: [-.2171, -.0063]), but was greatly attenuated.

Discussion

This study further investigated the role of lay beliefs about the impact of fake reviews on brand evaluations. Consistent with H3, the disconfirmation of lay beliefs regarding fake reviews attenuates the negative effect of the alert on brand evaluations, expected brand dishonesty, and perceived ratings bias was attenuated. This has two important implications. First, the attenuation suggests that consumers' everyday persuasion knowledge does consist of the lay belief that "fake reviews distort ratings," since disconfirmation resulted in attenuation. Second, the attenuation, but not elimination, of the negative effects of the alert is consistent with a meta-analysis that shows that persuasion knowledge attenuates, but does not eliminate, the effect of marketers' advertising tactics (Eisend and Tarrahi 2021). Thus, our work on fake reviews seems to align with work in the context of advertising.

In the next study, we examine H3 further by assessing whether the platform itself can influence consumers' reactions to the fake review alert by presenting disconfirmation

information directly on the brand's page within the platform. If so, a substantive implication would be that platforms could use such language to help consumers make more accurate decisions. We did not have a priori predictions about the effectiveness of this strategy. On the one hand, the platform would not be as independent a source as a news organization; however, given that the platform also issues the fake review alert, this information should be seen as credible. On the other hand, the disconfirmation information presented by the platform is concurrent with the alert rather than preceding the alert. This might also make a difference in terms of consumers' responses, though the direction of the effect is unclear.

Study 4: Moderation via First-Party PK Disconfirmation

Three-hundred and ninety-eight participants (M_{age} = 39.14; 55% female) completed the study on Amazon mTurk in exchange for a $0.50 payment. Participants were randomly assigned to a condition in a 2 (PK disconfirmation: absent, present) x 2 (fake review alert: absent, present) between-subjects design.

The study was similar to study 3 with a few key changes. First, the product category was hotels, and the target brand was West End Hotel, with an average product rating of 3.7 out of 5.0 and 82 reviews. Importantly, the PK disconfirmation information was embedded below the alert manipulation in a light green box; in the alert absent condition, the disconfirmation information appeared in the same place on the page. The box stated, "Fake Reviews Have Little Impact on this Site". It then described the same key information from the article used in the prior study: that the website's algorithm catches 95% of fake reviews automatically and that even if a fake review is missed it is likely to have little impact. In the disconfirmation absent condition, there was no

light green box and no mention of the impact of fake reviews. For the complete stimuli, see Web Appendix B.

Measures

After viewing the West End Hotel information, participants responded to the *alert manipulation check* (3 items; α = .817), followed by the *perceived product rating*, *brand intentions* (3 items; α = .901), *expected brand dishonesty* (3 items; α = .915), and *perceived ratings bias* (3 items; α = .958). Again, minor wording variations were used for the hotel category. To assess whether consumers saw West End as a cheater, we included additional measures of perceptions of the brand's manipulativeness (3 items; α = .938) with the items "West End Hotel is: manipulative/corrupt/misleading" on 5-point agreement scales. A subsequent factor analysis showed that brand manipulativeness loaded on the same factor as expected honesty, confirming that the brand was perceived as a cheater now and in the future. Because the results for brand manipulativeness did not differ from those for expected dishonesty, we report these results separately in Web Appendix D. Finally, we included a *PK disconfirmation manipulation check* with the item "to what extent do fake reviews affect the average ratings of all businesses on this website" on a 7-point scale, followed by demographics.

Results

Alert manipulation check. A 2 (PK disconfirmation) x 2 (alert) ANOVA on the alert manipulation check yielded a marginal effect of the alert ($F(1,394) = 2.778$; $p = .096$; $\eta^2_{partial} = .007$). Neither the effect of the disconfirmation ($F(1,394) = .069$; $p = .793$; $\eta^2_{partial} = 0$) nor the interaction ($F(1,394) = .386$; $p = .535$; $\eta^2_{partial} = .001$) were significant. Consistent with prior studies, though statistically weaker, the alert activated suspicion about reviews in general ($M_{present} = 2.65$; $M_{absent} = 2.52$). We speculate that these results may be weaker because the alert

was presented concurrently with the disconfirmation information, so that encoding may have been weaker (i.e., there was more information to process at the point of encoding).

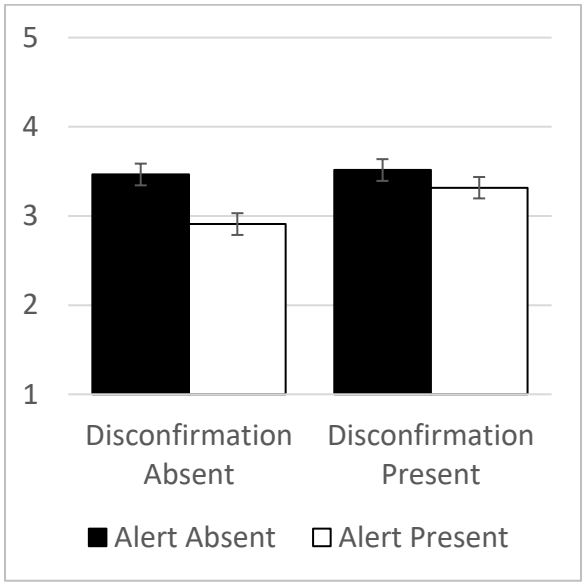
PK disconfirmation manipulation check. A 2x2 ANOVA on the PK disconfirmation manipulation check yielded a main effect of PK disconfirmation ($F(1,394) = 75.383$; $p < .001$; $\eta^2_{partial} = .161$). Neither the effect of the alert ($F(1,394) = .658$; $p = .418$; $\eta^2_{partial} = .002$) nor the interaction ($F(1,394) = .658$; $p = .418$; $\eta^2_{partial} = .002$) were significant. Participants perceived the impact of fake reviews to be much lower in the presence vs. absence of the PK disconfirmation message ($M_{present} = 3.28$; $M_{absent} = 4.75$). Thus, the disconfirmation manipulation was successful.

Brand evaluations. In this study, the results for brand intentions and average product rating were different. A 2x2 ANOVA on brand intentions yielded significant main effects of the alert $F(1,394) = 48.161$; $p < .001$; $\eta^2_{partial} = .109$) and PK disconfirmation ($F(1,394) = 11.914$; $p < .001$; $\eta^2_{partial} = .029$). Contrary to H3, the interaction was not significant ($F(1,394) = 1.793$; $p = .181$; $\eta^2_{partial} = .005$). Consistent with prior studies, brand intentions were lower in the presence (vs. absence) of a fake review alert ($M_{absent} = 4.07$, $M_{present} = 3.25$). Furthermore, brand intentions were lower in the absence vs. presence of PK disconfirmation ($M_{absent} = 3.46$, $M_{positive} = 3.86$).

A 2x2 ANOVA on the perceived product rating yielded significant main effects of the alert $F(1,394) = 38.841$; $p < .001$; $\eta^2_{partial} = .09$) and PK disconfirmation ($F(1,394) = 14.351$; $p < .001$; $\eta^2_{partial} = .035$) qualified by a significant interaction ($F(1,394) = 8.722$; $p = .003$; $\eta^2_{partial} = .022$). Planned contrasts demonstrated that in the absence of PK disconfirmation, a fake review alert significantly decreased the perceived product rating ($M_{absent} = 3.47$, $M_{present} = 2.91$; $F(1,394) = 41.978$; $p < .001$; $\eta^2_{partial} = .096$). Consistent with H3, in the presence of disconfirmation, the effect of the alert was attenuated ($M_{absent} = 3.64$, $M_{present} = 3.00$; $F(1,394) = 5.403$; $p = .021$; $\eta^2_{partial} = .014$).

Comparing simple effects of the PK disconfirmation within alert conditions, when the alert was absent, PK disconfirmation did not significantly impact the perceived product rating ($F(1,394) = .347$; $p = .556$; $\eta^2_{partial} = .001$). However, when the fake review alert was present, disconfirmation led to a significantly higher perceived product rating ($F(1,394) = 22.838$; $p < .001$; $\eta^2_{partial} = .055$). See figure 5. Thus, H3 was supported for perceived product rating but not for brand intentions.

Figure 5: Perceived product rating by PK disconfirmation and alert



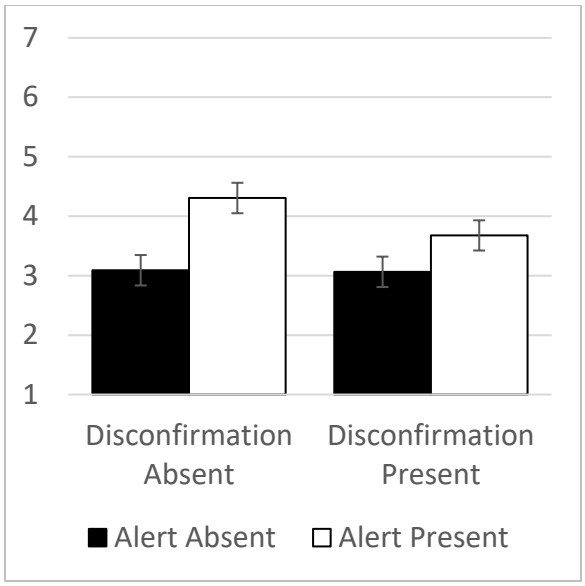
Expected brand dishonesty. A 2x2 ANOVA on expected brand dishonesty yielded significant main effects of the alert $F(1,394) = 75.632$; $p < .001$; $\eta^2_{partial} = .161$) and PK disconfirmation ($F(1,394) = 24.346$; $p < .001$; $\eta^2_{partial} = .058$). The interaction was not significant ($F(1,394) = 1.383$; $p = .24$; $\eta^2_{partial} = .003$). Consistent with prior studies, brand dishonesty was greater in the presence (vs. absence) of a positive alert ($M_{absent} = 3.13$, $M_{present} = 4.17$).

Furthermore, brand dishonesty was greater in the absence vs. presence of PK disconfirmation (M_{absent} = 3.95, $M_{positive}$ = 3.36). Thus, H3 is not supported for expected brand dishonesty.

Perceived ratings bias. A 2x2 ANOVA on the perceived ratings bias yielded significant main effects of the alert F(1,394) = 51.286; $p < .001$; $\eta^2_{partial} = .115$) and PK disconfirmation (F(1,394) = 6.62; $p = .01$; $\eta^2_{partial} = .017$) qualified by a significant interaction (F(1,394) = 5.578; $p = .019$; $\eta^2_{partial} = .014$). Planned contrasts demonstrated that in the absence of PK disconfirmation, a fake review alert significantly increased perceived ratings bias (M_{absent} = 3.09, $M_{present}$ = 4.31; F(1,394) = 45.122; $p < .001$; $\eta^2_{partial} = .103$). Consistent with H3, in the presence of PK disconfirmation, the effect of the alert was attenuated (M_{absent} = 3.06, $M_{present}$ = 3.68; F(1,394) = 11.576; $p = .001$; $\eta^2_{partial} = .029$). Thus, H3 is supported for perceived ratings bias.

Comparing simple effects of PK disconfirmation within alert conditions, when the alert was absent, PK disconfirmation did not significantly impact perceived ratings bias (F(1,394) = .022; $p = .882$; $\eta^2_{partial} = 0$). However, when the alert was present, disconfirmation led to significantly lower perceived bias (F(1,394) = 12.236; $p = .001$; $\eta^2_{partial} = .03$). See figure 6.

Figure 6: Perceived ratings bias by PK disconfirmation and alert



Mediation. Given that brand intentions and perceived product rating showed different effects, we summarize both results here and present the detailed analyses in Web Appendix D. Because the interaction of alert and PK disconfirmation on brand intentions was not significant, we examined the serial mediation of the alert on brand intentions using PROCESS Model 6 (as in study 2). Consistent with that study, we find evidence for a serial mediation effect (B = -.1086; $CI_{95\%}$: [-.1887, -.0508]), supporting H2. This means that the fake review alert increased expectations of a brand's subsequent dishonesty, which increased the perceived ratings bias, ultimately yielding lower brand intentions.

For the perceived product rating, we examined moderated serial mediation with PROCESS Model 85 (as in study 3). The index of moderated serial mediation via expected brand dishonesty and perceived ratings bias was not significant (B = .0155; $CI_{95\%}$: [-.0099, .0482]). In both the absence (B = -.0651; $CI_{95\%}$: [-.1093, -.027]) and presence (B = -.0496; $CI_{95\%}$: [-.0856, -.0206]) of the PK disconfirmation information, serial mediation remained significant, again

providing evidence for H2 but not H3. This demonstrates that persuasion knowledge disconfirmation by the platform did not effectively moderate the serial mediation effect.

Discussion

This study yielded mixed results. Although H3 was supported for perceived product rating and perceived bias, it was not supported for brand intentions and expected brand dishonesty. This suggests that PK disconfirmation by the platform is not as effective as third-party disconfirmation. This enriches our understanding of how the content of persuasion knowledge might be updated. It suggests that platforms can update the lay beliefs generated about fake reviews at the point of persuasion knowledge activation, thereby ensuring that consumers do not misapply their activated lay beliefs about the reviews on this platform. At the same time, it does not update the lay beliefs regarding the brand's expected dishonesty, which ultimately feed into their intentions to patronize the brand. Thus, platforms can ensure that consumers form accurate impressions of the reviews on their platform, while still allowing brands to be penalized for their behavior. Taken together with the results of study 3, it suggests that platforms need not attenuate attributions about the brand's dishonesty to avoid an overreaction by consumers regarding the veracity of the platform's content. If the platform provides information disconfirming lay beliefs about fake reviews, consumers can make more accurate judgments of the product rating and the perceived ratings bias, while still making negative attributions about the brand, lowering brand intentions.

In the next study, we examine H4, that changing the applicability of persuasion knowledge will moderate the effect of an alert. Specifically, we expected that the applicability of the lay belief that "cheaters will continue cheating" would be low in the presence of an ownership change, leading to attenuation of the negative effect of the alert.

Study 5: Moderation via Applicability of Persuasion Knowledge

Four-hundred and three participants (M_{age} = 25.59; 50% female) completed the study on Prolific in exchange for a $0.64 payment. Participants were randomly assigned to a 2 (applicability of PK: low, high) x 2 (fake review alert: absent, present) between-subjects design. The procedure was the same as in prior studies, with the product category as scooter rentals. Applicability of persuasion knowledge was manipulated through whether the establishment changed owners. As in study 4, the brand page contained a light green box that explained: "Please note that the ownership of this business recently changed. This means that older reviews may not be relevant to the business anymore. See below for a message from the new owners." In the same green box was also a message from the owners that read "Thank you for checking out our business. We promise to do better than the previous owners in order to regain your trust". See Web Appendix B for the stimuli.

Measures

Participants responded to the same measures: the *alert manipulation check* (3 items; α = .740), *average product rating*, *brand intentions* (3 items; α = .864), *expected brand dishonesty* (3 items; α = .878), and *perceived ratings bias* (3 items; α = .898), and *brand manipulativeness scale* (3 items; α = .890) with minor wording variations to account for the scooter rental category change. Again, we report the results of brand manipulativeness in the web appendix (E) for simplicity. We also assessed the *ownership manipulation check* with the item "which of the owners of this business do you think are more ethical?" (1 = the older owners are definitely more ethical, 7 = the new owners are definitely more ethical), followed by demographics.

Results

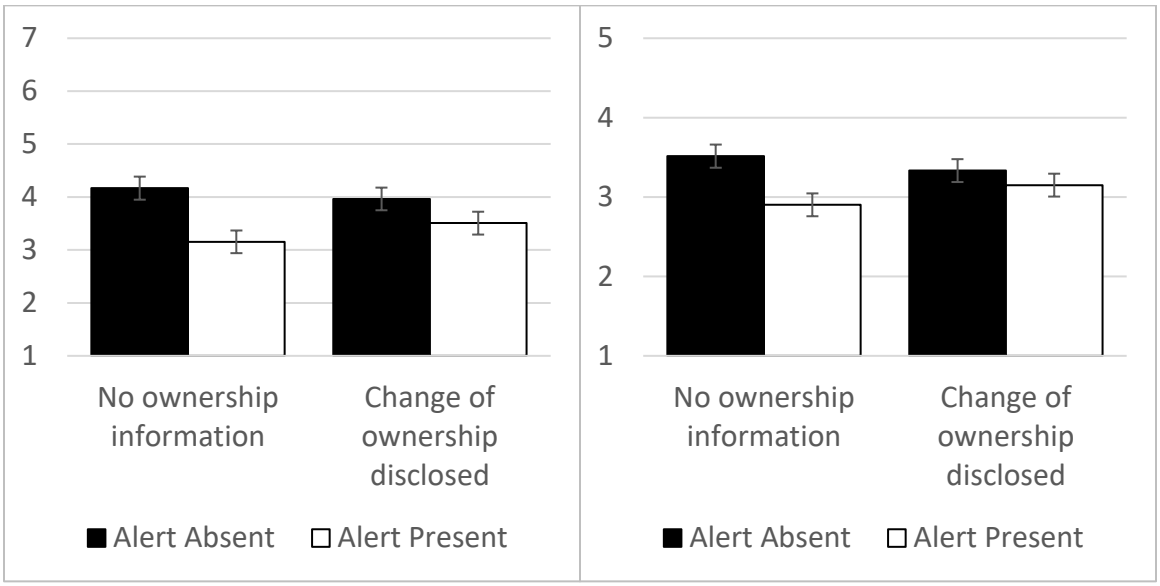
Alert manipulation check. A 2 (ownership change) x 2 (alert) ANOVA on the alert manipulation check yielded a main effect of the alert ($F(1,399) = 12.563$; $p < .001$; $\eta^2_{partial} = .031$). Neither the main effect of PK applicability ($F(1,399) = .88$; $p = .349$; $\eta^2_{partial} = .002$) nor the interaction ($F(1,399) = .007$; $p = .933$; $\eta^2_{partial} = 0$) were significant. Consistent with prior studies, the presence of an alert increased suspicion about reviews in general ($M_{present} = 2.92$; $M_{absent} = 2.66$).

Brand evaluations. H4 predicted a significant interaction effect on brand evaluations. A 2x2 ANOVA on the perceived product rating yielded a significant main effect of the alert [BI: $F(1,399) = 46.743$; $p < .001$; $\eta^2_{partial} = .105$; Perceived product rating: $F(1,399) = 30.479$; $p < .001$; $\eta^2_{partial} = .071$], qualified by the predicted interaction [BI: $F(1,399) = 6.7$; $p = .01$; $\eta^2_{partial} = .017$; Perceived product rating: ($F(1,399) = 8.877$; $p = .003$; $\eta^2_{partial} = .022$]. The main effect of applicability was not significant (p's > .49). When applicability was high, a fake review alert decreased brand intentions ($M_{absent} = 4.17$, $M_{present} = 3.15$; $F(1,399) = 44.306$; $p < .001$; $\eta^2_{partial} = .1$) and perceived product rating ($M_{absent} = 3.52$, $M_{present} = 2.90$; $F(1,399) = 36.034$; $p < .001$; $\eta^2_{partial} = .083$). When applicability was high, however, the effect of an alert was attenuated on both brand intentions ($M_{absent} = 3.96$, $M_{present} = 3.51$; $F(1,399) = 9.047$; $p = .003$; $\eta^2_{partial} = .022$) and product rating ($M_{absent} = 3.33$, $M_{present} = 3.15$; $F(1,399) = 3.238$; $p = .073$; $\eta^2_{partial} = .008$). This provides support for H4.

Comparing simple effects of PK applicability within alert conditions, when the alert was absent, applicability did not impact intentions ($F(1,399) = 1.796$; $p = .181$; $\eta^2_{partial} = .004$) and marginally decreased the perceived product rating ($F(1,399) = 3.168$; $p = .076$; $\eta^2_{partial} = .008$). When the alert was present, however, a change of ownership yielded a significantly higher

intentions (F(1,399) = 5.391; p = .021; $\eta^2_{partial}$ = .013) and product rating (F(1,399) = 5.926; p = .015; $\eta^2_{partial}$ = .015). See figure 7.

Figure 7: a) Brand intentions, and b) Perceived product rating by ownership change and alert

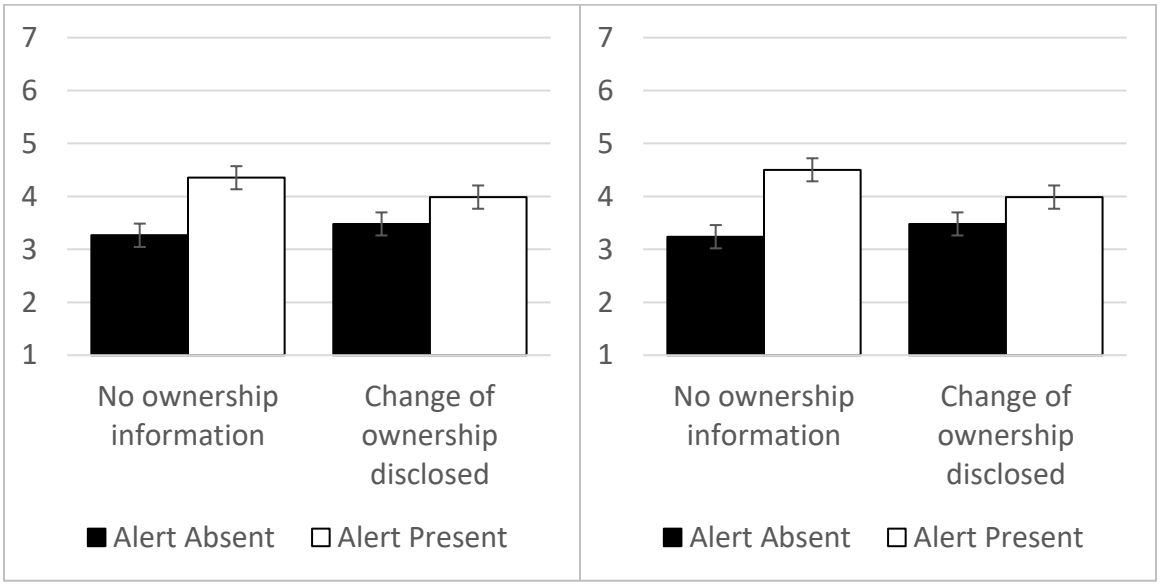


Expected brand dishonesty. Supporting H4, a 2x2 ANOVA on expected brand dishonesty yielded a significant main effect of the alert F(1,399) = 45.568; p < .001; $\eta^2_{partial}$ = .102), qualified by the predicted interaction (F(1,399) = 5.716; p = .017; $\eta^2_{partial}$ = .014). The main effect of applicability was not significant (F(1,399) = 1.232; p = .268; $\eta^2_{partial}$ = .003). When PK applicability was high, the alert increased expected brand dishonesty (M_{absent} = 3.27, $M_{present}$ = 4.35; F(1,399) = 41.675; p < .001; $\eta^2_{partial}$ = .095). When applicability was low, however, the effect of the alert was attenuated (M_{absent} = 3.42, $M_{present}$ = 3.94; F(1,399) = 9.527; p = .002; $\eta^2_{partial}$ = .023). This provides support for H4. See figure 8.

Perceived ratings bias. Supporting H4, a 2x2 ANOVA on perceived ratings bias yielded a significant main effect of the alert F(1,399) = 65.459; p < .001; $\eta^2_{partial}$ = .141), qualified by the

predicted interaction ($F(1,399) = 11.996$; $p = .001$; $\eta^2_{partial} = .029$). The main effect of PK applicability was not significant ($F(1,399) = 1.582$; $p = .209$; $\eta^2_{partial} = .004$). When applicability was high, the alert increased perceived ratings bias ($M_{absent} = 3.24$, $M_{present} = 4.50$; $F(1,399) = 66.581$; $p < .001$; $\eta^2_{partial} = .143$). When applicability was low, the effect of the alert was attenuated ($M_{absent} = 3.48$, $M_{present} = 3.99$; $F(1,399) = 10.732$; $p = .001$; $\eta^2_{partial} = .026$). Thus, H4 was supported. See figure 8b.

Figure 8: a) Expected brand dishonesty, and b) Perceived ratings bias by ownership and alert



Moderated serial mediation. Using the same model to investigate moderated serial mediation as in studies 3 and 4 (PROCESS model 85; Hayes 2017), we again found evidence of moderated serial mediation for both the perceived product rating (B = .0326; $CI_{95\%}$: [.0046, .0718]).) and brand intentions (B = .0559; $CI_{95\%}$: [.0079, .1211]). See web appendix E for full details. In short, in the absence of a disclosed ownership change, we found robust evidence for H2. The presence of a fake review alert increased expectations of a brand's dishonest behavior, which in turn increased the perceived ratings bias, ultimately lowering both the perceived

product rating (B = -.2538; $CI_{95\%}$: [-.3603, -.1588]) and brand intentions (B = -.1069; $CI_{95\%}$: [-.1763, -.0538]). In the presence of a disclosed ownership change, although these effects persisted for both the perceived product rating (B = -.0298; $CI_{95\%}$: [-.0578, -.009]) and brand intentions (B = -.051; $CI_{95\%}$: [-.0967, -.0171]), they did so at significantly reduced effect sizes. The serial effect was reduced by nearly 90% for the perceived product rating and over 50% for brand intentions, suggesting that a change of ownership is an effective way to curb the effect of a fake review alert. At the same time, the effect persisted, suggesting that new owners need to be aware of perceptions of the previous owners when taking over a business.

Discussion

This study was designed to further examine our process by determining whether the effects of fake review alerts are localized to the perceived agent of manipulation (i.e., the owners at the time of offense) or the brand itself. Our findings suggest that a change of ownership can attenuate, but not fully mitigate, the effect of a fake review alert on brand evaluations. This suggests that consumers still have some apprehensions about patronizing a brand that was engaged in manipulative behaviors even when new ownership arrives. This could be a function of a perceived relationship between old and new owners, though we did not explicitly examine this possibility. Nevertheless, we find that new ownership does attenuate the effect of an alert on the brand's expected dishonesty, which attenuates the perceived bias of the average ratings, ultimately attenuating the effect of an alert on brand evaluations. Thus, new ownership does not incur the full effect of a fake review alert but is subject to some residual effects. This suggests that new owners should be aware of any brand (mis)perceptions before they acquire it, as the previously activated persuasion knowledge may remain active.

GENERAL DISCUSSION

The objective of the paper was to examine the downstream consequences of fake review alerts on brand evaluations. Across five studies, we demonstrated that the presence of a fake review alert leads to lower perceptions of the product rating, brand intentions, and choice. The presence (vs. absence) of an alert informed consumers that the brand had attempted to persuade them by using deception. Consumers coped with this failed persuasion attempt by applying everyday persuasion knowledge in the form of lay beliefs about the brand and the effect of fake reviews. Specifically, consumers' everyday persuasion knowledge that "cheaters will continue cheating" increased expectations that the brand would behave dishonestly in the future. Moreover, everyday persuasion knowledge that "fake reviews distort ratings" led to higher perceptions of ratings bias. Both expected brand dishonesty and perceived ratings bias affected brand evaluations.

We identified two moderators of these effects. Studies 3 and 4 showed that disconfirming the lay belief that "fake reviews distort ratings" significantly attenuates these effects. This attenuation provides support for the proposed underlying process of everyday persuasion knowledge. Similarly, study 5 showed that changing the applicability of persuasion knowledge also leads to attenuation; when consumers learn that the brand has changed ownership, they are less likely to apply the "cheaters will continue cheating" belief to the brand. Although the new owners escape some of the punishment doled out to the previous owners for being deceptive, uncertainty about the ethics of the new owners remains, leading to attenuation but not elimination of the effect.

We demonstrated the effect of a fake review alert across a multitude of outcomes from actual Yelp ratings (study 1); product choice within a consideration set (study 2); perceived product ratings (studies 2-5); and intentions to patronize the brand (studies 3-5). In addition, the experimental stimuli spanned a variety of product categories: fruit snacks, hotel, restaurant, and scooter rental. Thus, we feel confident in the robustness of our findings. Next, we consider the implications of our work.

Theoretical Implications

We take a different approach to persuasion knowledge than prior work by focusing on the content, rather than activation, of everyday persuasion knowledge. It may be fairly obvious that a fake review alert will make consumers vigilant and suspicious. What is less obvious is what types of persuasion knowledge consumers access to cope with this persuasion. We focused on two lay beliefs, the first of which is general to all persuasion contexts (i.e., "cheaters will continue to cheat"); while the second is specific to the context of fake reviews (i.e., "fake reviews distort ratings"). Note that the general belief is much harder to change than the specific belief. The general belief, by definition, applies to a wide variety of persuasion contexts, so it may be based on actual experience as well as observation. As prior research using attribution theory shows (Reeder and Brewer 1979; Skowronski and Carlston 1987), individuals make dispositional inferences about dishonesty from a single dishonest behavior. Thus, when an entity (e.g., brand, person, company) is caught in deception, consumers will assume that the entity is dishonest. This is why, in study 5, we focused on changing the applicability of the belief rather than the belief itself.

In contrast, the belief that fake reviews distort product ratings is specific to the context of fake reviews. Based on the platform's algorithms, however, this belief may not be accurate, as the alert informs consumers that all fake reviews have been removed. Thus, this lay belief is likely more malleable than general beliefs about persuasion. Disconfirming this lay belief was effective in attenuating the negative effects of a fake review alert on brand evaluations. In fact, our approach of exploring the content of everyday persuasion knowledge allowed us to identify this lay belief as important to the review context. Although fake review alerts inform consumers that deceptive reviews are no longer present, it appears that the alerts may trigger inaccurate persuasion knowledge. If consumers adjust their perceptions of the average product rating because they perceive it to be biased, they are not making accurate judgments about the product rating.

We determined these two lay beliefs as important based on earlier research we had conducted involving open-ended responses. Of course, there may be other relevant beliefs that we have not identified. Future research may examine further the content of everyday persuasion knowledge about fake reviews. In addition, exploring everyday persuasion knowledge in other contexts may lead to useful insights.

Although we found evidence of serial mediation in studies 2-5, it could be argued that expected brand dishonesty and perceived ratings bias may work in parallel rather than sequentially. In conducting parallel (vs. serial) mediation analyses, we generally saw an increased effect size of the perceived ratings bias and decreased effect size of expected brand dishonesty. While parallel mediation somewhat simplifies the story, it overemphasizes the importance of consumers' perceptions of the reviews and underemphasizes the importance of their brand perceptions. Consumers who believe a brand is dishonest will expect them to engage

in subsequent dishonest behavior, including review manipulation. Without addressing the brand perceptions, it is extremely difficult to mitigate the negative effect of the alert on review perceptions. Thus, when persuasion knowledge is directed towards multiple sources, marketers need consider the relationship between these sources to fully understand the potential responses by consumers.

Managerial Implications

Our research provides some answers about whether fake review alerts are good for brands, consumers, and platforms. For the brand, the alert is clearly not beneficial. The alert increases perceptions of ratings bias and expectations of future dishonesty while decreasing perceptions of the average product rating and lowering brand intentions. Moreover, study 1 shows that it takes some time for the brand to recover from the negative effects of the alert. Thus, brands should be wary of using fake reviews given the consequences of being caught.

For consumers, fake review alerts have both positive and negative effects. In the absence of a fake review alert, consumers are unlikely to see the brand as deceptive and may be misled by dishonest brands. In pointing out that the brand has behaved deceptively, the alert allows consumers to evaluate the brand's future dishonesty and decreases brand intentions and choice. This is the positive effect of the alert. On the other hand, the presence of a fake review alert makes consumers suspicious of the existing reviews and the average product rating. As such, consumers might question the veracity of the authentic reviews, even though the reviews have been vetted by the platform. In this case, the alert leads consumers to make inaccurate judgments about the average product rating. This is the negative effect of the alert.

For the platform, the results are also mixed. On the one hand, the alert achieves the objectives of punishing the brand for deception and warns consumers about the brand's dishonesty. However, since the alert highlights the existence of fake reviews, it creates greater uncertainty for consumers about whether all the fake reviews have been detected and removed. The current platform strategy of stating that the platform has removed fake reviews is not sufficient to reassure consumers. Instead, we suggest that the platform implement a disconfirmation strategy like the one in study 4 by explaining why fake reviews do not distort the product ratings. From the platform's perspective, this strategy reassures consumers that they are seeing veridical ratings and reviews, which is beneficial for the platform. Thus, the platform can manage consumers' responses to fake review alerts. The platform's effort should be in changing consumers' persuasion knowledge that "fake reviews distort ratings." The platform could publicize its ability to detect and remove fake reviews, which might be covered by the media. As shown in study 3, independent disconfirmation is the most effective way to counter consumers' lay beliefs about the effects of fake reviews on ratings.

Another important issue is whether the fake review alert hurts all brands equally. In our studies, the average rating of the focal brand varied from 3.7 (study 4 and 5), 3.9 (study 3), and 4.3 (study 2) on a five-point scale, which may be seen as moderate-high ratings. We expect that when the brand is poorly rated, floor effects occur such that a fake review alert does not significantly decrease brand evaluations. So low-quality brands have less to lose. To test this boundary, we ran a study in which participants learned that the brand-in-question had either a low or high average rating once the fake reviews were removed. Although we replicated the results when the brand rating was high (4.0), there were no effects of a fake review alert when the brand rating was low (2.0). Thus, consumers who already hold a poor opinion of a brand are

unlikely to be swayed by the fake review alert. While this is intuitive, brands solicit fake reviews for various reasons, not just to increase their average rating. Fake reviews can make the brand appear more popular than it is due to an inflated review volume, and this metric is generally used platform algorithms for how they display search options. Lappas, Sabnis, and Valkanas (2016) found that merely 50 undetected fake reviews are enough for a brand to game a platform's algorithm and boost its visibility beyond its competitors. Future research might further explore whether this type of knowledge will lead consumers to incur additional costs (e.g., search time, travel distance, and price) to avoid falling prey to review manipulation.

Finally, an interesting question for future research has to do with consumer responses to other sources of fake reviews. Recently, some brands have received a fake review alert that blocks access to the reviews due to "high profile media coverage" of their business (Yelp 2020). These often arise from brands, or affiliates, engaging in politically or ideologically motivated behaviors; such as the bakery selling "build the wall" cookies (Brown 2019) or the restaurant whose owner's estranged sister was considered a right-wing firebrand (Wilson 2018).In these contexts, supporters (e.g., 5-star reviews) or opponents (e.g., 1-star reviews) of the ideological stance flood the platform (e.g. Yelp) with fake reviews. Thus, while the reviews are not reflective of true quality, they are used as a weapon to hurt or help a brand rather than a source of information. In this situation, the fake reviews are not solicited by the brand, leading to potential different assessments of expected brand dishonesty as consumers cannot apply the "cheater" lay belief to the brand. Future research should examine these different sources of fake reviews and alerts.

REFERENCES

Akoglu, Leman, Rishi Chandy, and Christos Faloutsos (2013), "Opinion Fraud Detection in Online Reviews by Network Effects," *Seventh International AAAI Conference on Weblogs and Social Media*.

Ahluwalia, Rohini (2002), "How Prevalent is the Negativity Effect in Consumer Environments?" *Journal of Consumer Research*, 29(2), 270-279.

Bassig, Migs (2020), "Yelp Factsheet: Stats your Business Needs to Know*,"* (accessed January 5, 2022), [available at https://www.reviewtrackers.com/yelp-factsheet/].

Boush, David M., Marian Friestad, and Gregory M. Rose (1994), "Adolescent Skepticism Toward TV Advertising and Knowledge of Advertiser Tactics," *Journal of Consumer Research*, 21(1), 165-175.

BrightLocal (2020), "*Local Consumer Review Survey,"* (accessed January 5, 2022), [available at https://www.brightlocal.com/learn/local-consumer-review-survey].

Brown, Andrea (2019), "*Edmonds Bakery Keeps Making 'Build the Wall' Cookies,*" (accessed January 14, 2022), [available at https://www.heraldnet.com/news/edmonds-bakery-inadvertently-ignites-a-fresh-controversy/].

Campbell, Margaret C. and Amna Kirmani (2000), "Consumers' Use of Persuasion Knowledge: The Effects of Accessibility and Cognitive Capacity on Perceptions of an Influence Agent," *Journal of Consumer Research*, 27(1), 69-83.

———, and ——— (2008), "I Know What You're Doing and Why You're Doing It," *Handbook of Consumer Psychology*, 549-574.

———, Gina S. Mohr, and Peeter W.J. Verlegh (2013), "Can Disclosures Lead Consumers to Resist Covert Persuasion? The Important Roles of Disclosure Timing and Type of Response," *Journal of Consumer Psychology*, 23(4), 483-495.

Chan, Elaine, and Jaideep Sengupta (2010), "Insincere Flattery Actually Works: A Dual Attitudes Perspective," *Journal of Marketing Research*, 47(1), 122-133.

Chevalier, Judith A., and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43(3), 345-354.

Darke, Peter R., and Robin J.B. Ritchie (2007), "The Defensive Consumer: Advertising Deception, Defensive Processing, and Distrust," *Journal of Marketing Research*, 44(1), 114-127.

De Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2015), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42(6), 817-833.

Eisend, Martin, and Farid Tarrahi (2021), "Persuasion Knowledge in the Marketplace: A Meta-Analysis," *Journal of Consumer Psychology*.

Feng, Song, Ritwik Banerjee, and Yejin Choi (2012), "Syntactic Stylometry for Deception Detection," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics.

Friestad, Marian, and Peter Wright (1994), "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts," *Journal of Consumer Research*, 21(1), 1-31.

———, and ——— (1995), "Persuasion Knowledge: Lay People's and Researchers' Beliefs about the Psychology of Advertising," *Journal of Consumer Research*, 22(1), 62-74.

Hardesty, David M., William O. Bearden, and Jay P. Carlson (2007), "Persuasion Knowledge and Consumer Reactions to Pricing Tactics," *Journal of Retailing*, 83(2), 199-210.

Hayes, Andrew F. (2017), "*Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach*," Guilford Publications.

Hennig-Thurau, Thorsten, Gianfranco Walsh, and Gianfranco Walsh (2003), "Electronic Word-of-Mouth: Motives for and Consequences of Reading Customer Articulations on the Internet." *International Journal of Electronic Commerce*, 8(2), 51-74.

Jain, Shailendra P., and Steven S. Posavac (2004), "Valenced Comparisons," *Journal of Marketing Research*, 41(1), 46-58.

Johar, Gita V., and Carolyn J. Simmons (2000), "The Use of Concurrent Disclosures to Correct Invalid Inferences," *Journal of Consumer Research*, 26(4), 307-322.

Kelley, Harold H., and John L. Michela (1980), "Attribution Theory and Research," *Annual Review of Psychology*, 31(1), 457-501.

Kirmani, Amna, and Margaret C. Campbell (2004), "Goal Seeker and Persuasion Sentry: How Consumer Targets Respond to Interpersonal Marketing Persuasion," *Journal of Consumer Research*, 31(3), 573-582.

———, and Rui Zhu (2007), "Vigilant Against Manipulation: The Effect of Regulatory Focus on the Use of Persuasion Knowledge," *Journal of Marketing Research*, 44(4), 688-701.

Lappas, Theodoros, Gaurav Sabnis, and Georgios Valkanas (2016), "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," *Information Systems Research*, 27(4), 940-961.

Luca, Michael, and Georgios Zervas (2016), "Fake it Till You Make it: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62(12), 3412-3427.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104(8), 2421-2455.

Mukherjee, Arjun, Vivek Venkataraman, Bing Liu, and Natalie S. Glance (2013), "What Yelp Fake Review Filter Might be Doing?" *ICWSM*.

Ott, Myle, Claire Cardie, and Jeffrey T. Hancock (2013), "Negative Deceptive Opinion Spam," *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Pechmann, Cornelia (1996), "Do Consumers Overgeneralize One-sided Comparative Price Claims, and are More Stringent Regulations Needed?" *Journal of Marketing Research*, 33(2), 150-162.

Reeder, Glenn D., and Marilynn B. Brewer (1979), "A Schematic Model of Dispositional Attribution in Interpersonal Perception," *Psychological Review*, 86(1), 61.

Rule, Brendan G., Gay L. Bisanz, and Melinda Kohn (1985), "Anatomy of a Persuasion Schema: Targets, Goals, and Strategies," *Journal of Personality and Social Psychology*, 48(5), 1127.

Schiffer, Zoe (2020), "*Amazon Takes Down a Five-star Fraud in the UK,"* (accessed September 8, 2020), [available at https://www.theverge.com/2020/9/4/21423429/amazon-top-reviewers-uk-fraud].

Skowronski, John J., and Donal E. Carlston (1987), "Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases," *Journal of Personality and Social Psychology*, 52(4), 689.

Steffel, Mary, Elanor F. Williams, and Ruth Pogacar (2016), "Ethically Deployed Defaults: Transparency and Consumer Protection through Disclosure and Preference Articulation," *Journal of Marketing Research*, 53(5), 865-880.

TripAdvisor (2020), "*What Happens if a Business is Found to have Fraudulent Reviews?"* (accessed September 21, 2020), [available at https://www.tripadvisorsupport.com/hc/en-us/articles/200614957-What-happens-if-a-business-is-found-to-have-fraudulent-reviews-].

Watson, Jared, Anastasiya Pocheptsova Ghosh, and Michael Trusov (2018), "Swayed by the Numbers: The Consequences of Displaying Product Review Attributes," *Journal of Marketing*, 82(6), 109-131.

Wilson, Andrew E., Peter R. Darke, and Jaideep Sengupta (2021), "Winning the Battle but Losing the War: Ironic Effects of Training Consumers to Detect Deceptive Advertising Tactics," *Journal of Business Ethics*, 1-17.

Wilson, Michael (2018), "*An Online Agitator, A Social Media Expose and the Fallout in Brooklyn*," (accessed January 14, 2022), [available at https://www.nytimes.com/2018/06/06/nyregion/amymek-mekelburg-huffpost-doxxing.html].

Wu, Tai-Yee, and Carolyn A. Lin (2017), "Predicting the Effects of eWOM and Online Brand Messaging: Source Trust, Bandwagon Effect and Innovation Adoption Factors," *Telematics and Informatics*, 34(2), 470-480.

Yelp (2019), *"Consumer Protection Initiative,"* (accessed April 11, 2019), [available at https://blog.yelp.com/category/en/news/consumer-protection-initiative].

——— (2020), *"New Consumers Alert on Yelp Takes Firm Stance Against Racism,"* (accessed January 14, 2022), [available at https://blog.yelp.com/news/new-consumer-alert-on-yelp-takes-firm-stance-against-racism/].

——— (2021), *"Yelp Trust & Safety Report 2020,"* (accessed January 5, 2022), [available at https://trust.yelp.com/wp-content/uploads/2021/02/Yelp-Trust-and-Safety-Report-2020.pdf].

Zhao, Yi, Sha Yang, Vishal Narayan, and Ying Zhao (2013), "Modeling Consumer Learning from Online Product Reviews," *Marketing Science*, 32(1), 153-169.

WEB APPENDIX - TABLE OF CONTENTS

WEB APPENDIX A – Additional Results for Study 1

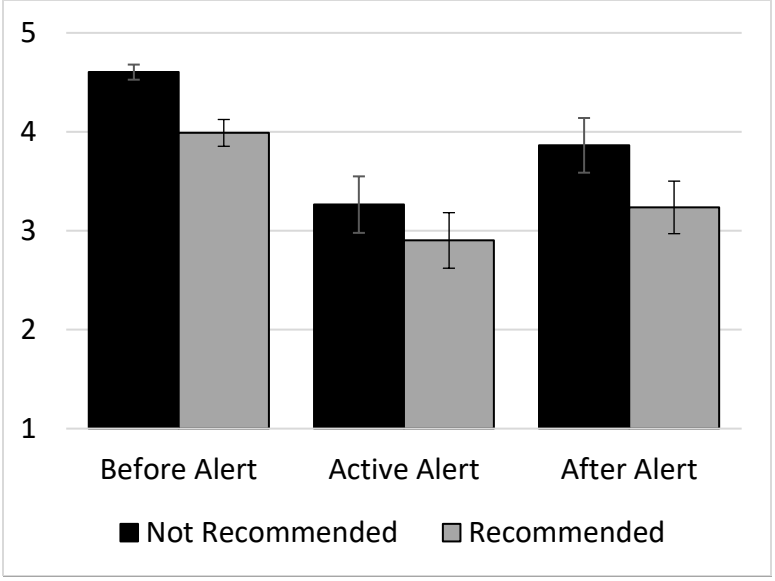
In addition to the fake review alerts, Yelp uses an algorithm to determine the likely veracity of reviews and automatically filters reviews into "recommended" and "not recommended" lists. Not recommended reviews are separated from the recommended reviews, are not prominently displayed (i.e., they are displayed on a separate page accessed via a discrete link), and do not impact the calculation of the brand's rating nor review volume. Although Yelp's sorting algorithm is proprietary, it is thought to include attributes such as the reviewer's IP address, the reviewer's review volume, and syntactical cues within the review. While reviews can be filtered for reasons other than being fraudulent (e.g., lacking usefulness), they do serve as a useful proxy for fake reviews. Thus, we will examine the effect of a fake review alert on both the recommended and not recommended reviews in the data as both provide useful insights for managers.

Additional Measures

Proportion of fake reviews. The alert had the intended effect of reducing the proportion of fake (i.e., not recommended) relative to authentic (i.e., recommended) reviews for the focal brand. A binary logistic regression of the presence of an alert on the proportion of recommended reviews yielded a significant omnibus effect (Wald $\chi^2(2) = 23.032$; $p < .001$; odds ratio = .755). Planned contrasts demonstrate that the proportion of fake reviews was significantly higher before an alert was active ($P_{before} = 58.33\%$) relative to during ($P_{during} = 51.12\%$; Wald $\chi^2(1) = 5.895$; $p = .015$; odds ratio = .992) or after it was removed ($P_{after} = 44.48\%$; Wald $\chi^2(1) = 20.855$; $p < .001$; odds ratio = .76), while the decline between during and after was marginal (Wald $\chi^2(1) = 3.093$; $p = .079$; odds ratio = 1.143). Thus, the presence of a positive alert improves the veracity of reviews for the target brand.

Average star rating. H1c suggests that a positive alert decreases brand evaluations relative to when an alert is absent. With the assumption that each review comes from a unique individual, we conducted a 2 (review type: not recommended, recommended) x 3 (time: before alert, active alert, after alert) ANOVA on average product ratings. The ANOVA yielded significant main effects of review type ($F(1,1990) = 43.23$; $p < .001$; $\eta^2_{partial} = .021$) and time period ($F(2,1990) = 104.187$; $p < .001$; $\eta^2_{partial} = .095$), while the interaction was not significant ($F(2,1990) = 1.05$; $p = .35$). See figure 2. As expected, the average "not recommended" review had a higher rating than the average "recommended" review ($M_{not\ recommended} = 4.28$, $M_{recommended} = 3.62$). This is consistent with the notion that Yelp's filter is identifying more fake reviews as positive versus negative. Furthermore, the pattern of not recommended reviews followed the same pattern of results as the recommended reviews (reported in the main text). See figure WA.1.

Figure WA.1 – Study 1: Average Rating of Reviews by Time Period and Recommendation Type



We also conducted a linear regression with brand fixed effects, demonstrating that the significant effects of review type and time period remained significant. See Table WA.1.

Table WA.1: Study 1 - Linear Regression Model with Brand Fixed Effects

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.686	.075		48.922	.000
	Review	-.667	.078	-.206	-8.573	.000
	Time	-.673	.062	-.321	-10.853	.000
	Review * Time	.084	.082	.030	1.033	.302
	d0	1.981	1.347	.027	1.471	.141
	d1	.786	.300	.050	2.621	.009
	d2	1.088	.377	.054	2.888	.004
	d3	1.585	.106	.313	14.924	.000
	d4	.716	.251	.054	2.847	.004
	d5	-.506	.285	-.033	-1.775	.076
	d6	.061	.259	.004	.237	.813
	d7	-.104	.259	-.008	-.402	.688
	d8	.353	.675	.010	.523	.601
	d9	-.020	.231	-.002	-.086	.932
	d10	1.323	.222	.119	5.952	.000

d11	-1.291	.254	-.097	-5.081	.000
d12	-.238	.157	-.029	-1.519	.129
d13	-1.992	.236	-.162	-8.445	.000
d14	-.687	.334	-.039	-2.061	.039
d15	.692	.219	.061	3.163	.002
d16	1.251	.203	.117	6.157	.000
d17	1.218	.260	.089	4.690	.000
d18	1.981	.779	.047	2.542	.011
d19	.686	.154	.088	4.446	.000
d20	.997	.257	.074	3.877	.000
d21	.641	.512	.023	1.251	.211
d22	2.569	1.349	.036	1.905	.057
d23	-1.073	.186	-.111	-5.764	.000
d24	-1.483	.674	-.041	-2.199	.028
d25	1.372	.479	.054	2.865	.004
d27	-.171	.675	-.005	-.253	.800
d28	1.571	.241	.124	6.522	.000
d29	-.232	.127	-.037	-1.822	.069
d30	-.218	.251	-.016	-.868	.386
d31	.671	.200	.065	3.365	.001
d32	.661	.195	.065	3.382	.001
d33	.665	.478	.026	1.390	.165

a. Dependent Variable: Rating

Table WA.2: Review Count by Brand by Condition Crosstabs

Brand			Time			Total
			Before Alert	During Alert	After Alert	
d0	Review type	Not Recommended				
		Recommended		1		1
	Total			1		1
d1	Review type	Not Recommended	17	2	2	21
		Recommended				
	Total			2	2	21
d2	Review type	Not Recommended	4	1	1	6
		Recommended	2	4	1	7
	Total		6	5	2	13
d3	Review type	Not Recommended	20	15	30	65
		Recommended	75	31	59	165
	Total		95	46	89	230

d4	Review type	Not Recommended	9	2	3	14
		Recommended	6	5	5	16
	Total		15	7	8	30
d5	Review type	Not Recommended	9	1		10
		Recommended	11	2		13
	Total		20	3		23
d6	Review type	Not Recommended	10	6	1	17
		Recommended	6	3	2	11
	Total		16	9	3	28
d7	Review type	Not Recommended	11	3		14
		Recommended	13	1		14
	Total		24	4		28
d8	Review type	Not Recommended	0	2		2
		Recommended	2	0		2
	Total		2	2		4
d9	Review type	Not Recommended	20	3	1	24
		Recommended	3	3	6	12
	Total		23	6	7	36
d10	Review type	Not Recommended	16		27	43
		Recommended				
	Total		16		27	43
d11	Review type	Not Recommended	8	4	9	21
		Recommended	2	3	4	9
	Total		10	7	13	30
d12	Review type	Not Recommended	14	12	3	29
		Recommended	25	15	14	54
	Total		39	27	17	83
d13	Review type	Not Recommended	7	17	4	28
		Recommended	4	3	0	7
	Total		11	20	4	35
d14	Review type	Not Recommended	6	8	3	17
		Recommended	1	1	0	2
	Total		7	9	3	19
d15	Review type	Not Recommended	3	16	5	24
		Recommended	3	13	1	17
	Total		6	29	6	41

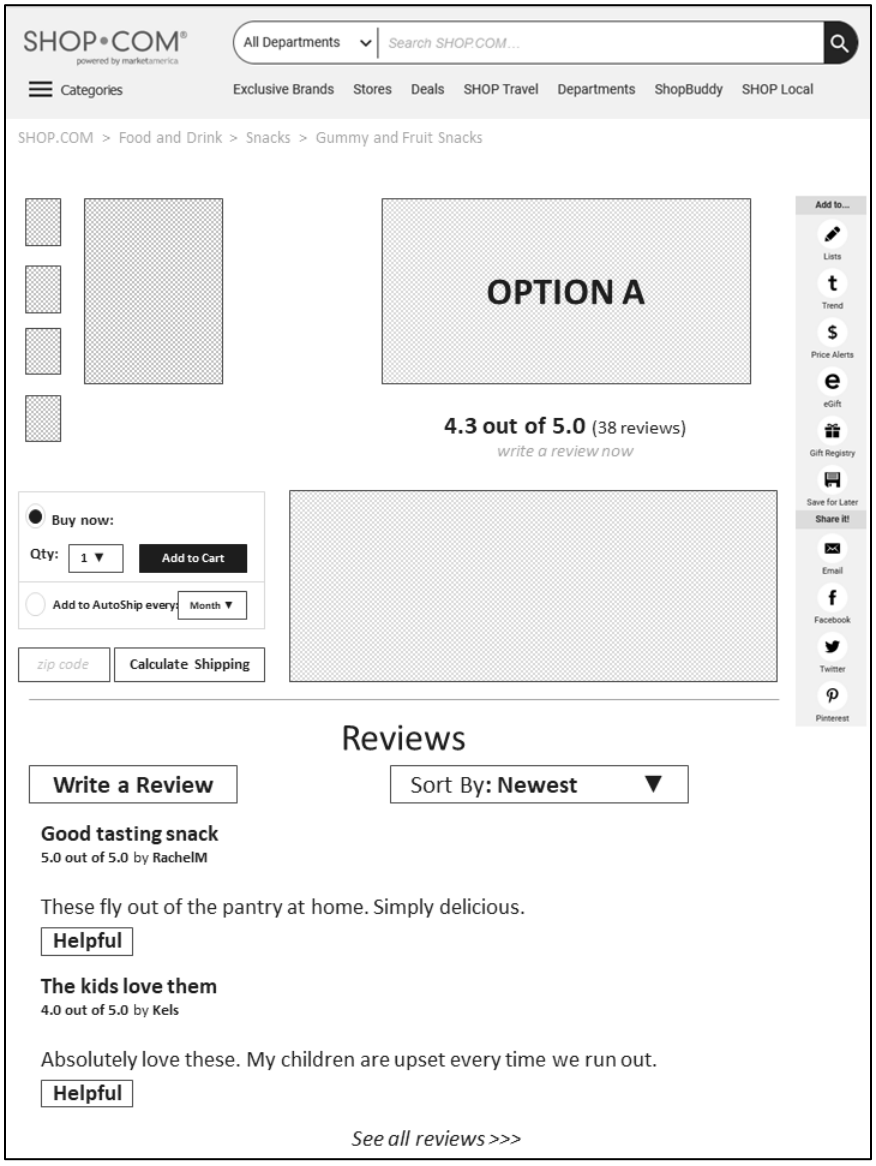
d16	Review type	Not Recommended	20	3	7	30
		Recommended	11	1	5	17
	Total		31	4	12	47
d17	Review type	Not Recommended	5	1	1	7
		Recommended	19	0	2	21
	Total		24	1	3	28
d18	Review type	Not Recommended				
		Recommended	1	1	1	3
	Total		1	1	1	3
d19	Review type	Not Recommended	77	3	1	81
		Recommended	7	1	0	8
	Total		84	4	1	89
d20	Review type	Not Recommended	7	4	8	19
		Recommended	7	2	1	10
	Total		14	6	9	29
d21	Review type	Not Recommended	7			7
		Recommended				
	Total		7			7
d22	Review type	Not Recommended				
		Recommended			1	1
	Total				1	1
d23	Review type	Not Recommended	33	11	1	45
		Recommended	8	2	2	12
	Total		41	13	3	57
d24	Review type	Not Recommended	2			2
		Recommended	2			2
	Total		4			4
d25	Review type	Not Recommended	1	0		1
		Recommended	6	1		7
	Total		7	1		8
d26	Review type	Not Recommended	320	13	14	347
		Recommended	283	58	56	397
	Total		603	71	70	744
d27	Review type	Not Recommended	3			3
		Recommended	1			1
	Total		4			4

d28	Review type	Not Recommended	5	1	4	10
		Recommended	12	1	10	23
	Total		17	2	14	33
d29	Review type	Not Recommended	39	26	21	86
		Recommended	21	18	14	53
	Total		60	44	35	139
d30	Review type	Not Recommended	7	9	0	16
		Recommended	6	3	5	14
	Total		13	12	5	30
d31	Review type	Not Recommended	37	8	4	49
		Recommended	1	0	0	1
	Total		38	8	4	50
d32	Review type	Not Recommended	35	11	2	48
		Recommended	1	1	2	4
	Total		36	12	4	52
d33	Review type	Not Recommended	4	0	1	5
		Recommended	2	1	0	3
	Total		6	1	1	8
Total	Review type	Not Recommended	756	182	153	1091
		Recommended	541	175	191	907
	Total		1297	357	344	1998

WEB APPENDIX B: Stimuli for Studies 2-5

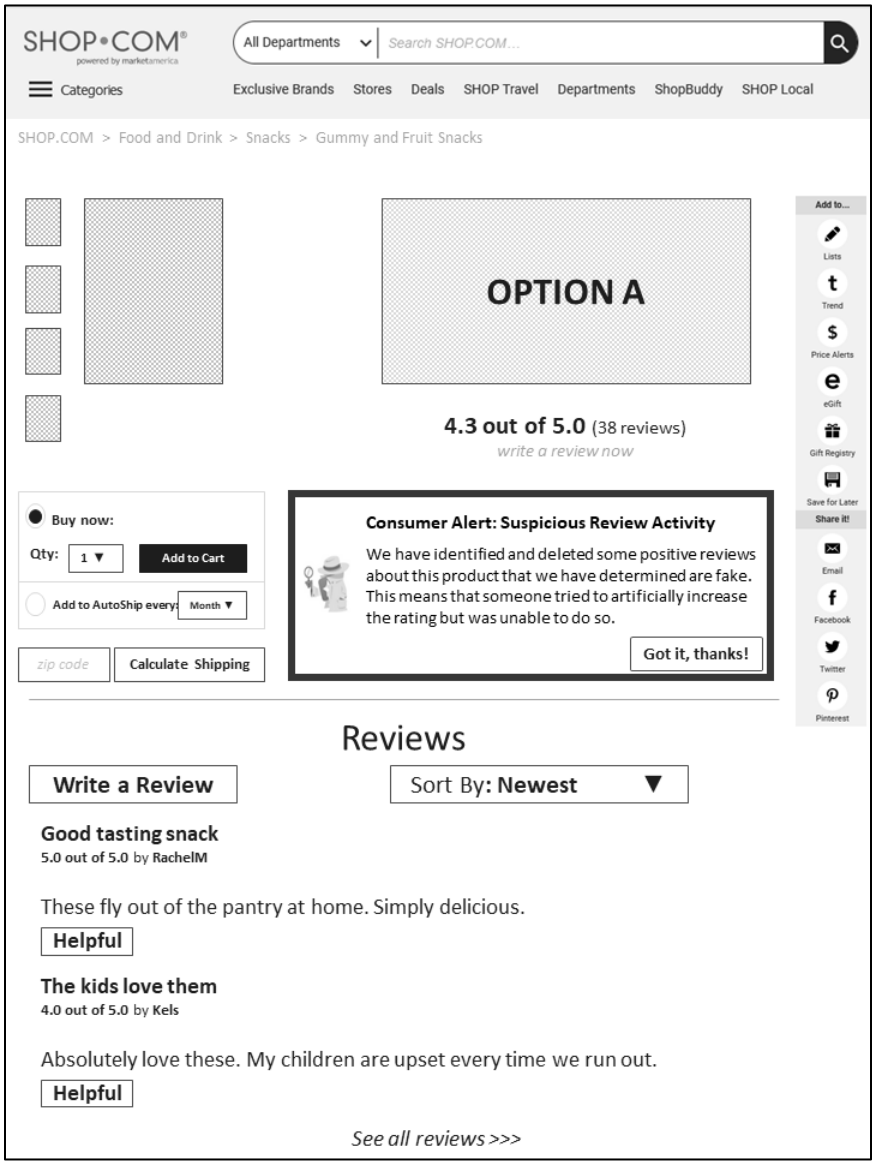
Study 2

Focal Option (alert absent condition)



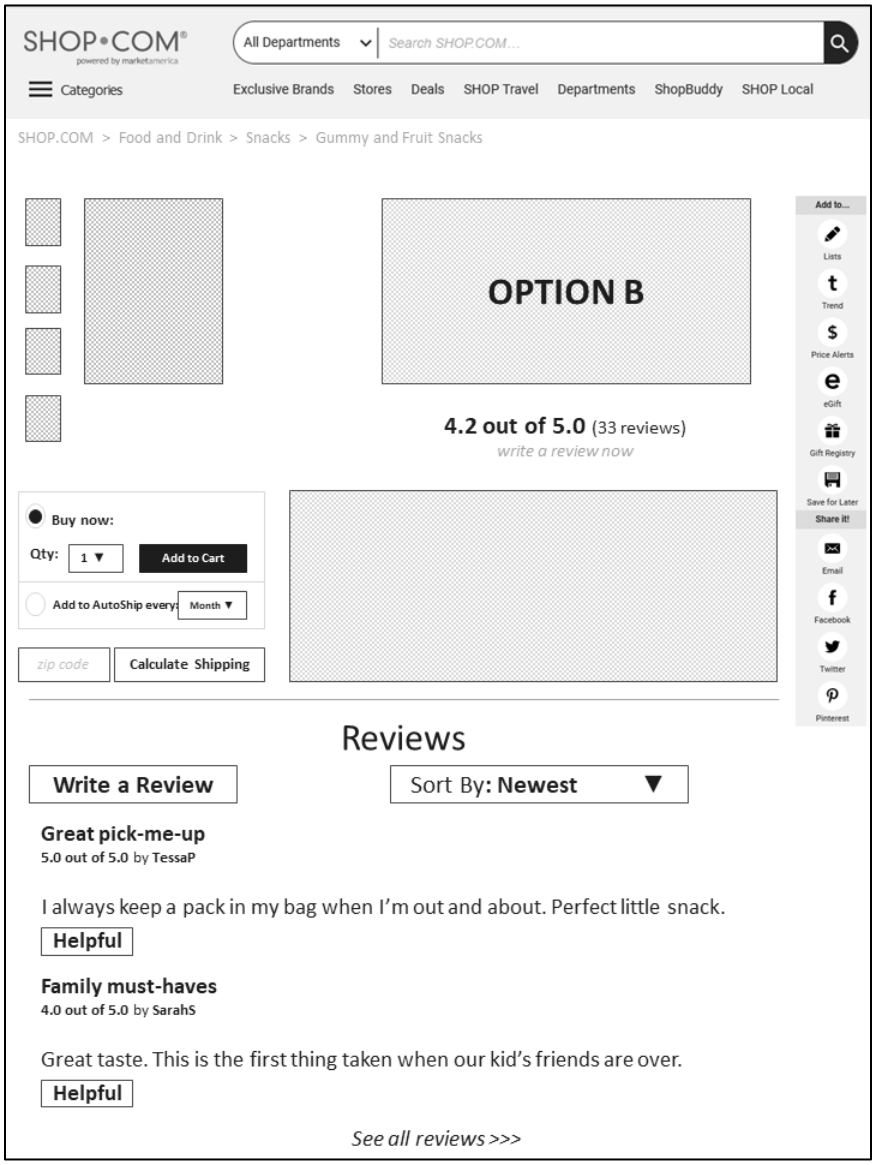
Study 2

Focal Option (alert present condition)



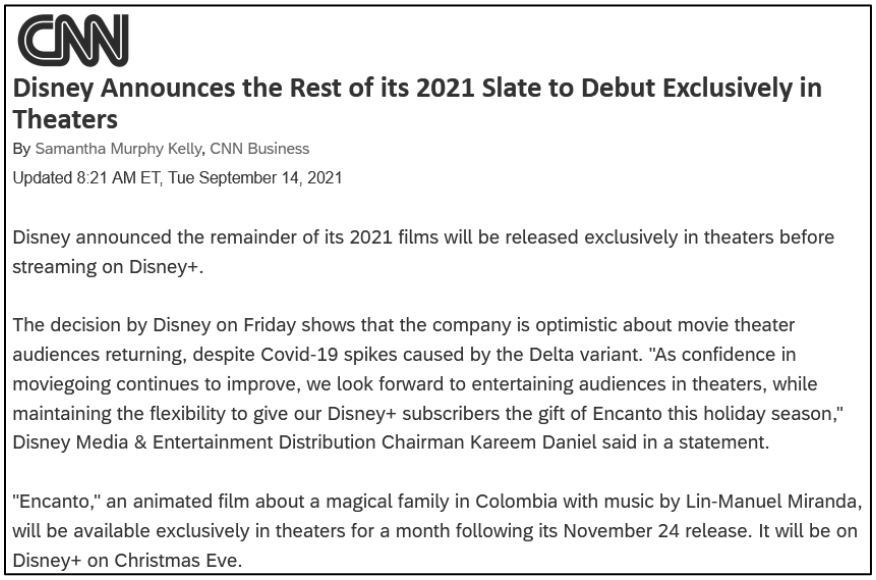
Study 2

Competing Option (all conditions)



Study 3

Control Prime



Disney Announces the Rest of its 2021 Slate to Debut Exclusively in Theaters

By Samantha Murphy Kelly, CNN Business

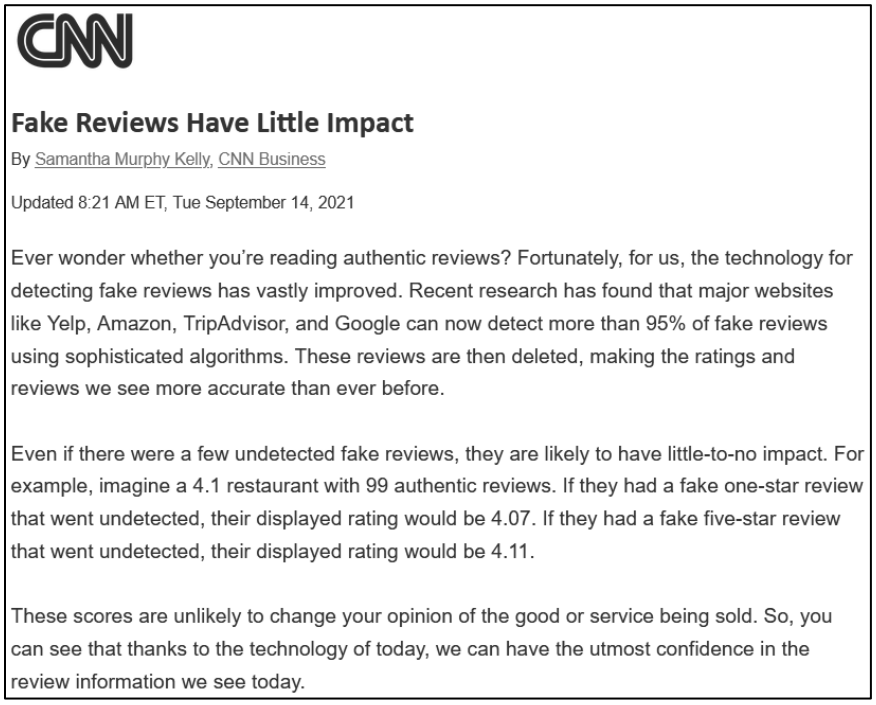
Updated 8:21 AM ET, Tue September 14, 2021

Disney announced the remainder of its 2021 films will be released exclusively in theaters before streaming on Disney+.

The decision by Disney on Friday shows that the company is optimistic about movie theater audiences returning, despite Covid-19 spikes caused by the Delta variant. "As confidence in moviegoing continues to improve, we look forward to entertaining audiences in theaters, while maintaining the flexibility to give our Disney+ subscribers the gift of Encanto this holiday season," Disney Media & Entertainment Distribution Chairman Kareem Daniel said in a statement.

"Encanto," an animated film about a magical family in Colombia with music by Lin-Manuel Miranda, will be available exclusively in theaters for a month following its November 24 release. It will be on Disney+ on Christmas Eve.

Medium PK Disconfirmation Prime



Fake Reviews Have Little Impact

By Samantha Murphy Kelly, CNN Business

Updated 8:21 AM ET, Tue September 14, 2021

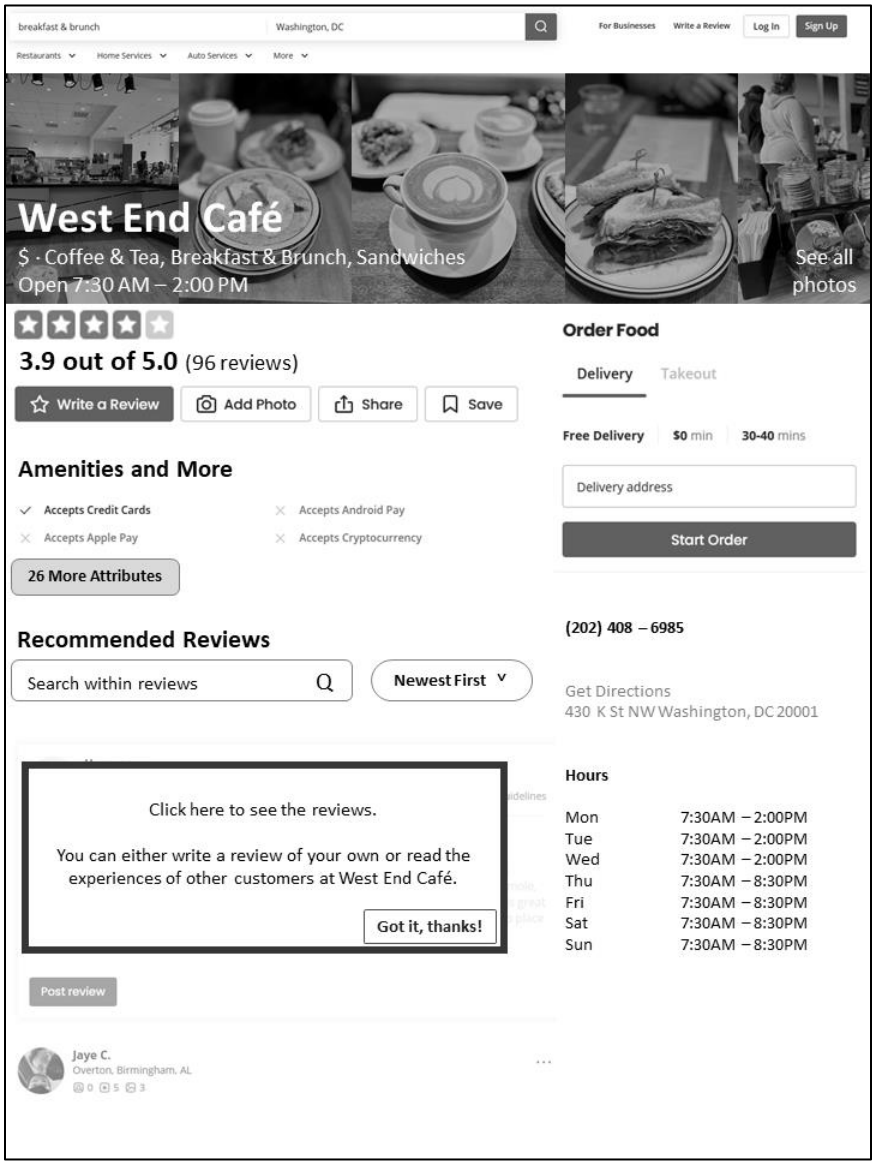
Ever wonder whether you're reading authentic reviews? Fortunately, for us, the technology for detecting fake reviews has vastly improved. Recent research has found that major websites like Yelp, Amazon, TripAdvisor, and Google can now detect more than 95% of fake reviews using sophisticated algorithms. These reviews are then deleted, making the ratings and reviews we see more accurate than ever before.

Even if there were a few undetected fake reviews, they are likely to have little-to-no impact. For example, imagine a 4.1 restaurant with 99 authentic reviews. If they had a fake one-star review that went undetected, their displayed rating would be 4.07. If they had a fake five-star review that went undetected, their displayed rating would be 4.11.

These scores are unlikely to change your opinion of the good or service being sold. So, you can see that thanks to the technology of today, we can have the utmost confidence in the review information we see today.

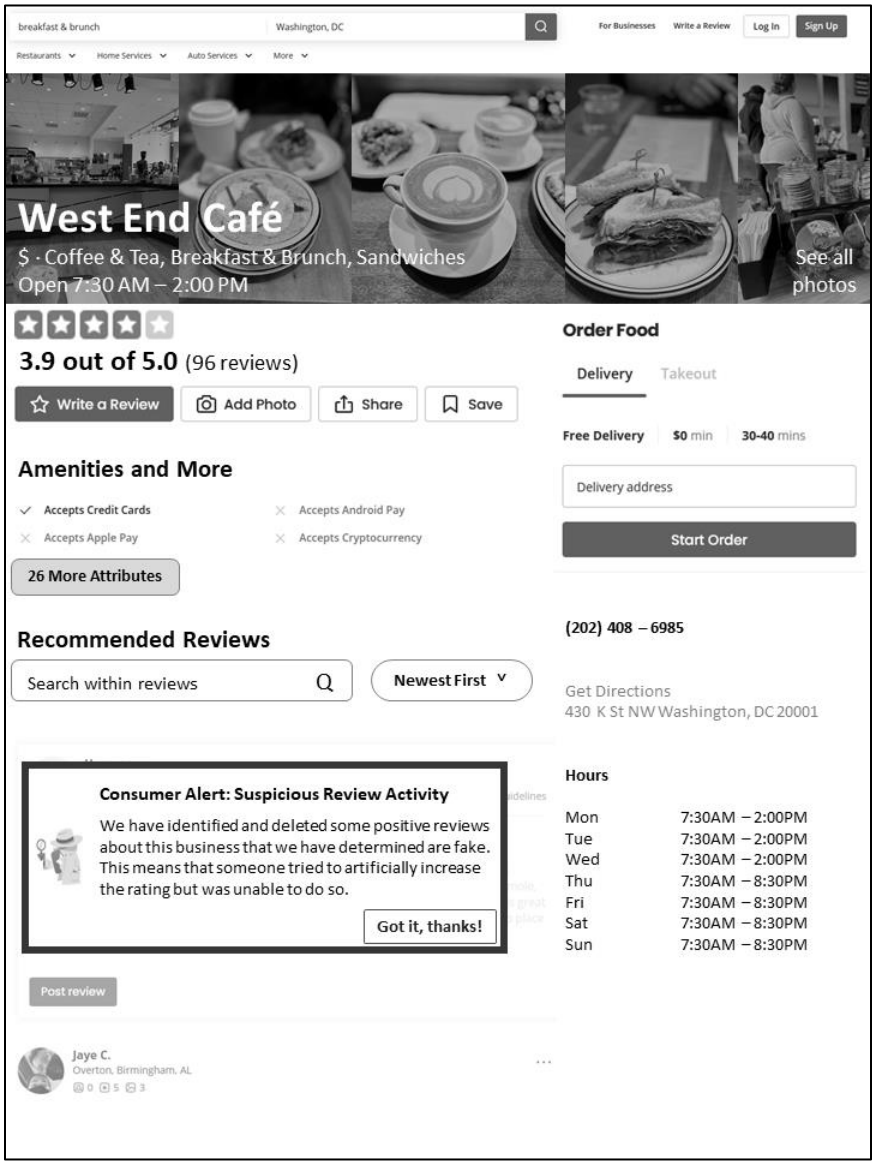
Study 3

Alert Absent Condition



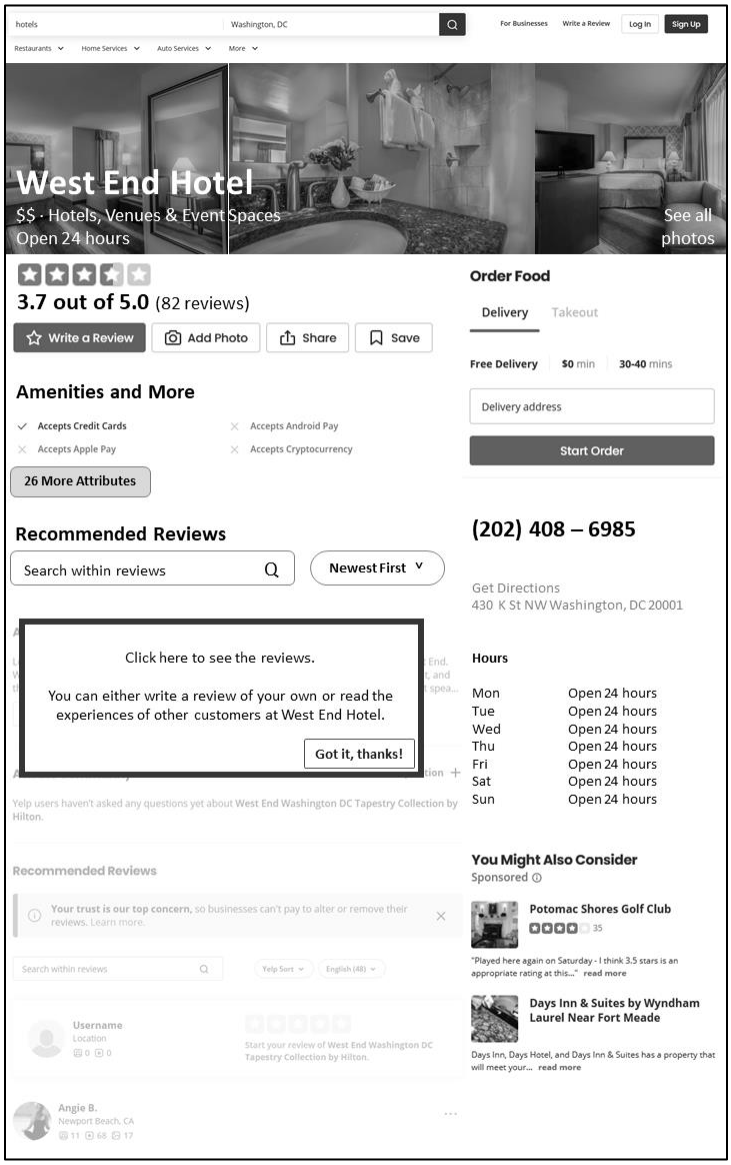
Study 3

Alert Present Condition



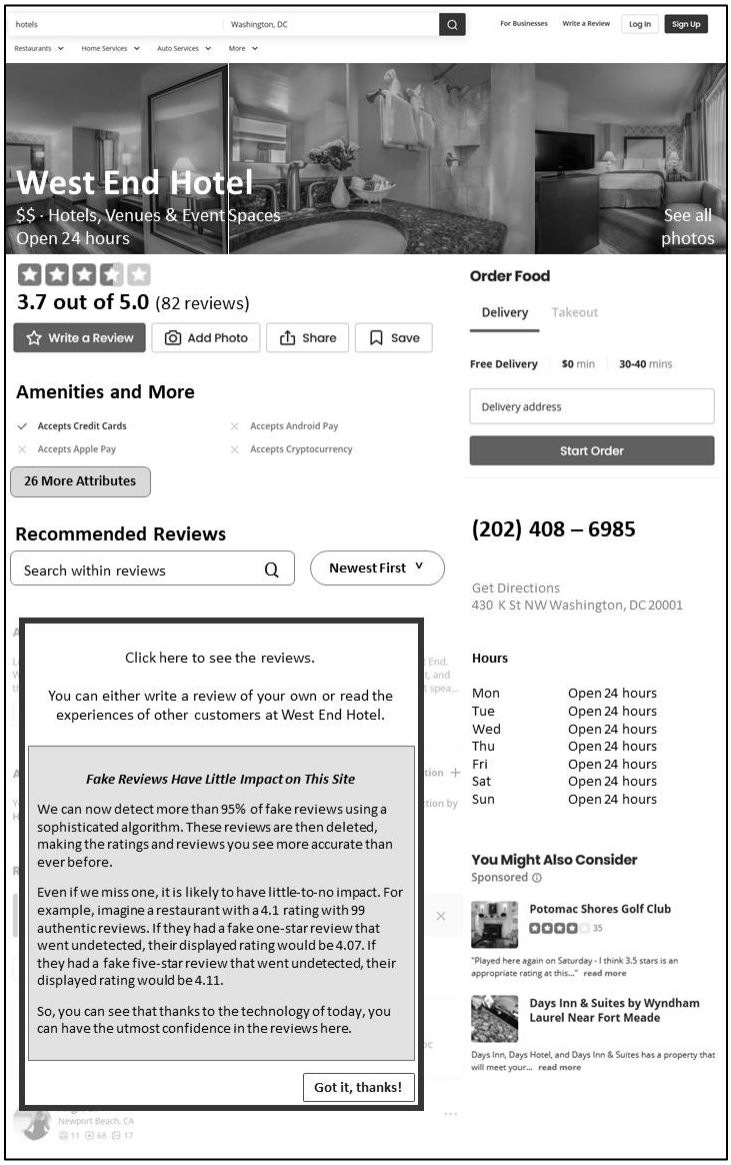
Study 4

PK Disconfirmation Absent, Alert Absent



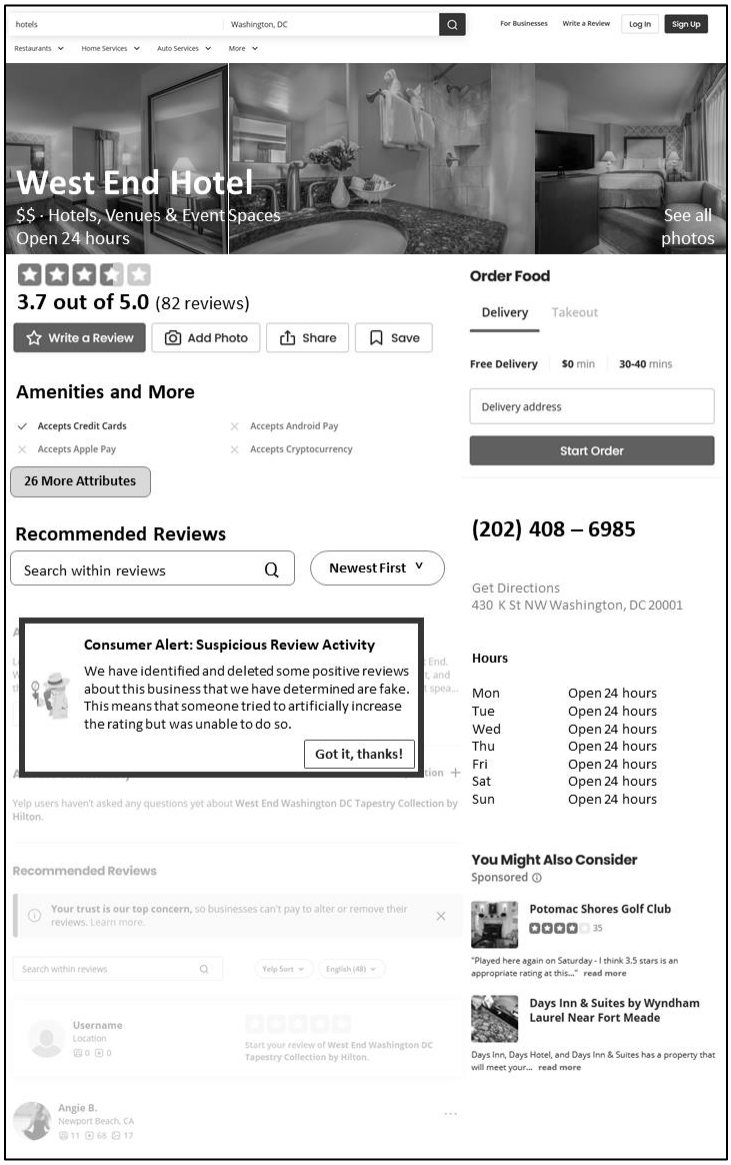
Study 4

PK Disconfirmation Present, Alert Absent



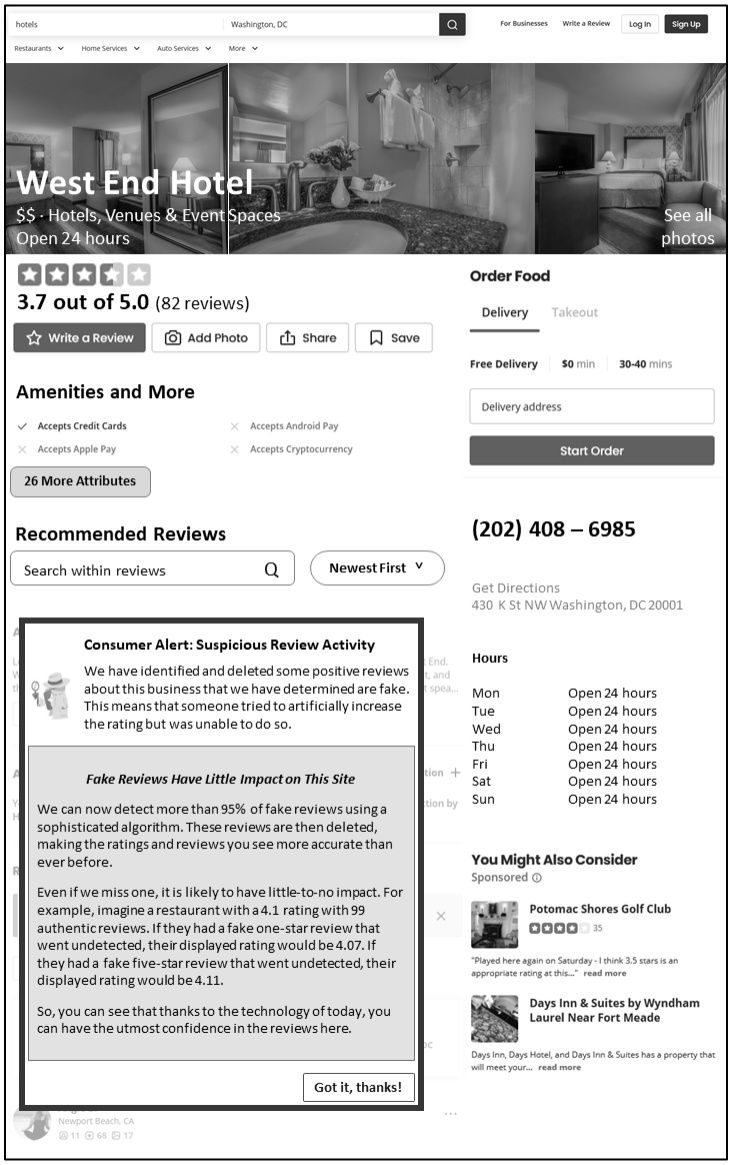
Study 4

PK Disconfirmation Absent, Alert Present



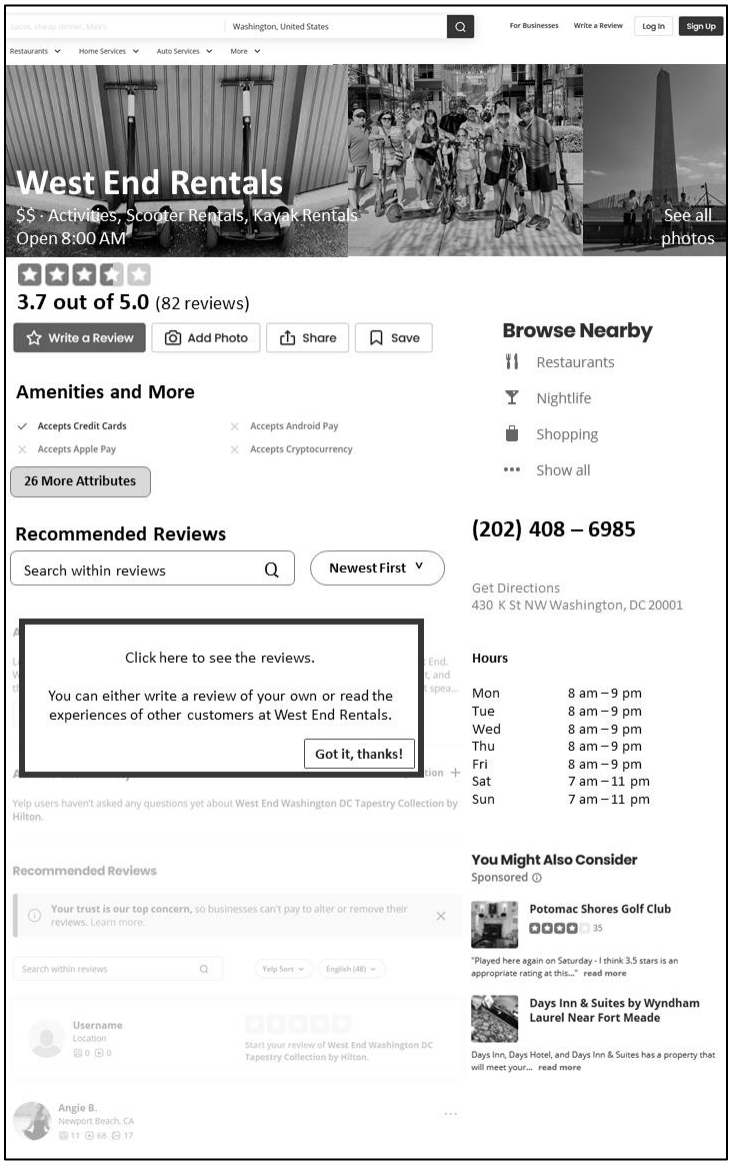
Study 4

PK Disconfirmation Present, Alert Present



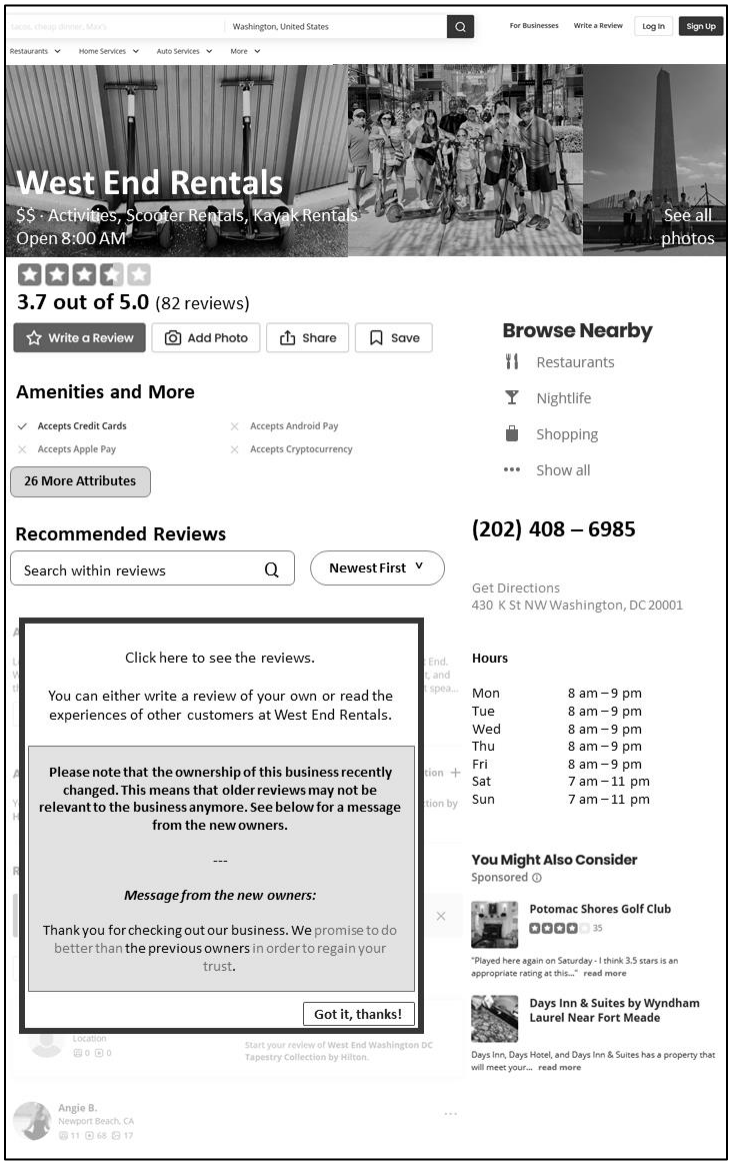
Study 5

PK Applicability High, Alert Absent



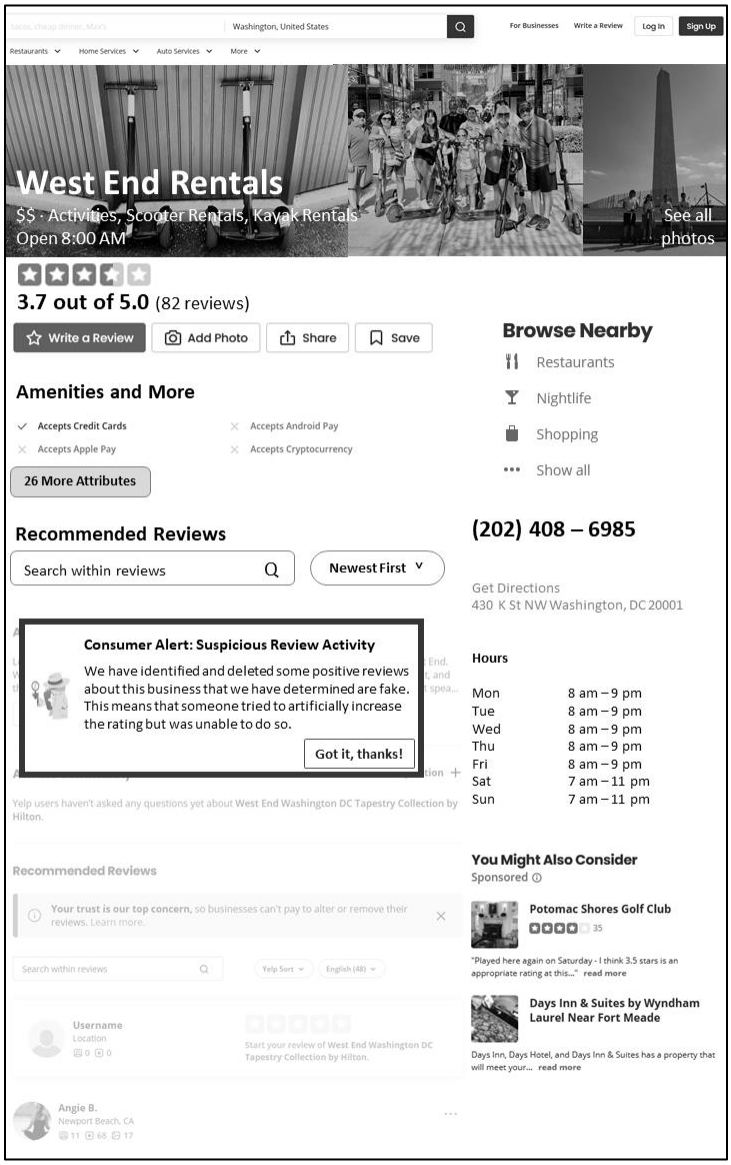
Study 5

PK Applicability Low, Alert Absent



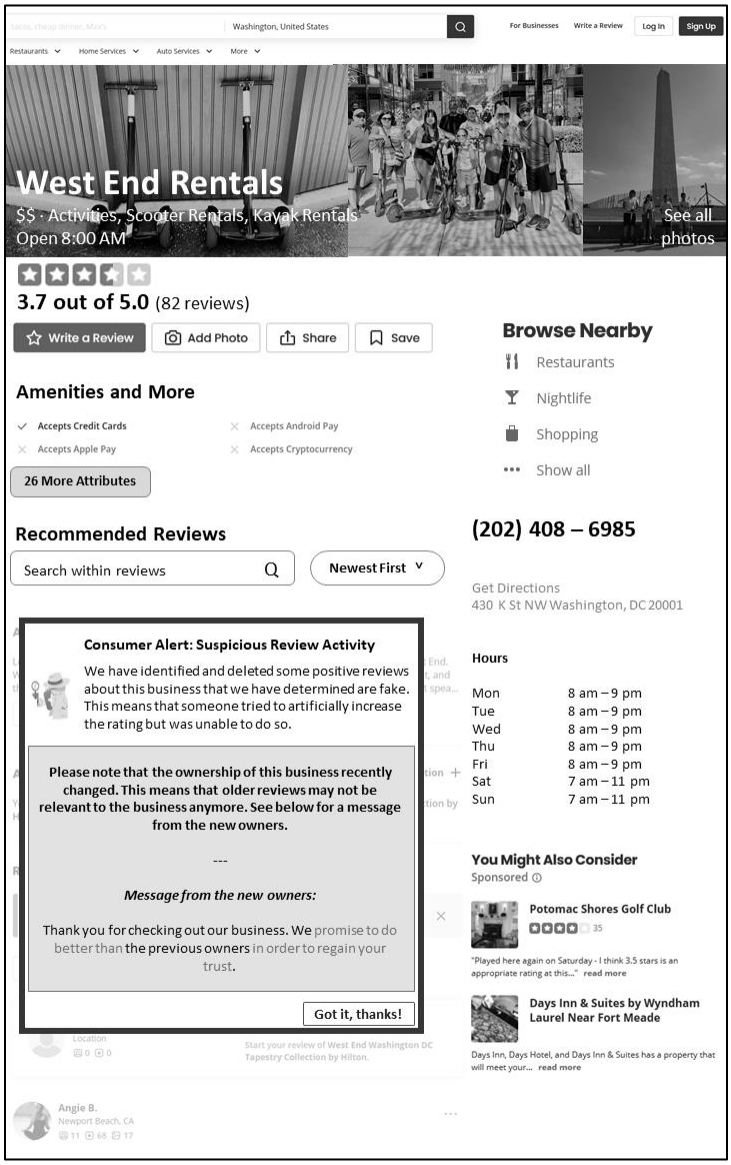
Study 5

PK Applicability High, Alert Present



Study 5

PK Applicability Low, Alert Present



WEB APPENDIX C – Additional Results for Study 3

Table WC.1 - Moderated serial mediation of perceived product rating via expected dishonesty and ratings bias (PROCESS macro (model 85; Hayes 2017)

	B	95% confidence interval
Index of Moderated Serial Mediation (expected brand dishonesty → perceived ratings bias)	.0792	[.0301, .1369]
Serial Mediation: PK Disconfirmation Absent	-.1217	[-.181, -.0733]
Serial Mediation: PK Disconfirmation Present	-.0425	[-.0807, -.0111]
Index of Moderated Mediation (expected brand dishonesty)	.05	[.003, .1251]
Simple Mediation: PK Disconfirmation Absent	-.0768	[-.1673, -.0068]
Simple Mediation: PK Disconfirmation Present	-.0269	[-.0652, -.0013]
Index of Moderated Mediation (perceived ratings bias)	.1672	[.0701, .2814]
Simple Mediation: PK Disconfirmation Absent	-.2244	[-.335, -.1277]
Simple Mediation: PK Disconfirmation Present	-.0572	[-.1168, -.004]

WEB APPENDIX D – Additional Results for Study 4

Additional Measures

Perceived brand manipulativeness. A 2x2 ANOVA on the perceived brand manipulativeness scale yielded significant main effects of the alert $F(1,394) = 100.055$; $p < .001$; $\eta^2_{partial} = .203$) and PK disconfirmation ($F(1,394) = 17.946$; $p < .001$; $\eta^2_{partial} = .044$) qualified by the interaction ($F(1,394) = 4.412$; $p = .036$; $\eta^2_{partial} = .011$). Planned contrasts demonstrate that in the absence of the PK disconfirmation, the presence of a positive alert significantly increased expectations of a brand's manipulativeness ($M_{absent} = 2.38$, $M_{positive} = 3.41$; $F(1,394) = 72.881$; $p < .001$; $\eta^2_{partial} = .156$). While in the presence of PK disconfirmation, the effect of the alert persisted but was attenuated ($M_{absent} = 2.20$, $M_{positive} = 2.87$; $F(1,394) = 31.38$; $p = .001$; $\eta^2_{partial} = .074$).

Comparing simple effects of PK disconfirmation within alert conditions, when the alert was absent, PK disconfirmation did not significantly influence expectations of manipulation ($F(1,394) = 2.27$; $p = .133$; $\eta^2_{partial} = .006$). However, when the positive alert was present, the PK disconfirmation led to reduced perceptions of manipulativeness ($F(1,394) = 20.177$; $p < .001$; $\eta^2_{partial} = .049$).

Full Mediation Results

Table WD.1 - Serial mediation of brand intentions via expected dishonesty and ratings bias (PROCESS macro (model 6; Hayes 2017)

	B	95% confidence interval
Index of Serial Mediation (expected brand dishonesty → perceived ratings bias)	-.1086	[-.1887, -.0508]
Simple Mediation (expected brand dishonesty)	-.4875	[-.6457, -.3499]
Simple Mediation (perceived ratings bias)	-.0868	[-.1653, -.0266]

Table WD.2 - Moderated serial mediation of perceived product rating via expected dishonesty and ratings bias (PROCESS macro (model 85; Hayes 2017)

	B	95% confidence interval
Index of Moderated Serial Mediation (expected brand dishonesty → perceived ratings bias)	.0155	[-.0099, .0482]
Serial Mediation: PK Disconfirmation Absent	-.0651	[-.1093, -.027]
Serial Mediation: PK Disconfirmation Present	-.0425	[-.0856, -.0206]
Index of Moderated Mediation (expected brand dishonesty)	.0461	[-.033, .126]
Simple Mediation: PK Disconfirmation Absent	-.1935	[-.2805, -.1178]
Simple Mediation: PK Disconfirmation Present	-.1474	[-.2326, -.0799]
Index of Moderated Mediation (perceived ratings bias)	.0535	[.0016, .1167]
Simple Mediation: PK Disconfirmation Absent	-.0739	[-.1367, -.0257]
Simple Mediation: PK Disconfirmation Present	-.0205	[-.0693, -.0176]

WEB APPENDIX E – Additional Results for Study 5

Additional Measures

Perceived brand manipulativeness. A 2x2 ANOVA on ulterior brand motives yielded a significant main effect of the alert $F(1,399) = 77.33$; $p < .001$; $\eta^2_{partial} = .162$), qualified by the interaction ($F(1,399) = 7.18$; $p = .008$; $\eta^2_{partial} = .018$). The main effect of the ownership change was not significant ($F(1,399) = 2.235$; $p = .136$; $\eta^2_{partial} = .006$). In the absence of ownership change, a positive alert increased expected brand manipulativeness ($M_{absent} = 2.43$, $M_{positive} = 3.32$; $F(1,399) = 65.651$; $p < .001$; $\eta^2_{partial} = .141$). When ownership change was salient, the effect of an alert was attenuated ($M_{absent} = 2.53$, $M_{positive} = 3.00$; $F(1,399) = 18.74$; $p < .001$; $\eta^2_{partial} = .045$).

Comparing simple effects of an ownership change within alert conditions, when the alert was absent, ownership change did not impact expectations of brand manipulation ($F(1,399) = .403$; $p = .403$; $\eta^2_{partial} = .002$). When the positive alert was present, a change of ownership yielded significantly lowered the expected brand manipulation ($F(1,399) = 8.736$; $p = .003$; $\eta^2_{partial} = .021$).

Full Mediation Results

Table WE.1 - Moderated serial mediation of perceived product rating via expected dishonesty and ratings bias (PROCESS macro (model 85; Hayes 2017)

	B	95% confidence interval
Index of Moderated Serial Mediation (expected brand dishonesty → perceived ratings bias)	.0326	[.0046, .0718]
Serial Mediation: PK Disconfirmation Absent	-.0624	[-.1071, -.0293]
Serial Mediation: PK Disconfirmation Present	-.0298	[-.0578, -.009]
Index of Moderated Mediation (expected brand dishonesty)	.1328	[.0211, .2528]
Simple Mediation: PK Disconfirmation Absent	-.2538	[-.3603, -.1588]
Simple Mediation: PK Disconfirmation Present	-.121	[-.2111, -.0412]
Index of Moderated Mediation (perceived ratings bias)	.0901	[.0226, .1761]
Simple Mediation: PK Disconfirmation Absent	-.1423	[-.2259, -.073]
Simple Mediation: PK Disconfirmation Present	-.0522	[-.1089, -.0064]

Table WD.2 - Moderated serial mediation of brand intentions via expected dishonesty and ratings bias (PROCESS macro (model 85; Hayes 2017)

	B	95% confidence interval
Index of Moderated Serial Mediation (expected brand dishonesty → perceived ratings bias)	.0559	[.0079, .1211]
Serial Mediation: PK Disconfirmation Absent	-.1069	[-.1763, -.0538]
Serial Mediation: PK Disconfirmation Present	-.051	[-.0967, -.0171]
Index of Moderated Mediation (expected brand dishonesty)	.2398	[.0403, .4443]
Simple Mediation: PK Disconfirmation Absent	-.4584	[-.6341, -.2969]
Simple Mediation: PK Disconfirmation Present	-.2186	[-.3742, -.0761]
Index of Moderated Mediation (perceived ratings bias)	.1543	[.0403, .2806]
Simple Mediation: PK Disconfirmation Absent	-.2438	[-.3572, -.146]
Simple Mediation: PK Disconfirmation Present	-.0895	[-.1756, -.0132]