

How Fake Review Alerts Help the Platform

Jared Watson, Assistant Professor of Marketing, Leonard N. Stern School of Business, New York University, 40 W 4th St, Tisch 809, New York, NY 10012, jw5798@stern.nyu.edu

Ted Matherly, Assistant Professor of Marketing, Price School of Business, University of Oklahoma, 1320 Brookside Drive, Norman, OK 73072, ted@ou.edu

Amna Kirmani, Ralph J. Tyser Professor of Marketing, Robert H. Smith School of Business, University of Maryland, 3467 Van Munching Hall, College Park, MD 20742, akirmani@umd.edu

Corresponding Author: Jared Watson

Author Notes: N/A

Acknowledgements: The authors would like to thank many parties for the current manuscript –the NYU marketing group, UMD marketing group, and those who have provided insights at ACR/SCP/CBSIG conferences, and brownbag and seminar presentations.

Statements and Declarations

Ethical Considerations & Consent to Participate

The study was approved by the New York University Institutional Review Board (IRB FY2018-2303) on September 13, 2018. All participants provided written informed consent prior to participating in each survey.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding Statement

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data Availability Statement

The data that support the findings of this article are available upon request.

How Fake Review Alerts Help the Platform

Abstract

Some review platforms, such as Yelp and TripAdvisor, use fake review alerts in their fight against review fraud. These are disclosures by the platform that a business has attempted to manipulate the reviews on a platform but were caught by the platform. This paper employs an empirics-first approach, first documenting the scope of review manipulation for businesses that receive alerts, then documenting the effect of these alerts on various outcomes including business and platform judgments, and subsequent review quality. We find that alerted businesses are penalized by consumers while platforms receive reputational benefits. Most interestingly, consumers who write reviews for an alerted business go on to write higher-quality reviews for subsequent businesses. This benefits the platform, the business which receives the review, and consumers who read it. This effect is robust to various measures of subjective and objective review quality, and replicates across field and lab data. Thus, the potential long-term benefits for the platform and its stakeholders from disclosing attempted review fraud suggest that more platforms should actively do so rather than employing strategies commonly used by many platforms today, such as silent removal of fraudulent reviews.

Keywords: product reviews, platforms, opinion leaders, opinion seekers, influencers, online word-of-mouth

Introduction

“We have noticed suspicious review activity for this business. This sort of activity can take many forms, including when a number of positive reviews originate from the same IP address ... we wanted to call this to your attention because someone may be trying to artificially inflate the rating for this business.” – Yelp (2025)

Fraudulent reviews are an endemic problem for digital commerce platforms. In a typical year, popular service review platforms Yelp and TripAdvisor remove up to 9% of the submitted reviews for being fraudulent or for lacking credibility (Yelp 2025, TripAdvisor 2025). Meanwhile, industry estimates suggest that around 30% of online reviews may be fake, leading to consumer losses of over \$770 billion worldwide (CapitalOne 2025). While there is some debate as to what constitutes a fake review, it is widely recognized that reviewers may have conflicts of interest due to financial incentives (Ham et al. 2021) or relational factors with businesses (Kim, Chung, and Lim 2019) that may degrade the quality or trustworthiness of reviews. Moreover, the U.S. Federal Trade Commission (FTC) defines fake reviews as those written with the intent to mislead or deceive, including reviews where compensation is provided conditional on a particular sentiment (positive or negative). These practices are illegal in the U.S. under FTC rules (FTC 2024) and cause harm to all stakeholders. For consumers, fake reviews increase the risk of acting on unreliable or misleading information, affecting over \$300 billion in annual U.S. sales, equivalent to \$2,400 per U.S. household (Huffman 2024). Simultaneously, honest businesses are at a distinct disadvantage when competitors use fake reviews to increase their own ratings, pressuring these honest businesses to engage in fraudulent behavior to compete. Finally, platforms rely on their reputation for providing accurate information; to the extent that customers doubt the veracity of the information, fake reviews represent an existential threat to platforms.

To address the problem of fake reviews, leading platforms use sophisticated algorithms to identify and remove such reviews. Although this approach can stop approximately 75% of fake reviews from reaching consumers (TripAdvisor 2023), many argue that platforms are not doing enough and should be held liable (Dean 2024). As an additional measure, some platforms supplement their efforts to remove fraudulent reviews with fake review alerts. Fake review alerts appear on a business' subpage within the platform and highlight that the platform has identified and removed fraudulent reviews for that business. For example, Yelp's "Consumer Alerts" and TripAdvisor's "red penalty badge" programs both inform consumers that the platform has identified potential review fraud related to a specific business and denote this on that business's page within their platform. However, other platforms, such as Amazon, quietly remove fraudulent reviews without alerting consumers. This raises the question of the extent to which these alerts are beneficial to consumers and to the platform. On the one hand, fake review alerts can inform consumers that the platform has done its due diligence and is punishing offending businesses, which should make consumers more confident about the veracity of reviews. On the other hand, such alerts also raise the salience of review fraud, which may increase consumer suspicion about the veracity of reviews on the platform. Thus, understanding the impact of fake review alerts has important theoretical and practical implications.

In this paper, we use an empirics-first approach (Golder et al. 2023) to examine the consequences of using fake review alerts for different stakeholders: the review platform, businesses on the platform, and consumers who use the platform. We focus on the following questions: How does review fraud for businesses who receive fake review alerts unfold before the alert is displayed, while it is displayed, and after it is removed? Do alerts benefit or harm perceptions of the platform? Do alerts change how consumers process business information? Do

they change how consumers who contribute reviews to the platform behave subsequently? The answers to these questions would inform practitioners about the advisability of using fake review alerts.

Using analyses of secondary data as well as experiments, we find that fake review alerts affect the business and platform differently. For businesses that receive alerts, ratings decrease immediately. Thus, the alert helps consumers avoid potentially dishonest businesses. In terms of the platform, the alert benefits the platform in several ways: 1) consumers perceive the platform as having higher integrity; 2) reviewers exposed to the alert subsequently post higher quality reviews for other businesses on the platform, which suggests that the quality of reviews on the platform improves.

Besides the practical implications, the paper contributes to the literature on fake reviews. Researchers in computer science have focused on developing detection algorithms (for a review, see Wu et al. 2020) as well as investigating the ability of humans and artificial intelligence to identify fake reviews (Meng et al. 2025; Walther et al. 2023). In the marketing literature, researchers have sought to identify factors that increase the likelihood of a business committing review fraud. For instance, independent (vs. chain) businesses 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, the effects are often exacerbated when independent businesses have a weak reputation (Luca and Zervas 2016). We also extend experimental work by Beck, Wuyts, and Jap (2023), various signals of review (e.g., authenticity verification) and reviewer credibility (e.g., credibility badges) increase platform trust, by showing that these disclosures can have beneficial second-

order effects for the platform by encouraging quality contributions from reviewers. By pointing out the benefits of such a strategy, we enhance our understanding of fake reviews in general.

Fake Reviews and Alerts

Fake or fraudulent reviews include those that are intended to mislead consumers, such as those that are biased due to undisclosed incentives, those not based on actual experience, and those generated by bots. A fake review alert informs consumers that the platform has found and removed fake reviews from the business' page. Yelp began to use fake review alerts in 2012 and, in 2023, published a list of all businesses which had received an alert since the program's inception. What has been the consequence of these alerts? Since most fake reviews are positive (Sullivan 2022), these alerts have likely led to the removal of mostly fake positive reviews, thereby decreasing the business's rating. But alerts only last for 90 days on Yelp, so it is unclear what happens once that alert is removed. Does the business engage in subsequent review fraud or does the alert effectively staunch the production of more fake reviews? How long-lasting are the effects of the alert? Does the alert change consumer behavior? What are the effects on the platform?

From the consumer perspective, a fake review alert is likely to increase suspicion of the business, thereby activating persuasion knowledge and leading to perceptions that the business is untrustworthy (Campbell and Kirmani 2000; Darke and Ritchie 2007). This might lead consumers to avoid the business and to abandon the platform that hosts fake reviews. Although fake review alerts attempt to reassure consumers that the platform took the initiative to identify and remove fraudulent reviews, they might have the opposite effect. On the one hand, fake review alerts may lower platform perceptions. A persuasion knowledge account (Friestad and Wright 1994; Kirmani and Zhu 2007) would suggest that fake review alerts may harm platforms

by making the potential presence of fake reviews salient. In other words, even if consumers were not thinking about fraud, the alert makes fraud salient, raising suspicion and leading to defensive processing. With consumers becoming increasingly sensitive to the presence of fake reviews (Pitman 2022), a fake review alert could have negative consequences for the platform by indicating that the reviews of all businesses are biased.

On the other hand, the alert may increase perceptions of the platform's integrity because it has publicized and addressed the fake reviews. Prior experimental work has shown that awareness that the platform may publicize fraudulent review activity enhances trust (Beck, Wuyts, and Jap 2023). But seeing an alert on a page may have a stronger effect, as it is effectively a costly signaling action (Kirmani and Rao 2000) where the platform is asserting that it has successfully addressed the problem of fraudulent reviews. This implies that the platform expects the alerts will lead to more positive perceptions of the platform. Given the competing possibilities, we examined secondary data on fake review alerts.

Study 1: Effect of an Alert on Business Ratings

The goal of the first study was to contextualize fake review alerts. Specifically, we intended to investigate the nature of businesses that receive alerts, to compare them against other businesses operating on the platform, and to explore the dynamics of ratings surrounding the issuing of an alert. We focused our analysis on Yelp, which released a list of all businesses on its platform which had a received fake review alert. When Yelp issues an alert for a business, the offending reviews are removed from the average rating calculation and partitioned from the other reviews (though they are technically still accessible on a different area on the platform).

We gathered all available reviews for businesses that had received fake review alerts between 2012 and 2023. Yelp distinguishes between two types of fake review alerts:

Compensated Activity and Suspicious Review Activity. A Compensated Activity alert warns that Yelp has caught “someone offering payment in the form of cash, discounts, gift certificates or other incentives in exchange for someone to write, change, prevent or remove reviews...” (Yelp n.d.). A Suspicious Review Activity alert warns that Yelp has detected “a number of positive reviews originate from the same IP address, or when we’ve identified reviews resulting from a possible deceptive review ring...” (Yelp n.d.). Given that the data identified these two alert types (Yelp 2023), we decided to examine whether they led to different effects.

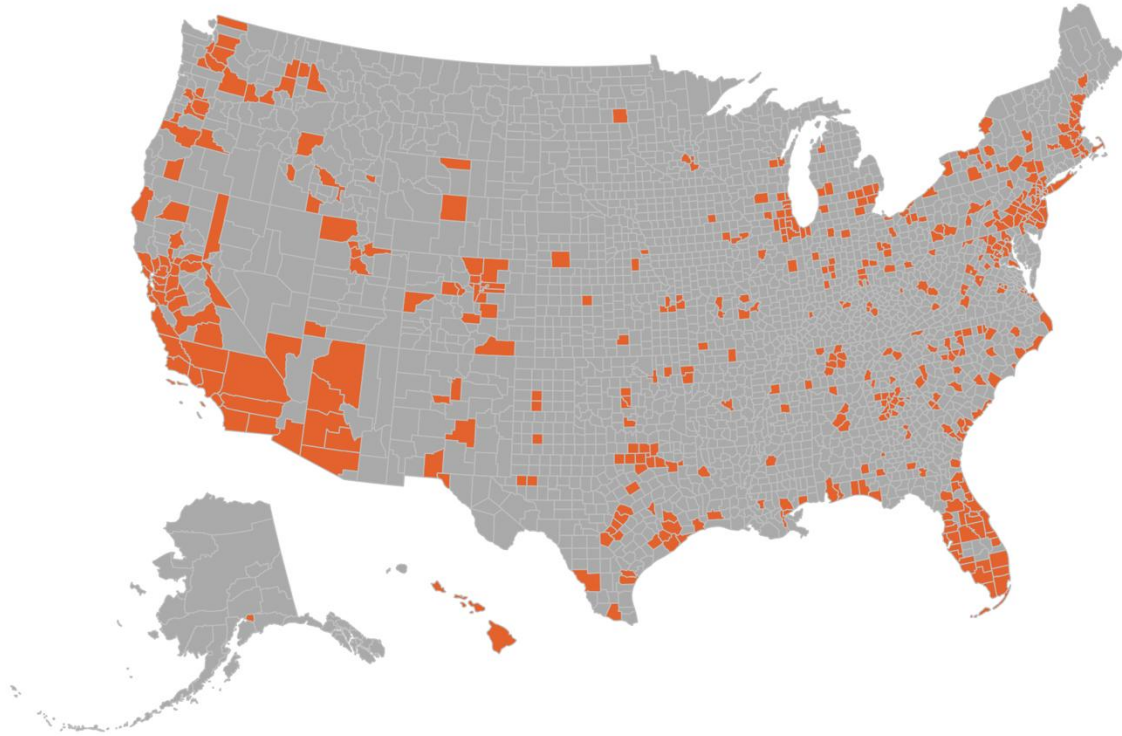
Methodology

Data

We gathered all available reviews for the 4963 businesses in the focal set from within the Yelp platform, totaling 749,514 reviews. We begin with model-free analyses of the data by characterizing the distribution of the alerts across geography and time. As shown in Figure 1, businesses in 480 counties representing 48 U.S. states and the District of Columbia received alerts (no businesses in Montana, South Dakota or Vermont received alerts). Consistent with prior findings, increased competition yielded increased review fraud (Luca and Zervas 2016; Mayzlin, Dover, and Chevalier 2014). Specifically, our analysis found that counties that had businesses that received alerts were more densely populated ($M = 503$ per square kilometer) than both the U.S. average ($M = 106$) and that of other counties where businesses were represented on Yelp but did not receive alerts ($M = 148$). However, the alerts were otherwise widespread, suggesting that the phenomenon was not-region specific.

Figure 1: Geographic Distribution of Businesses that Received Yelp Fake Review Alerts

Counties with businesses that **received alerts** were widely distributed across the United States



Examining the alerts over time, the earliest alerts were issued in October 2012, and the last observed alert in August 2023, covering a period of more than 10 years. Alerts were generally issued once per quarter (see Figure 2), on the first day of the middle month, though they were sometimes issued at other times in the observation period. The general trend was a steady increase over time, and there was a large spike in alerts issued on February 1, 2020, though the reason for this increase was unclear. Aside from this specific period, the alerts were well distributed across time periods. Taken together, these findings suggest that the issuing of fake review alerts was unlikely to be driven by factors outside of the platform’s ecosystem.

Figure 2: Number of Yelp Fake Review Alerts Issued per Month During Observation Window

Alerts were generally issued once per quarter, and slowly increased in number over time

	J	F	M	A	M	J	J	A	S	O	N	D
2012										9	20	
2013				59			3	50		2		109
2014		39			35				60			
2015	60	26			1	45			44			
2016		53			48			107		2	124	
2017		79			112			130			128	
2018		143			150			130	1	2	186	
2019		160			190			207			158	
2020	138	561			155			96			91	
2021		103			95			106		65	41	
2022		98			90			118			120	
2023		88			122	17	23	164				

Next, we turned our focus to the businesses that received alerts. Prior work identifies several characteristics common to businesses that engage in review manipulation: they are independent businesses that are not part of chains (Mayzlin, Dover, and Chevalier 2014); they have relatively few reviews prior to engaging in manipulation (He, Hollenbeck, and Proserpio 2021); and they have poor reputations prior to doing so (Luca and Zervas 2016). Testing for the presence of these characteristics necessitates comparison against a set where it is unlikely that review manipulation is present. To this end, we employed the Yelp Open Dataset, which is provided by Yelp for research, and characterized as “real-world data” including almost seven million reviews of 150,000 businesses. As these reviews were selected by Yelp to be representative of the platform, it is likely that these are high quality reviews for businesses

unlikely to be engaged in manipulation of the platform. We used this set as a baseline for both the descriptive analysis and the modeling reported subsequently.

We first examined the percentage of businesses that were chains in the two sets. Among those that received alerts, 10.4% of businesses were coded as chains, while 7.4% of those in the comparison set were chains. While this is inconsistent with the relationship observed in Mayzlin, Dover, and Chevalier (2014), it is possible that this difference is due to the broader set of businesses within the present data. While the Yelp data covers a wide variety of industries, Mayzlin and colleagues (2014) focus on hotels, where 83% of the businesses had an affiliation with a large hotel chain.

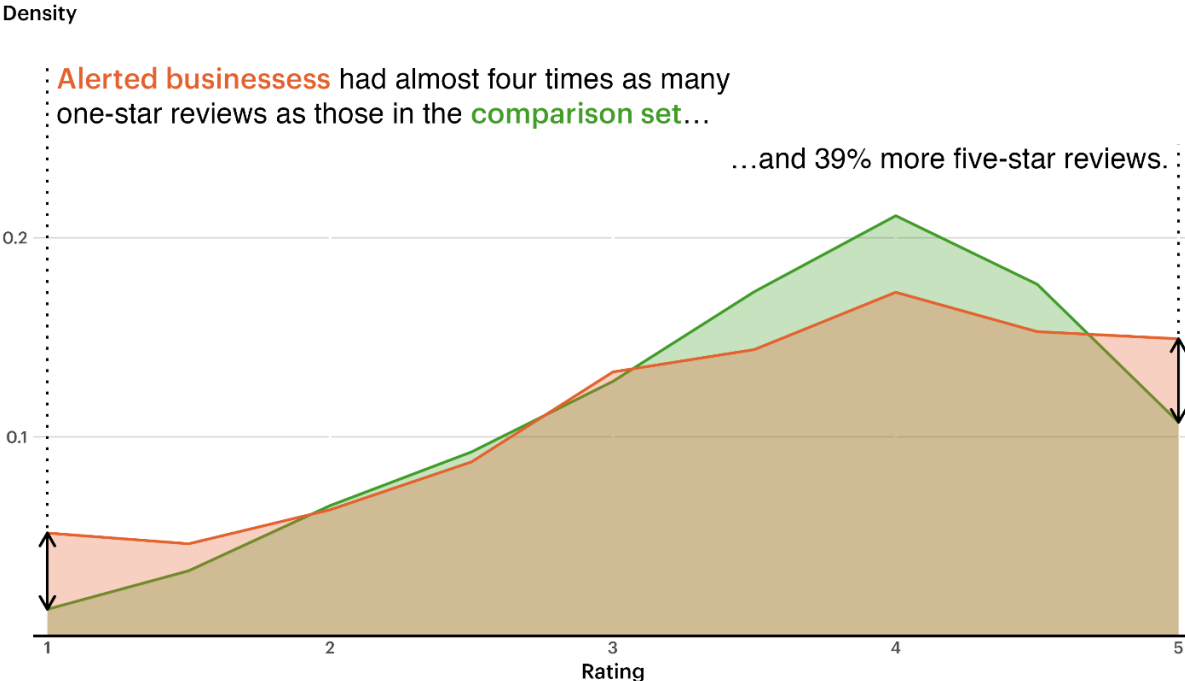
Next, we considered the “cold start” problem, where businesses use fraudulent reviewing activities to build their reputations initially. As prior work has shown, businesses using fraudulent reviewing activity to establish their reputation are likely to have a high number of initial positive reviews (He, Hollenbeck, and Proserpio 2021). In our data, 51.0% of the first reviews received by alerted businesses were five-star reviews; in contrast, the overall rate of five-star reviews among the comparison set was 43.0%, suggesting that alerted businesses were more likely to have five-star reviews as their first review than the base rate among businesses unlikely to be engaged in review manipulation. Further, 22.2% of alerted businesses had all five-star reviews for their first five reviews and received 17.5% of their total reviews within one year of their first review, patterns consistent with these businesses engaging in review manipulation to establish their reputation on the platform.

Finally, we explore the reputations of the alerted businesses prior to the alerts. Alerted business had lower average ratings ($M = 3.49$) compared to businesses that did not receive alerts in the comparison set ($M = 3.59$). Although the difference in ratings was relatively small, the

distributions were markedly different, as shown in the density plot in Figure 3. Extreme ratings were much more common among businesses that received alerts, which were almost four times more likely to have one-star ratings than those in the comparison set (5.17% vs. 1.34%), and were also one and a half times more likely to have five-star ratings (14.9% vs. 10.7%).

Figure 3: Distributions of Ratings for Businesses that Received Alerts and the Comparison Set

Alerted businesses had more extreme reviews than the comparison set



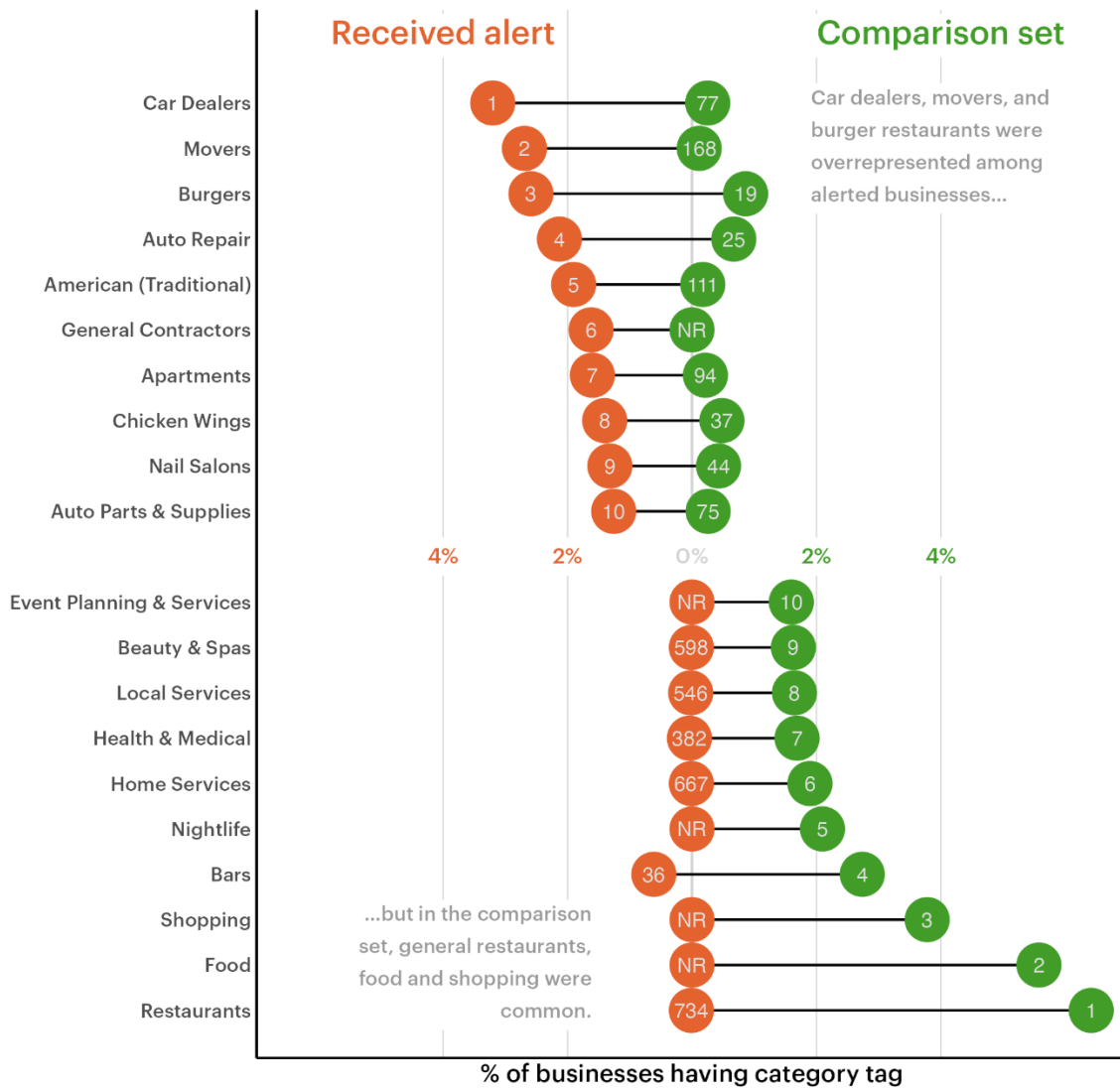
Finally, a comparison of the business category tags between the alerted businesses and the comparison set suggests that those that received alerts operated in different categories. As shown in Figure 4, the most common category tags among businesses that received alerts were car dealers, movers and burger restaurants, while none of the ten most common categories for alerted businesses ranked higher than 19th among the tags in the comparison set. Similarly, the most common categories among the comparison set were rare among the alerted businesses, supporting the conclusion that the alerted businesses represented a distinct subset of businesses on the platform. Moreover, the average rating of businesses from the comparison set in the most

common categories among alerted businesses was 3.54, substantially below the overall average in the comparison set of 3.75. This suggests that the categories within which the alerted businesses operated may have a worse reputation compared to other businesses on Yelp, providing an incentive for these businesses to engage in review manipulation.

Taken together, these relationships suggest that the businesses that received Yelp alerts were systematically different from others on the platform. However, these differences are largely consistent with the findings of prior work characterizing firms engaged in fraudulent review activity. To the extent that these differences, including the businesses' locations, their chain status, the categories they operate in, and their underlying quality – are time-invariant, business fixed effects should be sufficient to remove these differences as a source of variation in our model. We adopted this approach in the formal model which we describe in the next section, which allows us to estimate the dynamic nature of the review manipulation activity and the subsequent effect of the alert on a business's ratings.

Figure 4: Comparison of the Business Category Tags between Alerted Businesses and the Comparison Set.

The categories of businesses that received alerts were different from the comparison set



Model

To model the dynamics of reviews for alerted businesses before and after the alerts were issued, we adopted an event study model (Autor 2003; Miller 2023) using the comparison set as the baseline. To ensure an adequate number of observations for each time period, we focused on a subset of the overall data in a multi-year window around the time of the alert being posted, in which the 90-day period where the alert was active was standardized as time 0. We limited the

event panel to 20 quarters before this period and 12 quarters after this period. Combined with the comparison set, this yielded 7,576,829 total reviews for 154,566 businesses after removing singletons (Correia 2015). The outcome of interest in our model was the Yelp rating as recorded on a five-point scale, with higher values corresponding to more favorable ratings. We estimated the following model:

$$y_{it} = \mathbf{Pre}'_{it}\boldsymbol{\beta}_1 + \mathbf{Post}'_{it}\boldsymbol{\beta}_2 + \gamma R_{it} + \lambda_i + \tau_t + \epsilon_{it}$$

where

- y_{it} is the rating of the business i at time t ,
- $\mathbf{Pre}'_{it} = [d_{i,q}]$ is a vector of period indicators, with $q \in \{-20, \dots - 1\}$ and $d_{i,q} = 1$ during the q th quarter in the period prior to the alert and zero otherwise,
- $\mathbf{Post}'_{it} = [d_{i,q}]$ is a vector of period indicators, with $q \in \{0, \dots 12\}$ and $d_{i,q} = 1$ during the q th quarter in the period after the alert and zero otherwise,
- a_i is an indicator such that $a_i = 1$ if business i received a Compensated Activity alert, and $a_i = 0$ if it was a Suspicious Activity alert,
- R_{it} was the running average review for business i up to time $t - 1$, such that $R_{it} = \frac{\sum_{t=0}^{t-1} y_{it}}{t-1}$,
- $\boldsymbol{\beta}_{1-2}$ were vectors of parameters capturing the difference in reviews between the alert period and all other time periods,
- γ was a parameter to be estimated,
- λ_i and τ_t were fixed effects for business i and time t ,
- ϵ_{it} was the disturbance term.

The 0th quarter corresponds to the displayed alert period (i.e., when the alert appeared publicly). We include the running average review to control for the effect of the current rating of

the business on reviewers' evaluation. As described above, we include business fixed effects to control for time-invariant business heterogeneity. We also include year-month fixed effects to account for potential seasonal or annual variations in reviewing. We estimate the model using OLS, with heteroskedasticity-robust standard errors clustered on businesses and report the results around the questions of interest.

Results

Question 1: How does the manipulation of reviews manifest on the platform?

The full results are reported in Table 1. We first present the results using the overall data in Column 1 and visualize the ratings and volume of reviews in Figure 5. Using the comparison set as the baseline, we observe no differences between the alerted and comparison businesses prior to 15 quarters ($q = -15$) before the alert was issued. From this point until the alert was issued, the increase in the ratings of alerted businesses became positive and significant and grew steadily more favorable over time. The ratings peaked two quarters prior to the alert being issued in a difference of .317, representing an 8.4% increase over the baseline. Immediately after the alert was issued ($q = 0$), ratings returned to the baseline levels, and no significant differences were observed in any of the 13 quarters following the alert. Thus, fake review alerts seem to curb the effect of fake reviews.

Question 2: Does the type of alert matter?

Next, we consider potential differences in the effects of the alert for businesses issued Suspicious ($N = 3214$) and Compensated Activity ($N = 1749$) alerts. For clarity, we split the data set and employ the comparison set as the baseline in both, though we obtain similar results when modeling these jointly with alert type by relative quarter interactions. First, we examine the patterns around Suspicious Activity fake review alerts, with the results reported in column 2.

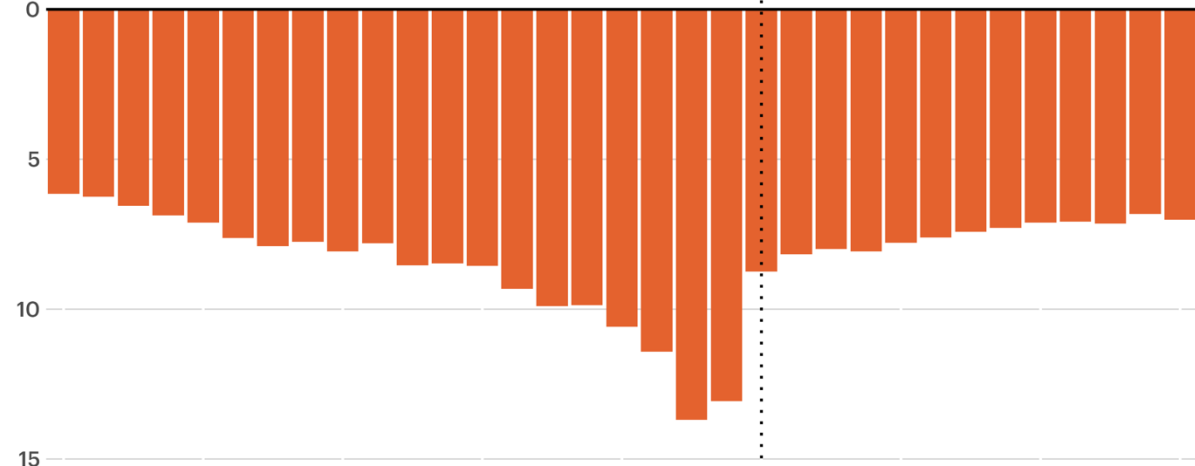
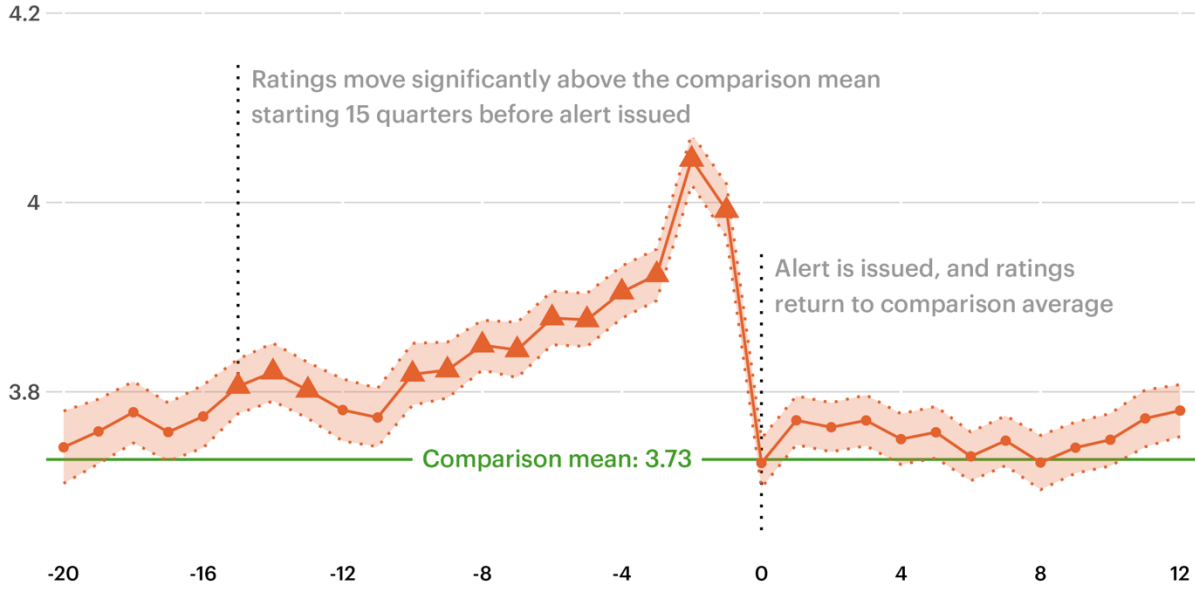
Here, we observe a first significant effect at $q = -14$, though this difference disappears until six quarters before the alert. We observe significant positive effects following the alert in $q = 1$, as well as the final two quarters, $q = 11$ and $q = 12$, though the size of these effects is modest compared to those in the year immediately prior to the alert.

In our last analysis, we only include those businesses which received Compensated Activity fake review alerts and report these results in column 3. We observe the first significant effect fifteen quarters before the alert was issued ($q = -15$), resuming in $q = 11$ and building until the alert is issued. We do not observe any significant effects after the alert period ends. These results suggest that the businesses which received Compensated Activity alerts began engaging in fraudulent review practices earlier than those that received Suspicious Activity fake review alerts, and that their solicitation of fake reviews had a larger impact on their ratings. Yet, the results demonstrate that consumer response to Compensated Activity alerts is virtually indistinguishable from that of Suspicious Activity alerts, suggesting that the presence of the alert is more influential than the content of the alert.

Figure 5: Predicted Ratings and Number of Reviews for Businesses Surrounding a Yelp Fake Review Activity Alert

Businesses that received alerts quickly return to the overall average after the alert is issued

Business rating



Number of reviews

Shaded area represents ± 1 standard error around the model predictions.

Table 1: Study 1 Regression Results

Variables	(1)	(2)	(3)
	Overall	Suspicious Alerts	Compensated Alerts
q = -20	0.0129 (0.0385)	-0.0526 (0.0457)	0.123 (0.0662)
q = -19	0.0296 (0.0346)	-0.00274 (0.0466)	0.0875 (0.0523)
q = -18	0.0499 (0.0331)	0.0260 (0.0428)	0.0955 (0.0537)
q = -17	0.0289 (0.0317)	0.0201 (0.0398)	0.0491 (0.0535)
q = -16	0.0455 (0.0338)	0.0417 (0.0392)	0.0578 (0.0629)
q = -15	0.0772** (0.0297)	0.0628 (0.0378)	0.111* (0.0487)
q = -14	0.0923** (0.0315)	0.0943* (0.0388)	0.0977 (0.0549)
q = -13	0.0731* (0.0306)	0.0691 (0.0379)	0.0899 (0.0533)
q = -12	0.0523 (0.0338)	0.0201 (0.0387)	0.117 (0.0631)
q = -11	0.0443 (0.0319)	0.00764 (0.0418)	0.115* (0.0500)
q = -10	0.0900** (0.0338)	0.0627 (0.0424)	0.145* (0.0567)
q = -9	0.0944** (0.0307)	0.0494 (0.0399)	0.180*** (0.0491)
q = -8	0.121*** (0.0275)	0.0604 (0.0346)	0.231*** (0.0450)
q = -7	0.116*** (0.0304)	0.0557 (0.0414)	0.229*** (0.0439)
q = -6	0.149*** (0.0294)	0.0746* (0.0375)	0.273*** (0.0469)
q = -5	0.147*** (0.0293)	0.0890* (0.0367)	0.253*** (0.0477)
q = -4	0.177*** (0.0288)	0.125*** (0.0379)	0.270*** (0.0449)
q = -3	0.195*** (0.0285)	0.129*** (0.0369)	0.310*** (0.0432)
q = -2	0.317*** (0.0282)	0.286*** (0.0384)	0.385*** (0.0411)
q = -1	0.263*** (0.0292)	0.227*** (0.0409)	0.327*** (0.0418)
q = 0	-0.00376 (0.0273)	0.0219 (0.0357)	-0.0204 (0.0426)
q = 1	0.0413 (0.0270)	0.0736* (0.0355)	0.0102 (0.0408)
q = 2	0.0341 (0.0271)	0.0316 (0.0341)	0.0501 (0.0444)
q = 3	0.0412 (0.0280)	0.0528 (0.0358)	0.0353 (0.0453)
q = 4	0.0215 (0.0279)	0.0323 (0.0365)	0.0160 (0.0432)
q = 5	0.0287 (0.0284)	0.0568 (0.0365)	-0.00566 (0.0449)
q = 6	0.00329 (0.0269)	0.0357 (0.0343)	-0.0407 (0.0433)
q = 7	0.0199 (0.0273)	0.0574 (0.0352)	-0.0324 (0.0433)
q = 8	-0.00321 (0.0296)	0.00904 (0.0385)	-0.0198 (0.0457)
q = 9	0.0125 (0.0279)	0.0445 (0.0361)	-0.0396 (0.0434)
q = 10	0.0208 (0.0284)	0.0700 (0.0362)	-0.0732 (0.0447)
q = 11	0.0434 (0.0307)	0.103** (0.0400)	-0.0648 (0.0456)
q = 12	0.0517 (0.0281)	0.0874* (0.0353)	-0.0215 (0.0459)
Running Average	-0.0136** (0.00464)	-0.0341*** (0.00436)	-0.0490*** (0.00404)
Observations	7,576,829	7,285,368	7,122,706
R-squared	0.253	0.253	0.251
Business Fixed Effects	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes

Discussion

Empirical investigation of fake review alerts yielded several interesting insights. First, we note the long-term nature of the solicitation of fake reviews, with observable differences beginning 15 quarters prior to alerts being issued. The 336,084 reviews posted for the alerted businesses during this period collectively received 510,691 “useful” votes, 111,986 “funny” votes and 147,788 “cool” votes from other Yelp users. This suggests a tremendous scale for the impact of the fake reviews, as at least half a million platform consumers read and engaged with reviews for businesses that were manipulating the platform. To the extent that exposure to fraudulent reviews may impact trust in the platform, addressing this through the alert system can have a considerable impact on users.

Second, we note that alerts have an immediate and substantial effect on ratings. Following years of measurable increases in ratings, ratings immediately return to their original levels after alerts are issued, and we also observe an immediate decline in review volume. This pattern is consistent with that observed in He, Hollenbeck, and Proserpio (2021), where, because firms soliciting fake reviews were of low quality, both ratings and the review volume fell sharply after solicitation of fake reviews ended. In our context, this suggests that the fake reviews did not enable the firm to build a sustainable advantage in attracting reviews through engaging in manipulation of the platform.

Lastly, we note that while businesses that received Compensated Activity alerts had significant ratings inflation that began earlier than those that received Suspicious Review Activity alerts, the maximal extent of their effects on ratings were similar. Moreover, platform consumers did not appear to discern between the alert types when it came to their post-alert evaluations. Thus, from the consumer point of view, this may suggest that platforms need only

consider one type of alert; however, platforms may prefer different wording for other reasons, such as legal distinctions.

This study is not without limitations. As noted, the most prominent among these is that alerts were not assigned randomly, and instead businesses effectively selected into potential treatment by engaging in fraudulent review manipulation. While this limits the ability of our findings to generalize to other contexts, the consistency in the characteristics of the firms in the present study to those in prior work suggests that the alerted businesses are reasonably typical of businesses that engage in review fraud. A more significant issue is the imprecise nature of the criteria under which Yelp chose to issue alerts to these businesses. Because this is not known, it is possible that the effects observed here are an artifact of this potential selection.

While Study 1 showed how business' ratings changed during and after a fake review alert, it provided little insight into other ways in which consumers might respond to the alert. For instance, do consumers change their reading behavior in the presence of an alert? Do they think the removal of the fake reviews makes the remaining reviews more authentic? To what extent does the presence of an alert affect their opinion of the platform? In the next study, we conduct an experiment to examine the effect of the alert on consumers' impressions of the business and the platform.

Study 2: Additional Consequences of Alerts

Procedure

We recruited 300 participants from CloudResearch Connect and receive 303 completed responses ($M_{\text{age}} = 41.32$; 144 male, 151 female, 6 nonbinary/third gender, 2 prefer not to disclose) in exchange for payment. Participants were randomly assigned to one of two between-subjects conditions (fake review alert: absent, present). Participants were told, "This study will

ask for your evaluations of a business and review platform. You will see some information about a café and then will be asked some questions.” In the control condition, participants were then shown the presumed Yelp page for a business (West End Café) (see Appendix WA). In the fake review alert treatment, participants first encountered an alert with the text slightly adapted from Yelp’s Compensated Review Activity wording (see Table 2) before viewing the same Yelp business page as in the control condition. They then responded to a series of measures.

Table 2: Language of Fake Review Alert in Studies

Consumer Alert: Compensated Review Activity
We caught someone offering up payment in the form of cash, discounts, gift certificates or other incentives in exchange for someone to write, change, prevent or remove reviews for this business. We wanted you to know because these actions not only hurt consumers, but also honest businesses who play by the rules.

Given that a fake review alert increases suspicion, we first assessed a manipulation check about the *expected trustworthiness* of the reviews for the business by measuring agreement with the item “The available reviews for this product will be trustworthy” (1 – 5: strongly disagree – strongly agree). Next, to measure the extent to which the alert might affect reading behavior, participants could choose to read some or all of the ten reviews that were available for the business (i.e., *reviews read*). If they indicated yes, they were shown a single review drawn from a representative distribution of 10 reviews (i.e., six 5-star, one 4-star, two 2-star, one 1-star, which yielded a 3.9 out of 5.0 average; see Appendix WA for the specific reviews). This process occurred until all 10 reviews were exhausted or until they indicated “no”.

Once participants indicated “no”, they proceeded to our focal measures (See Table 3) of *business intentions* (3 items; $\alpha = .90$) and *perceived platform integrity* (3 items; $\alpha = .88$). See table 3 for individual items. Business intentions were measured to assess whether the findings for

average business ratings in Study 1 were replicated with the downstream measure of intentions. *Perceived platform integrity* assessed whether the alert increased or decreased consumers' views of the platform. These measures were followed by two others: *review quality motivations* (3 items; $\alpha = .73$); and *perceived ratings bias* (3 items; $\alpha = .96$), followed by age and gender. Review quality motivations was an exploratory measure of whether seeing a fake review alert makes consumers likely to change how they might write reviews in the future. Finally, whereas the manipulation check of review trustworthiness measured expectations of the reviews, perceived ratings bias measured consumers' impressions of the degree to which they thought the reviews were authentic *after* reading the reviews. As in study 1, we organize the results around the relevant questions.

Table 3: Items Used for Each Measure

Business Intentions (1-7: extremely unlikely – extremely likely)	<ul style="list-style-type: none"> • “If you were searching for hotels, how likely would you be to select this option?” • “If you had a 25% off coupon, how likely would you be to try this option?” • “If you were recommending nearby options to a friend, how likely would you be to recommend this option?”
Review Platform Integrity (1-7: strongly disagree – strongly agree)	<ul style="list-style-type: none"> • “Yelp is doing enough to protect their consumers from fake reviews” • “Yelp deserves praise for their efforts to protect their consumers from fraudulent information” • “Yelp’s transparency with their consumers is admirable”
Review Quality Motivations (1-7: strongly disagree – strongly agree)	<ul style="list-style-type: none"> • When I write reviews, I should focus more on making sure that they are high-quality • I should write reviews often to help out those who seek information • I need to do a better job when I write reviews
Perceived Ratings Bias (1-7: not at all – extremely; reverse-coded)	<p>To what extent do you think the available reviews for West End Café were:</p> <ul style="list-style-type: none"> • Accurate • Believable • Authentic

Results

Manipulation check

A one-way ANOVA of the alert condition on expected review trustworthiness yielded a significant effect ($F(1,301) = 47.819, p < .001, \eta^2_p = .137$). Relative to the control ($M = 3.65$), exposure to a fake review alert decreased expected trustworthiness of reviews ($M = 2.94$). Thus, the alert reduced trustworthiness of the reviews on the business' subpage.

Question 1: Does a fake review alert decrease business intentions?

A one-way ANOVA of the alert condition on business intentions yielded a significant effect ($F(1,301) = 16.145, p < .001, \eta^2_p = .051$). Relative to the control ($M = 4.84$), exposure to a fake review alert decreased business intentions ($M = 4.26$). This result is consistent with Study 1's finding that average business ratings decreased during the period of the alert.

Question 2: Does an alert affect perceptions of the platform?

A one-way ANOVA of the alert condition on perceived platform integrity yielded a significant effect ($F(1,301) = 29.504, p < .001, \eta^2_p = .089$). Relative to the control ($M = 4.48$), exposure to a fake review alert increased perceived platform integrity ($M = 5.17$). Thus, participants thought more favorably of the platform in the presence vs. absence of a fake review alert.

Question 3: Does an alert change review reading behaviors?

A binary logistic regression demonstrated that the presence of an alert significantly decreased the likelihood of reading reviews ($P_{\text{control}} = .78, P_{\text{present}} = .65, \text{Wald}(1) = 6.582, p = .01; \text{OR} = .513$). While 78% of respondents read one or more reviews in the control condition, only 65% did so in the alert condition. This suggests that participants were engaged in defensive

processing (Darke and Ritchie 2007) by avoiding reading reviews from an untrustworthy business.

Among those who chose to read the reviews, however, the alert did not affect the number of reviews read ($M = 4.53$, $F(1,215) = 675$, $p = .412$, $\eta^2_p = .003$) nor did it affect the average valence of reviews read ($M = 3.89$, $F(1,215) = .048$, $p = .827$, $\eta^2_p = 0$). This suggests that while the alert reduced the likelihood of reading reviews, it did not make participants more likely to read certain types of reviews (or to stop reading after exposure to different types of reviews).

Question 4: Does a fake review alert increase perceived ratings bias?

A one-way ANOVA of the alert condition on perceived ratings bias yielded a significant effect ($F(1,301) = 15.653$, $p < .001$, $\eta^2_p = .049$). Relative to the control ($M = 2.97$), exposure to a fake review alert increased perceived bias ($M = 3.56$). Thus, another platform consequence of the alert was that it increased suspicion of reviews of the fraud-producing business. However, the mean perceived ratings bias did not reach the scale midpoint, suggesting that consumers did not see the bias as egregious.

To explore the perceived ratings bias further, we wanted to understand if it differentially impacted those who chose to read reviews versus those who did not. A 2 (alert: control, present) x 2 (read reviews: no, yes) x revealed significant main effects of the alert ($F(1,299) = 7.25$, $p < .007$, $\eta^2_p = .024$) and the decision to read reviews ($F(1,299) = 17.69$, $p < .001$, $\eta^2_p = .056$), with no significant interaction ($F(1,299) = .57$, $p = .45$, $\eta^2_p = .002$). Interestingly, participants who chose to read reviews perceived the reviews to be less biased than those who avoided the reviews ($M_{\text{did not read}} = 3.80$, $M_{\text{read}} = 3.05$). Thus, reading the reviews made participants see the reviews as less biased and thus more credible. To assess whether this was due to self-selection (i.e., those who read the reviews already trusted the reviews more), we ran the same 2-way ANOVA on the

manipulation check measure (i.e., perceived review trustworthiness), which was taken prior to the opportunity to read the reviews, and found no effect of reading the reviews ($F(1,299) = 2.57$, $p = .11$; $\eta^2_p = .009$) nor the interaction ($F(1,299) = .01$, $p = .915$; $\eta^2_p = 0$). Thus, the effect of reading the reviews on perceived ratings bias was not due to self-selection, but to the actual reviews changing people's perceptions of bias. This is good news for the platform as readers perceived less bias in the reviews.

Question 5: Does the alert change consumers' motivation to write reviews?

A one-way ANOVA of the alert condition on review quality motivations yielded a nonsignificant effect ($M = 4.64$, $F(1,301) = .681$, $p = .410$, $\eta^2_p = .002$). Thus, the alert did not spur consumers to consciously think of improving the quality of their own future reviews.

Discussion

This study extended the consequences of fake review alerts to consumers' reading behaviors as well as perceptions of the platform and reviews. A fake review alert lowered perceptions of the business but increased perceptions of the platform's integrity, as participants viewed the platform as being transparent and honest. In terms of reading behavior, fewer participants chose to read the business's reviews after encountering the alert. However, they read the same number and valence of reviews as did participants in the control condition. Moreover, those who did read the reviews perceived the reviews as less biased than those who did not read the reviews in both the control and alert condition. This is encouraging for the platform, as it suggests that participants perceived the reviews to be more authentic after reading the actual reviews. These results are consistent with those of Beck and colleagues (2023), showing consumers perceive the platform as having more integrity when they are told that the platform communicates review credibility. However, our context provides a more naturalistic exposure to

a fake review alert (rather than reviewer authentication or credibility badges) and subsequently demonstrates additional behavioral consequences of the alert through a decreased likelihood of reading reviews. Although exposure to the alert did not affect participants' motivation to improve the quality of their own reviews, this could be due to our sample. The proportion of platform users who write reviews is quite small; industry estimates suggest only 5 – 10% of users write reviews (Weise 2017), while most individuals consume them. This means that very few participants in Study 2 were likely to be review writers. We speculated that this might explain why the fake review alert did not affect potential review writing behavior. There may be a difference between those who write reviews and those who read them.

A potential framework for understanding this distinction is provided by Flynn, Goldsmith, and Eastman (1996), who delineate opinion leaders, who provide information to others, from opinion seekers, who search for information from others. In general, we expected that those who write reviews are likely to be opinion leaders (in a given domain), while those who read reviews are likely to be opinion seekers. Opinion leaders are likely to be more highly involved or knowledgeable in the product category than are opinion seekers (Flynn, Goldsmith, and Eastman 1996). In addition, their underlying motivations might also differ. Although reviewers write reviews for several different reasons (Babic Rosario, de Valck, and Sotgiu (2020), the most common motivation is to help others make good choices (Hennig-Therau et al. 2004), while they may also be driven by community-building and status (Beck, Wuyts, and Jap 2023). To such reviewers, a fake review alert indicates that readers have been getting poor information. In other words, a fake review alert threatens the veracity of the platform, which is important to such reviewers. As a result, opinion leaders may want to offset fake reviews by improving the quality of their own reviews. In contrast, opinion seekers want to get accurate

information rather than disseminating the information. Therefore, they may not end up writing reviews.

Since Study 2 examined the behavior of opinion seekers, we designed the next study to assess whether the behavior of opinion leaders (i.e., review writers) is affected by a fake review alert. Specifically, the study involved participants writing reviews following an actual experience. We then employed text analysis of the written reviews to determine their quality.

Study 3: Effect of Fake Review Alerts on Opinion Leaders

Procedure

We recruited 600 participants from Amazon mTurk. Ultimately, 599 participants ($M_{\text{age}} = 44.88$; 263 male, 327 female, 8 non-binary or other, 3 prefer not to answer) completed the survey in exchange for payment. Participants were randomly assigned to one of two between-subjects conditions (alert prime: absent, present). Participants were told, “This study consists of two parts. In the first, we’ll ask you to read a short article, then we’ll check for comprehension with a few questions about the content. In the second, we’ll ask you to participate in a breathing exercise and write a review for the instructor of the video.” Part 1 manipulated the alert. In the alert absent condition, participants read a brief article titled “Collective Craze: Labubu’s Plush Monsters Go Mainstream”, which featured an image of Labubu plushies. In the alert present condition, participants read a brief article titled “Fake Reviews Craze: Calls for Honest Feedback” which featured an image of a fake review alert. For the full articles, see Appendix WC. As a cover, participants were asked to recall the article topic and report its “interestingness”.

Next, participants moved on to Part 2, where all participants watched a two-minute breathing exercise video. After the video, they were asked to rate the experience (“How would you rate the experience?”; [1 – 5 stars]) and to write a review (“Next, please write a review for

Scott Schwenk's Breathwork Tutorial. Tell us about your experience and what you think others might like to know about this video. (*Reviews need to be at least 85 characters; copy & paste functionality is not available*”). We then measured a manipulation check and two exploratory measures (see Web Appendix WC), and demographics.

Quality Measures

For each review, we extracted a set of features to capture the quality of the reviews, adapted from prior work (Crossley 2020; Kiefer 2019). These review features captured aspects of the length of the text, lexical diversity, and overall complexity. Longer text is seen as a marker of higher quality (Blumenstock 2008; Fleckenstein et al. 2020), and we included the typical measure of word counts, along with character and sentence counts. Lexical diversity captures the use of varied words within the text and is also associated with quality (Yang, Yap, and Ali 2023). We measured this through the number of unique words used in the review. Finally, the complexity of a text reflects the density of information it provides (Pilgrim et al. 2024; Aceves and Evans 2024) and was measured through Shannon entropy, which captures the extent to which the text is “predictable” and repetitive. We employed multiple dependent variables for robustness. In Table 4, we present two reviews from the experiment and their scores on each of the five metrics.

Table 4: Example Reviews and Quality Metrics

Condition	Review Text	(1) Words	(2) Characters	(3) Sentences	(4) Unique Words	(5) Entropy
Control Prime	I like this tutorial very much because it helps me to relax my mind and also it helps me to refresh my body. I feel this tutorial is very useful for everyone	32	126	2	23	.947

Alert Prime	I recently watched Scott Schwenk's Breathwork Tutorial, and I found it both insightful and accessible. Scott has a calm, reassuring presence that makes the techniques feel approachable, even if you're new to breathwork	33	187	2	29	1.451
-------------	--	----	-----	---	----	-------

Results

The full results are reported in Table 5. Regression analysis with the alert prime as an independent variable showed that those in the alert prime (vs. control) condition wrote reviews that were longer [words: $b = 3.15$, $t(597) = 2.07$, $p = .039$; characters: $b = 15.19$, $t(597) = 1.99$, $p = .047$; sentences: $b = .21$, $t(597) = 2.32$, $p = .021$], more lexically diverse [unique words: $b = 2.06$, $t(597) = 2.16$, $p = .031$] and marginally more complex [entropy: $b = .10$, $t(597) = 1.89$, $p = .059$]. Taken together, these results suggest that exposure to the alert prime led participants to produce reviews that were more effortful and of higher quality.

Table 5: The Effect of an Alert Prime on Review Quality

Variables	(1) Words	(2) Characters	(3) Sentences	(4) Unique Words	(5) Entropy
Alert	3.148* (1.523)	15.19* (7.631)	.213* (.0920)	2.057* (.953)	.102^ (.0538)
Constant	37.89*** (1.110)	189.20*** (5.560)	2.520*** (.0671)	28.43*** (.694)	1.275*** (.0392)
Observations	599	599	599	599	599
R-squared	.007	.007	.009	.008	.006

Discussion

Analyzing the quality of reviews written by participants in the lab, we found that exposure to a fake review alert prime led participants to write higher quality reviews across a variety of metrics commonly used to capture the quality of text. The reviews were longer, with more words, characters, and sentences; more lexically diverse; and marginally more complex

than reviews written in the absence of a fake review alert prime. This suggests that improvement in the quality of reviews by opinion leaders is another benefit to platforms of posting fake review alerts, though the study does not speak to the underlying mechanism.

In the final study, we again employ secondary data from the Yelp platform to assess whether these effects can be replicate with “true” opinion leaders (i.e., online review-writers), and without the potential demand effect of reviews being written immediately after alert exposure.

Study 4: Effect of Fake Review Alerts on Online Opinion Leaders

Methodology

To examine online review quality after exposure to a fake review alert, we returned to the 4,963 businesses that received Yelp alerts in Study 1. We focused on the reviews posted to these businesses’ Yelp pages during the 90-day period, which were contributed by 26,999 unique Yelp users. These users would have seen the fake review alert when posting their review and, thus, were aware of the steps being taken to improve the quality of reviews posted on the platform. We refer to this group as the *treated* users. We then collected all the available reviews posted by these users, including before and after the alert period, totaling 1,551,396 reviews.

To construct an appropriate counterfactual, it was necessary to identify a group of platform users who were (1) representative of typical platform users, (2) active on the platform during the same period as the alert, and (3) less likely to have been exposed to a fake review alert compared to the treated users. We again employed the Yelp Open data set as a counterfactual against alerted businesses, as the users it contains should represent platform users of sufficient quality. In the Yelp Open data set, there were 1.8 million users who posted a review of a business that had not received an alert during an alert period. In the interests of tractability and operating

within application programming interface (API) rate limitations, we drew a simple random sample of users from this set, following from prior work to limit the sample to 50,000 users, approximately double the size of the treated group (Villanova and Matherly 2024). We collected all available reviews from these users, totaling 2,233,337 reviews, and matched them into cohorts with those who had reviewed a business during an alert period, yielding the *untreated* set. Combined, this yielded 3,054,526 total reviews from 532,151 businesses after removing singletons. Posting a review requires visiting the Yelp page of the business, so we can observe when treated users would be exposed to an alert; however, we cannot directly observe what business pages users visit without posting a review. Therefore, we cannot be certain that users in the untreated set did not see a review alert. However, the probability is necessarily less-than-or-equal-to those in the treated set who would see the alert when posting their review of the alerted business. Therefore, our estimate of the effect of having seen the alert on review quality derived from observed differences between the treated and untreated groups is likely to underestimate the magnitude of the true effect.

Model

For each review, we recorded the rating assigned to the business, the number of useful votes the review received (normalized by age of the review), and the body text of the review. The number of useful votes received is a parallel measure of quality (Ceylan, Diehl, and Proserpio 2024; Zhang, Zhou and Andrews 2026), and we normalized the raw value to account for the reviews' age. We additionally calculated the same measures of text quality as in Study 3. To model the effect of having seen an alert on these outcomes, we estimated the following model:

$$y_{ijt} = \beta X_{it} + \lambda_i + \gamma_j + \tau_t + \epsilon_{ijt},$$

- Where y_{ijt} was the outcome,

- X_{it} was a treatment indicator, such that $X_{it} = 1$ if reviewer i had seen an alert while reviewing a business before time t , and 0 otherwise,
- λ_i , γ_j , and τ_t were fixed effects for reviewer, business and week, respectively, and
- ϵ_{ijt} was the disturbance term

The model was estimated with robust standard errors clustered on the business.

Results

The full results are presented in Table 5. Replicating Study 3, following exposure to a fake review alert, reviewers wrote significantly longer reviews (word count: $b = 1.01$, $t(532,151) = 3.85$, $p < .001$; character count: $b = 5.02$, $t(532,151) = 4.07$, $p < .001$; sentence count: $b = .08$, $t(532,151) = 4.70$, $p = .008$), with more unique words ($b = .805$, $t(532,151) = 7.27$, $p < .001$) and greater entropy ($b = .04$, $t(532,151) = 4.47$, $p < .001$). Thus, the quality of the reviews was higher when reviewers had been exposed to a fake review alert than when they had not been exposed.

The “usefulness” votes told a similar story. Reviews written after exposure to the alert received more usefulness votes ($b = .224$, $t(532,151) = 16.53$, $p < .001$) than those without alert exposure. Overall, these findings suggest that treated users produced higher quality reviews for subsequently reviewed businesses compared to those who wrote reviews for businesses that were not issued an alert during the same period as the affected businesses.

Table 6: The Effect of Alert Exposure on Subsequent Review Quality

Variables	(1) Words	(2) Characters	(3) Sentences	(4) Unique words	(5) Entropy	(6) Usefulness Votes
Alert	1.009*** (0.262)	5.018*** (1.234)	0.0408** (0.0153)	0.805*** (0.111)	0.0377*** (0.00844)	0.224*** (0.0136)
Constant	140.9*** (0.0469)	665.5*** (0.221)	9.075*** (0.00274)	80.59*** (0.0198)	4.512*** (0.00151)	0.654*** (0.00243)

Observations	3,054,526	3,054,526	3,054,526	3,054,526	3,054,526	3,054,526
R-squared	0.557	0.561	0.531	0.581	0.561	0.426

Discussion

Replicating and extending the findings of Study 3, the results of Study 4 provide real-world evidence for fake review alerts’ potential to increase review quality for businesses that are subsequently reviewed on the platform. Thus, businesses who do not engage in review fraud appear to benefit from Yelp’s fake review alerts, as review quality for these businesses improves after reviewers’ exposure to an alert from a different business. This suggests that the long-term health of the platform may improve as the quality of its content improves.

General Discussion

Across four studies, spanning the field and the lab, we demonstrated that fake review alerts decrease fake review prevalence and penalize businesses that engage in fraud. Moreover, platforms see multiple benefits, from increased perceptions of platform integrity to improved review quality from those exposed to an alert. Notably, after writing a review for an alerted business, reviewers subsequently contribute reviews that are longer, more effortful, and rated more useful, improving the review ecosystem for the platform.

Our findings are robust across contexts. The secondary data spans thousands of businesses across dozens of categories while the lab experiments utilize commonly reviewed domains to investigate the effects in-depth. Moreover, this work considers the various stakeholders of interest. From the business’s decision to post fake reviews to the platform’s issuance of an alert, we consider how both opinion seekers (review readers) and opinion leaders (review writers) respond to the presence of a fake review alert. Since many of the findings have

implications for various stakeholders, Table 6 summarizes each finding under the key stakeholder. We discuss the implications of these findings below.

Table 6: The Effects of a Fake Review Alert on the Platform, Businesses, and Consumers

Stakeholder	Key Findings
Platform	<ul style="list-style-type: none"> • Decreases review fraud [Study 1] • Able to inform consumers of platform strategy [Study 1, 2, 4] • Incur increased integrity perceptions [Study 2] • Does not alter consumers' experience for subsequent businesses [Study 3] • Increases review quality on platform [Study 3, 4]
Alerted Business	<ul style="list-style-type: none"> • Incur a ratings penalty while the alert is active [Study 1] • Decreases consumer intentions [Study 2]
Unalerted Business	<ul style="list-style-type: none"> • Receives improved review quality [Study 3, 4] • No effect on business evaluation [Study 3, 4]
Opinion Seekers	<ul style="list-style-type: none"> • Reduces likelihood of reading reviews [Study 2] • No effect on which reviews are read (no change in the stopping rule or average valence read) [Study 2] • Decreases intentions towards an alerted business [Study 2]
Opinion Leaders	<ul style="list-style-type: none"> • Increases review quality for subsequent businesses [Study 3, 4] • Leaders' reviews receive more positive reinforcement (i.e., usefulness votes) from seekers [Study 4]

Implications for the Platform

The paper provides several implications for platforms that employ fake review alerts. Overall, fake review alerts appear to be effective in achieving the objectives of warning consumers as well as reducing fraud behavior of the target business. Perhaps the biggest contribution of the paper is in demonstrating that exposure to a fake review alert improved the quality of reviews written by opinion leaders. In Studies 3 and 4, those who saw the fake review alert (vs. not) wrote reviews that were longer, with more words, characters, and sentences; more lexically diverse; and more complex. Note that this is not a result we had anticipated, and it was the result of the empirics first approach.

Two other findings increase our understanding of the effects of fake review alerts on consumers and implications for platforms. First, the wording of the alert (suspicious versus compensated) did not differentially affect business ratings during and after the alert period, nor did the wording materially affect other business or platform judgments (Study 2). To ensure that we could generalize this effect, we ran a controlled study, reported in Web Appendix WB, which included both compensated and suspicious alerts. We found similar results across both alert types giving us confidence that the effect is robust to various wordings of the alert. This is also consistent with the response to alerts observed in Study 1. We believe that the wording does not especially matter because the warning itself raises sufficient suspicion such that different plausible reasons for the platform's detection system do not matter to consumers. Of course, the platform has myriad options of alert wordings to use. So, it is possible that other wordings may have unique effects on consumers.

The second finding is that the alert only partly affects consumers' review reading behavior. In Study 2, although fewer participants chose to read the business's reviews after encountering the alert, those who did read the reviews read the same number and valence of reviews as did participants in the control condition. While the study did not allow consumers to choose specific review valences to read, it did not change the stopping rules employed by readers, suggesting that this may not have made a difference. Moreover, readers perceived the reviews as less biased than those who did not read the reviews in both the control and alert condition. For the platform, this suggests that participants perceived the reviews to be more authentic after reading the actual reviews. However, we cannot claim this as an empirical generalization as it was only assessed in one study. Future research should examine opinion seekers' reading behavior in the presence of a fake review alert.

Implications for Businesses and Public Policy

Our research provides some insight into the effects of fake review alerts for businesses. For the focal business, the alert is clearly not beneficial. A fake review alert hurts business evaluations (ratings) and behavioral intentions for the alerted business, but not unalerted businesses. Given the importance of ratings to sales and product choice (Chevalier and Mayzlin 2006; Watson, Ghosh, and Trusov 2018), businesses who solicit fake reviews may be underestimating the consequences of getting caught and failing to consider the long-term effects this may have on their business. For honest businesses, fake review alerts are beneficial. They penalize dishonest competitors while simultaneously increasing the review quality by those who encounter an alert and write a review for their business. While this does not necessarily mean that the rating of the experience improves, it means that the reviews became more useful and persuasive for opinion seekers, pushing them towards high quality businesses. Thus, businesses which engage in honest behavior should request that more platforms utilize a fake review alert system.

From a policy perspective, this work gives organizations such as the FTC a building block to develop regulatory policy. Our work suggests that fake review alerts could and should be encouraged on the policy front. In doing so, policy bodies might consider whether developing standardized fake review alert wording across platforms would be beneficial to consumers, and whether a shared database of fraudsters would help platforms to ensure the review integrity on their platform when an offender was flagged on another platform.

These claims should be qualified by the understanding that Yelp is a review platform, where the lion's share of the revenue comes from advertising rather than on-site purchases. That is, Yelp's primary role in the marketplace is that of an information provider rather than a retailer.

While the benefits of fake review alerts in this context are plentiful, it is not obvious what the implications would be for a retailer platform (e.g., Amazon). He, Hollenbeck, and Proserpio (2022) identified large review fraud rings on Amazon, observing that the sales of promoted products crater without the fake reviews, but the authors stop short of addressing interventions for long-term mitigation. We would expect that fake review alerts would yield similar sales-stopping effects for the focal business. More interesting, the full marketplace dynamics are less clear. Would a fake review alert push consumers to purchase from another (honest) business or opt for a different retailer entirely? We leave this question to future work.

Implications for Consumers

The implications for consumers are fairly straightforward. Fake review alerts serve as a marker of transparency from the platform. This leads opinion seekers to rely more on the alert than reviews in their judgments of alerted businesses. Alerts allow opinion seekers to avoid businesses that lack trustworthiness, but this does not necessitate that other businesses are more trustworthy. Indeed, there exist likely hotbeds of fraudulent review activity where many businesses with similar profiles may engage in review fraud (Mayzlin, Dover, and Chevalier 2014). Moreover, the tradeoff consumers may make between ethical practices and expected quality may be context-specific, but understanding when consumers may still prefer an alerted business over an unalerted business is an interesting question for future research. Will consumers choose an objectively inferior, presumed-honest business over a superior, but immoral business? Work by Kirmani et al. (2017) shows that consumers generally weigh competence more than morality when choosing a service provider; however, if the immorality is directly harmful to the consumer, consumers will avoid an immoral provider. Since the business that posts fake reviews

is harming consumers by giving them false information, it is likely that morality might prevail. However, future research can examine the specific trade-off in this context.

Another contribution of the paper is to highlight the motivational differences between opinion leaders and opinion seekers, which lead to different outcomes. The distinction between opinion leaders and opinion seekers is similar to the distinction between posters and lurkers (Schlosser 2005). By our definition, posters, who like to share their experiences online, are opinion leaders. Lurkers, who read others' posts, are opinion seekers. Schlosser (2005) argues that posters, but not lurkers, are likely to have self-presentation motives. This leads posters to adjust their opinions in the presence of others' negative reviews. Although Schlosser (2005) focused on product ratings, our research suggests that the quality of reviews may also be affected by the type of information encountered by the opinion leader. We speculate that these self-presentation motives may affect how opinion leaders (i.e., posters) write reviews after being exposed to a fake review for several reasons. Specifically, a fake review alert may make salient that the platform is monitoring the reviews, so opinion leaders become more careful in writing the reviews, as they do not want their reviews to be classified as fake, thereby damaging their reputation. More likely, however, are other motives that drive opinion leaders. Opinion leaders may be trying to help consumers by providing better information to offset the misinformation provided by bad players (i.e., fake review writers). Since helping others and influencing others are among the most common motivations for writing reviews (Berger 2014; Hennig Thureau et al 2004), opinion leaders realize that fake reviews can mislead consumers into making poor choices. By improving the quality of their reviews, they make the reviews more useful. Second, opinion leaders may be trying to help the platform, which has been attacked by bad players. To the extent that opinion leaders' identity is aligned with the platform, they may want to ensure the

platform's success. While there is little research on whether online reviewers identify with a platform, helping the company is one of the motivations for writing reviews (Hennig-Thurau et al 2004). Thus, reviewers may have developed a connection to the platform, enforcing its implicit and explicit rules. Extrinsic rewards, such as badges and other forms of recognition may also help develop connection to the platform. Future research can investigate these different mechanisms underlying review writing behavior.

Notably, however, opinion seekers and leaders are not mutually exclusive. Opinion leadership likely emerges from an opinion seeker finding value in the opinions of others and choosing to also produce useful content for others. Thus, the seeker/leader distinction is context-dependent. While this work finds evidence that consumers become more selective with opinion leadership after alert exposure, it would be interesting to understand further how it affects more vs. less prominent opinion leaders. Or, if there are cross-platform implications where someone may become more or less vocal on one platform given the fake review alert usage on another platform.

Limitations

As with any research, limitations exist. Given the structure of our field data, we only observed individuals who wrote reviews, not those who encountered the alert and chose to not write a review (Studies 1, 4). As such, this presents a necessarily limited understanding of consumer response because we do not observe how those who encounter the alert, but do not review or respond in the field. Our experiments (Studies 2, 3) attempted to address this concern by treating populations with various information environments, but this too, may lack the full external validity of a natural information environment. Without that environmental context in the data, there is some potential for unexplained variance in the effects of alerts.

Furthermore, while the fake review alerts investigated in this paper are increasingly commonplace, new alerts are emerging which capture unique occurrences. For example, Yelp has new contextual fake review alerts like a “Racist Behavior Alert” and “Unusual Activity Alert”. Rather than indicating review fraud by the business, these alerts state that actions by the business have resulted in likely review fraud by the public. Thus, unique from the alerts studied in this paper, these other alerts often also bring additional societal values into the context of evaluating businesses. This may yield quite different results from the alerts studied in this paper as a function of consumers’ own values. Future work may also consider how to best mitigate fake reviews from these consumers by attenuating their motivation to post beforehand.

Finally, while reviews are a mainstay of online word-of-mouth, influencer marketing is becoming more central business marketing strategy (Leung, Gu, and Palmatier 2022). This is because influencers are a relatively low-cost way to raise awareness and increase sales (Leung et al. 2022). As such, the relative power of traditional product reviews may be losing ground to influencers. However, regulation in the influencer marketing space is evolving, and the low cost of entry makes recommendation fraud nearly as easy as fake reviews. As such, this suggests that social media platforms may find interest in our work as they consider disclosure and correction policies for fraudulent content on their platforms. As the use of synthetic influencers rise (DualMedia 2025), social media platforms must navigate how to penalize the creators and businesses when disclosure is not apparent, and our work may shed light on processes for platforms to navigate this coming challenge.

References

- Aceves, Pedro, and James A. Evans (2024), "Human Languages with Greater Information Density have Higher Communication Speed but Lower Conversation Breadth," *Nature Human Behaviour* 8, (4): 644-656.
- 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*.
- Autor, David H. (2003), "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing," *Journal of Labor Economics*, 21 (1), 1–42.
- Beck, Ben B., Stefan Wuyts, and Sandy Jap. (2023), "Guardians of Trust: How Review Platforms Can Fight Fakery and Build Consumer Trust," *Journal of Marketing Research*, 61(4), 682-699.
- Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs (2001), "Bad is Stronger than Good," *Review of General Psychology*, 5 (4), 323-370.
- Blumenstock, Joshua E. (2008), "Size Matters: Word Count as a Measure of Quality on Wikipedia," In *Proceedings of the 17th international conference on World Wide Web*, 1095-1096.
- Broun, Kenneth S., Paul C. Giannelli and Robert P. Mosteller (2014), *Evidence*, 8th Edition. Eagan, MN: West Academic Publishing.
- 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.

- Campbell, Margaret C., Gina S. Mohr, and Peeter WJ 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.
- CapitalOne (2025), "Fake Review Statistics," (accessed January 2, 2026), [available at: <https://capitaloneshopping.com/research/fake-review-statistics/>].
- Ceylan, Gizem, Kristin Diehl, and Davide Proserpio (2024), "Words Meet Photos: When and Why Photos Increase Review Helpfulness," *Journal of Marketing Research*, 61 (1), 5-26.
- Chen, Yubo, and Jinhong Xie (2008), "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science*, 54 (3), 477-491.
- 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.
- Correia, Sergio (2015), "Singletons, cluster-robust standard errors and fixed effects: A bad mix," *Technical Note, Duke University*, November 2015. Accessed January 29, 2026 from <https://scorreia.com/research/singletons.pdf>.
- Crossley, Scott A (2020), "Linguistic Features in Writing Quality and Development: An Overview," *Journal of Writing Research* 11 (3), 415-443.
- 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.
- Darke, Peter R., Laurence Ashworth, and Robin JB Ritchie (2008), "Damage from Corrective Advertising: Causes and Cures," *Journal of Marketing*, 72 (6), 81-97.

- Dean, Kathryn (2024), "FTC rules on fake reviews aren't enough: Hold sites accountable," (accessed September 15, 2024), [available at: <https://thehill.com/opinion/technology/4861444-fake-reviews-online-review-fraud/>].
- De Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817-833.
- DualMedia (2025), "The Rise of Synthetic Influencers: What It Means for Digital Marketing," (accessed January 13, 2026), [available at: <https://www.dualmedia.com/rise-synthetic-influencers/>].
- Eisend, Martin, Eva A. van Reijmersdal, Sophie C. Boerman, and Farid Tarrahi (2020), "A Meta-Analysis of the Effects of Disclosing Sponsored Content," *Journal of Advertising*, 49 (3), 344-366.
- 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.
- Fleckenstein, Johanna, Jennifer Meyer, Thorben Jansen, Stefan Keller, and Olaf Köller (2020), "Is a Long Essay Always a Good Essay? The Effect of Text Length on Writing Assessment," *Frontiers in Psychology*, 11, 562462.
- Floyd, Kristopher, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling (2014), "How Online Product Reviews Affect Retail Sales: A Meta-Analysis," *Journal of Retailing*, 90 (2), 217-232.

- Flynn, Leisa Reinecke, Ronald E. Goldsmith, and Jacqueline K. Eastman (1996), "Opinion Leaders and Opinion Seekers: Two New Measurement Scales," *Journal of the Academy of Marketing Science*, 24 (2), 137-147.
- Friestad, Marian, and Peter Wright (1994), "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts," *Journal of Consumer Research*, 21 (1), 1-31.
- FTC (2024), "16 CFR Part 465: Trade Regulation Rule on the Use of Consumer Reviews and Testimonials (Final Rule)," (accessed September 15, 2024), [available at: <https://www.ftc.gov/legal-library/browse/federal-register-notices/16-cfr-part-465-trade-regulation-rule-use-consumer-reviews-testimonials-final-rule>].
- Gilbert, Daniel T., Douglas S. Krull, and Brett W. Pelham (1988), "Of Thoughts Unspoken: Social Inference and the Self-Regulation of Behavior," *Journal of Personality and Social Psychology*, 55 (5), 685.
- Ham, Sung H., Ingrid Koch, Noah Lim, and Jiabin Wu (2021), "Conflict of Interest in Third-party Reviews: An Experimental Study," *Management Science*, 67 (12), 7535-7559.
- Hayes, Andrew F. (2017), "*Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach*," Guilford Publications.
- He, Sherry, Brett Hollenbeck, and Davide Proserpio (2022), "The Market for Fake Reviews," *Marketing Science*.
- Huffman, Mark (2024), "Here's how much fake reviews could be costing you," (accessed December 15, 2024), [available at: <https://www.consumeraffairs.com/news/heres-how-much-fake-reviews-could-be-costing-you-120924.html>].

- Jones, Edward E., and Keith E. Davis (1965), "From Acts to Dispositions the Attribution Process in Person Perception," In *Advances in Experimental Social Psychology*, 2, 219-266. Academic Press.
- Kelley, Harold H., and John L. Michela (1980), "Attribution Theory and Research," *Annual Review of Psychology*, 31 (1), 457-501.
- Kiefer, Cornelia (2019), "Quality Indicators for Text Data," In *BTW (Workshops)*, 145-154.
- Kim, Keehyung, Kevin Chung, and Noah Lim (2019), "Third-party Reviews and Quality Provision," *Management Science*, 65 (6), 2695-2716.
- Kirmani, Amna, 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.
- Leung, Fine, Flora Gu, and Robert Palmatier (2022), "Online Influencer Marketing," *Journal of the Academy of Marketing Science*, 50 (2), 226-251.
- Leung, Fine, et al. (2022), "Influencer Marketing Effectiveness," *Journal of Marketing*, 86 (6), 93-115.
- Luca, Michael, and Georgios Zervas (2016), "Fake it Till You Make it: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62 (12), 3412-3427.
- Mayer, Roger C., James H. Davis, and F. David Schoorman (1995), "An Integrative Model of Organizational Trust," *Academy of Management Review*, 20 (3), 709-734.

- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8), 2421-2455.
- McCarthy, Philip M., and Scott Jarvis (2010), "MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment," *Behavior Research Methods* (42) 2, 381-392.
- Meng, Weiyao, John Harvey, James Goulding, Chris James Carter, Evgeniya Lukinova, Andrew Smith, Paul Frobisher, Mina Forrest, and Georgiana Nica-Avram (2025), "Large Language Models as 'Hidden Persuaders': Fake Product Reviews are Indistinguishable to Humans and Machines," *arXiv preprint arXiv: 2506.13313*.
- Miller, Douglas L. (2023), "An Introductory Guide to Event Study Models," *Journal of Economic Perspectives*, 37 (2), 203–30.
- 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*.
- Pilgrim, Charlie, Weisi Guo, and Thomas T. Hills (2024), "The Rising Entropy of English in the Attention Economy," *Communications Psychology* 2, 70.
- Pitman, Michael (2022), "Local Consumer Review Survey 2022," (accessed September 28, 2022), [available at: <https://www.brightlocal.com/research/local-consumer-review-survey/>].

- Prweb (2021), “*STUDY: 67% of US Consumers Say Fake Online Reviews a Growing Problem,*” (accessed September 28, 2022), [available at: <https://martechseries.com/sales-marketing/customer-experience-management/uberall-67-of-us-consumers-say-fake-online-reviews/>].
- Reeder, Glenn D., and Marilyn B. Brewer (1979), "A Schematic Model of Dispositional Attribution in Interpersonal Perception," *Psychological Review*, 86 (1), 61.
- ReviewTrackers (2021), “*Online Reviews Statistics and Trends: A 2022 Report by ReviewTrackers,*” (accessed September 28, 2022), [available at: <https://www.reviewtrackers.com/reports/online-reviews-survey/>].
- Rojas, Warren (2018), “*D.C.'s Red Hen Gets Caught in Trump Aides' Dining Drama,*” (accessed November 8, 2022), [available at <https://dc.eater.com/2018/6/25/17500794/red-hen-dc-sarah-huckabee-sanders>].
- Rozin, Paul, and Edward B. Royzman (2001), "Negativity Bias, Negativity Dominance, and Contagion," *Personality and Social Psychology Review*, 5 (4), 296-320.
- Salehi-Esfahani, Saba, and Ahmet Bulent Ozturk (2018), "Negative Reviews: Formation, Spread, and Halt of Opportunistic Behavior," *International Journal of Hospitality Management*, 74, 138-146.
- 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.
- Snyder, Aaron I., and Zakary L. Tormala (2017), "Valence Asymmetries in Attitude Ambivalence," *Journal of Personality and Social Psychology*, 112 (4), 555.

- Spranca, Mark, Elisa Minsk, and Jonathan Baron (1991), "Omission and Commission in Judgment and Choice," *Journal of Experimental Social Psychology*, 27 (1), 76-105.
- Steward, Michelle D., Alvin C. Burns, Felicia N. Morgan, and Michelle L. Roehm (2020), "Credible Effects: The Impact of Disclosure of Material Connections within Online Product Reviews," *Journal of Public Policy & Marketing*, 39 (3), 353-368.
- Streitfeld, David (2012), "Buy Reviews on Yelp, Get Black Mark," (accessed September 1, 2022), [available at <https://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html>].
- Sullivan, Laurie (2022), "Fake Reviews Surged in 2021," (accessed January 19, 2026), [available at <https://www.mediapost.com/publications/article/374082/fake-reviews-surged-in-2021.html>].
- TripAdvisor (2023), "Tripadvisor Review Transparency Report," (accessed May 6, 2025), [available at <https://www.tripadvisor.com/TransparencyReport2023>].
- TripAdvisor (2025), "Tripadvisor Review Transparency Report," (accessed May 6, 2025), [available at <https://www.tripadvisor.com/TransparencyReport2025>].
- Villanova, Daniel, and Ted Matherly (2024), "For Shame! Socially Unacceptable Brand Mentions on Social Media Motivate Consumer Disengagement," *Journal of Marketing*, 88 (2), 61-78.
- Walther, Michelle, Timo Jakobi, Steven James Watson, and Gunnar Stevens (2023), "A Systematic Literature Review about the Consumers' Side of Fake Review Detection—Which Cues do Consumers Use to Determine the Veracity of Online User Reviews?" *Computers in Human Behavior Reports*, 10, 100278.

- 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.
- Weiner, Bernard (1985), "An Attributional Theory of Achievement Motivation and Emotion," *Psychological Review*, 92 (4), 548.
- World Economic Forum (2021), "Fake Reviews Cost \$152 Billion a Year. Here's How e-Commerce Sites Can Stop Them," (accessed August 30, 2022), [available at: <https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/>].
- Wu, Yuanyuan, Eric WT Ngai, Pengkun Wu, and Chong Wu (2020), "Fake Online Reviews: Literature Review, Synthesis, and Directions for Future Research," *Decision Support Systems*, 132, 113280.
- Yang, Yang, Ngee Thai Yap, and Afida Mohamad Ali (2023), "Predicting EFL expository writing quality with measures of lexical richness." *Assessing Writing* 57, 100762.
- Yelp (n.d.), "Consumer Alerts," (accessed January 29, 2026), [available at: <https://trust.yelp.com/consumer-alerts/>]
- Yelp (2023), "Compensated & Suspicious Review Activity Alerts," (accessed September 9, 2024) [available at: <https://trust.yelp.com/consumer-alerts/quarterly-alerts/>].
- Yelp (2025), "Yelp Trust & Safety Report 2024," (accessed April 5, 2025), [available at <https://trust.yelp.com/trust-and-safety-report/2024-report/>].
- Zhang, Zheng, Wenjun Zhou, and Michelle Andrews (2026), "Displaying the Amount of Consumption Time in Online Reviews can Affect Helpful Votes," *Journal of Marketing*, 90 (1), 91-105.

Web Appendix

How Fake Review Alerts Help the Platform

Web Appendix A – Study 2 Stimuli.....	2
Web Appendix B – Supplemental Study.....	6
Web Appendix C – Study 3 Stimuli & Additional Measures.....	10

These materials have been supplied by the authors to aid in the understanding of their paper. The AMA is sharing these materials at the request of the authors.

Web Appendix A

Study 2: stimuli (brand info, alert, reviews)

[Compensated Activity Alert; treatment only]



Consumer Alert: Compensated Activity Alert

We caught someone offering up payment in the form of cash, discounts, gift certificates or other incentives in exchange for someone to write, change, prevent or remove reviews for this business.

We wanted you to know because these actions not only hurt consumers, but also honest businesses who play by the rules.

Got it, thanks!

[Brand Stimulus; all participants]

West End Café
\$\$ - Restaurant, Full-service meals & Coffee
Open 6am - 8p

3.9 out of 5.0 (82 reviews)

Order Food
Delivery | Takeout

Free Delivery 50 min 30-40 mins

Delivery address

Start Order

(202) 408 - 6985

Get Directions
430 K St NW Washington, DC 20001

Hours
Mon 6AM - 8PM
Tue 6AM - 8PM
Wed 6AM - 8PM
Thu 6AM - 8PM
Fri 6AM - 8PM
Sat 8AM - 6PM
Sun 8AM - 6PM

Recommended Reviews

Search within reviews [Q] Newest First [v]

Click here to see the reviews.
You can either write a review of your own or read the experiences of other customers at West End Café.
Got it, thanks!

Recommended Reviews

Your trust is our top concern, so businesses can't pay to alter or remove their reviews. Learn more.

Search within reviews [Q] Help [v] English [v]

Start your review of West End Washington DC Tapestry Collection by Hilton.

Angie B.
Newport Beach, CA

You Might Also Consider
Sponsored

Potomac Shores Golf Club
4.5 stars (25 reviews)
"Played here again on Saturday - I think 3.5 stars is an appropriate rating at this..." read more

Days Inn & Suites by Wyndham Laurel Near Fort Meade
Days Inn, Days Hotel, and Days Inn & Suites has a property that will meet your... read more

[Reviews]

<p>Keith 5.0 out of 5 stars Great café Reviewed in the United States on May 6, 2023 Verified Purchase Great iced coffee. Really lively outdoor setting. Picnic tables and cornhole. Cute indoor area too. Very pleasant all the way around.</p>
<p>Ophal Sutton 5.0 out of 5 stars Cozy place Reviewed in the United States on May 9, 2023 Verified Purchase Outdoor friendly, shaded seating, trendy and relaxed. Staffing was very friendly. Secluded and right at the heart of downtown.</p>
<p>NC 5.0 out of 5 stars Great stuff Reviewed in the United States on April 1, 2023 Verified Purchase This place is apart of the Bode Hotel/Condo we were staying at. We went here to get coffee the morning we checked out. Service is fast, the coffee and bagel was really good. Place is very clean and cozy. Walked over at 10am and they went packed.</p>
<p>SuSu 5.0 out of 5 stars Good Reviewed in the United States on March 12, 2023 Verified Purchase Visited three times over my long weekend stay in town. Lavender latte with oat milk and yogurt parfait were both delicious. Staff was very friendly and accommodating.</p>
<p>Adam 5.0 out of 5 stars Exactly what you'd expect Reviewed in the United States on February 19, 2023 Verified Purchase Great outdoor experience with a coffee shop/ hotel vibe on the interior. Would go again.</p>
<p>Ms. W. 5.0 out of 5 stars Lots of space Reviewed in the United States on January 26, 2023 Verified Purchase This place was so cute! They have tons of outside seating which is beautiful and the inside is beautiful too with lots of plants. I would totally come back here.</p>
<p>GApeach 4.0 out of 5 stars Good treats Reviewed in the United States on February 8, 2023</p>

<p>Verified Purchase What a great coffee shop. Beautiful indoor and outdoor space including bean bag toss games. Coffee was very good and the ricotta toast was perfect. Will stop here on every visit to Nashville.</p>
<p>Harry Busey 2.0 out of 5 stars Mixed bag Reviewed in the United States on January 22, 2023 Verified Purchase Only go here for coffee/tea. My drink was good but the food was awful. The breakfast sandwich was the worst I've ever had - super dry, hard as a rock, only 2 tiny pieces of bacon, and the smallest driest egg I've ever seen.</p>
<p>Tyler Murray 2.0 out of 5 stars Could be a lot better Reviewed in the United States on January 22, 2023 Verified Purchase Staff was really nice but the owners should really do better.</p>
<p>Karen Verner 1.0 out of 5 stars Rude and confused Reviewed in the United States on June 14, 2023 Verified Purchase Looked for a dog-friendly breakfast place on google. Picked this place. After a long walk, we were greeted with a sharp we don't allow dogs per the owner. Nobody on patio and a quiet little dog at a picnic table and we were most unwanted. Fine. Outdoor table at Starbucks was great!</p>

Web Appendix B

Supplemental Study

This study was designed as a follow-up to Study 2. While Study 1 yielded similar consumer responses to both alert types, we wanted to assess if consumer response on any of our measures were different based on the alert type.

Procedure

We recruited 300 participants from Amazon mTurk. Ultimately, 301 participants ($M_{age} = 43.68$; 46% female) completed the survey in exchange for payment. Participants were randomly assigned to one of three between-subjects conditions (fake review alert: absent, suspicious, compensated). Participants followed the same procedure as Study 2. Treated participants viewed one of the alerts (see Table WB1), while control participants saw nothing. After this, they all encountered the same stimulus providing summary information about the brand (See Appendix WA). After viewing the summary information, participants responded to a series of measures.

TABLE W1: FAKE REVIEW ALERT LANGUAGE

<p>Consumer Alert: Suspicious Review Activity</p> <p>We have noticed suspicious review activity for this product. This sort of activity can take many forms, including when a number of positive reviews originate from the same IP address or when we've identified resulting from a possible deceptive review ring. Our automated recommendation software has taken this suspicious activity into account in choosing which reviews to display, but we wanted to call this to your attention because someone may be trying to artificially inflate the rating for this product.</p>
<p>Consumer Alert: Compensated Review Activity</p> <p>We caught someone offering up payment in the form of cash, discounts, gift certificates or other incentives in exchange for someone to write, change, prevent or remove reviews for this business. We wanted you to know because these actions not only hurt consumers, but also honest businesses who play by the rules.</p>

We utilized the same core measures as Study 2: a defensive processing *manipulation check*, the decision to *read reviews*, *brand intentions* (3 items; $\alpha = .88$), *review platform integrity* (3 items; $\alpha = .88$),

and *perceived ratings bias* (3 items; $\alpha = .95$). We also measured some exploratory items around *brand distrust* (3 items; $\alpha = .91$) and *fraud perceptions* ($\alpha = .81$; measured only in the suspicious and compensated alert conditions). Brand distrust was captured with agreement to the statements that “West End Café seems like a place that might: overcharge their customers when they can/try to cheat their customers/lie about their ingredients” (1 – 7: strongly disagree – strongly agree). *Fraud consequences* was assessed with the following items: “The fake reviews on West End Cafe's page was probably a _____ problem for how consumers trust the brand” (1 – 7: very small – very big) as well as agreement with the following statements: “West End Café should receive a strict punishment for the fake reviews”, “West End Café is definitely responsible for the fake reviews”, “West End Café will lose a lot of customers because of this situation”, and “Just by having fake reviews on their Yelp page, West End Café will have a major trust problem with their customers” (1 – 7: strongly disagree – strongly agree).

Results

Manipulation check. Consistent with expectations, a one-way ANOVA of the alert condition on the manipulation check yielded a significant effect ($F(2, 297) = 29.266; p < .001$). Relative to the control ($M = 3.68$), planned contrasts determined that both the suspicious ($M = 2.90; t(297) = -6.16; p < .001$) and compensated alerts ($M = 2.79; t(297) = -7.009; p < .001$) yielded decreased review credibility perceptions. There was no significant difference between the suspicious and compensated alerts ($t(297) = -.916; p = .361$).

Decision to read reviews. On average, 69% of participants chose to read some reviews. A binary logistic regression of the decision to read reviews on the alert condition yielded no significant effect (Wald (2) = 3.131; $p = .209$) reading likelihood.

Number of reviews read and average valence read. Next, we considered only those who chose to read some reviews to assess how many they read and the average valence of the reviews they read. On average, participants read 4.51 reviews and the average valence of those reviews was 3.85 (out of 5.0). One-way ANOVAs of the alert condition yielded nonsignificant effects on both the number of reviews

read ($F(2, 204) = .493; p = .611$) and the average valence read ($F(2, 204) = 1.552; p = .214$). Thus, we do not consider these results in subsequent analyses.

Review platform integrity. A one-way ANOVA of the alert condition on review platform integrity yielded a significant effect ($F(2, 297) = 4.862; p = .008$). Relative to the control ($M = 4.604$), both the suspicious ($M = 5.13; t(297) = 3.011; p < .003$) and compensated alerts ($M = 4.99; t(297) = 2.208; p = .028$) improved review platform integrity. There was no significant difference between the suspicious and compensated conditions ($t(297) = -.776; p = .438$).

Brand intentions. A one-way ANOVA of the alert condition on brand intentions yielded a significant effect ($F(2, 297) = 4.605; p = .011$). Relative to the control ($M = 4.84$), both the suspicious ($M = 4.48; t(297) = -2.005; p = .046$) and compensated alerts ($M = 4.30; t(297) = -2.975; p = .003$) decreased brand intentions. Mirroring the pilot study's effect on ratings, there was no significant difference between the suspicious and compensated alerts on brand intentions ($t(297) = -.995; p = .321$).

Brand distrust. A one-way ANOVA of the alert condition on brand distrust yielded a significant effect ($F(2, 297) = 4.039; p = .019$). Relative to the control ($M = 3.00$), both the suspicious ($M = 3.35; t(297) = 1.84; p = .067$) and compensated alerts ($M = 3.53; t(297) = 2.796; p = .006$) increased distrust. There was no significant difference between the suspicious and compensated alerts ($t(297) = .978; p = .329$).

Perceived ratings bias. A one-way ANOVA of the alert condition on perceived ratings bias yielded a significant effect ($F(2, 297) = 12.27; p < .001$). Relative to the control ($M = 2.88$), both the suspicious ($M = 3.45; t(297) = 3.216; p = .001$) and compensated alerts ($M = 3.75; t(297) = 4.87; p < .001$) increased perceptions of biased ratings. Interestingly, there was also a marginal difference between suspicious and compensated conditions ($t(297) = 1.695; p = .091$). While not hypothesized, it does suggest that suspicious alerts may yield marginally less perceived bias than their compensated counterparts.

Fraud consequences. This measure was assessed in the suspicious and compensated alert conditions only. A one-way ANOVA of the alert condition on the measure yielded no significant effect ($F(1,198) = 2.625; p = .107$).

Discussion

This study replicated the findings of Study 2, finding that the presence of alerts harms brand intentions but improves platform perceptions. Across key metrics, the suspicious and compensated alerts performed similarly, though a suspicious alert exhibited somewhat weaker effects. This makes some sense given the more circumstantial nature of the alert, which does not definitively identify the brand as the culprit. Somewhat interestingly, the alert yielded no effect on reading behavior in this study, suggesting that its effect on reading may be weaker than initially thought.

Web Appendix C

Study 3: stimuli and additional measures

Stimuli

Control (suspicion absent)

Collectible Craze: Labubu's Plush Monsters Go Mainstream

A small, wide-eyed creature with sharp ears and a mischievous grin is taking the world by storm. Known as Labubu, the character was created by artist Kasing Lung and brought to life by Pop Mart, the company behind the popular "blind box" trend. Each box hides a surprise version of the plush figure, turning every purchase into a playful gamble. Limited-edition runs have fueled collector enthusiasm, with certain designs selling out within hours online. Fashion influencers and celebrities have been spotted carrying Labubu charms on bags, helping push the toy from niche collectible to mainstream icon. For fans, it's more than a plush—it's a symbol of creativity and community.

Image of Labubu plush monsters:

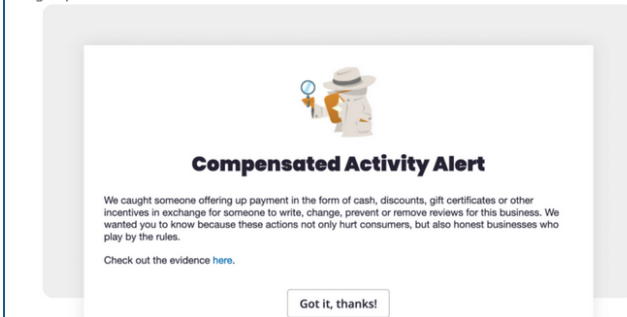


Suspicion Prime

Fake Reviews Craze: Calls for Honest Feedback

Online reviews shape where people shop, eat, and travel—but experts warn that fake reviews are on the rise. Businesses are increasingly tempted to buy positive ratings or flood competitors with negatives, leaving customers unsure who to trust. Consumer advocates say the best defense is simple: people need to write more thoughtful, detailed, and truthful reviews. Honest accounts make it harder for fraudulent ones to dominate. Yelp has also stepped in with its *Consumer Alerts* program, which posts warnings on business pages when suspicious activity is detected. These alerts, often visible for 90 days, give shoppers an extra layer of protection against misleading information.

Image of a Consumer Alert:



Video Link: https://www.youtube.com/watch?v=A9_nFLKhIB0

Additional Measures

Manipulation Check [1 – 7: Strongly disagree – Strongly agree]

1. Review platforms are monitoring the content of reviews

The prime yielded a marginal effect ($M_{\text{control}} = 4.91$, $M_{\text{prime}} = 5.12$; $F(1,599) = 3.693$; $p = .055$).

Exploratory items [1 – 7: Strongly disagree – Strongly agree]

1. It is important to write accurate reviews

The prime yielded a nonsignificant effect on the statement ($M_{\text{control}} = 6.28$, $M_{\text{prime}} = 6.37$; $F(1,599) = 1.799$; $p = .180$).

2. It is important to help others through writing good reviews

The prime yielded a nonsignificant effect on the statement ($M_{\text{control}} = 5.56$, $M_{\text{prime}} = 5.64$; $F(1,599) = .643$; $p = .423$).

