Fake or Real? The Role of Genuineness and Completeness in Social Media Brand Apologies

INTRODUCTION

Corporate crises on social media unfold rapidly and visibly, exposing brands to public scrutiny and potential reputational damage. In such settings, organizational apologies are critical tools to rebuild trust and mitigate negative reactions (Lewicki et al., 2016; Staubach & Wannow, 2024). Yet, many corporate statements adopt the form of an apology without demonstrating genuine remorse or responsibility – so-called fake apologies. Despite their prevalence in digital communication, little is known about their effectiveness compared to genuine apologies. This study addresses this gap by experimentally comparing real versus fake and incomplete versus complete apologies in a social media crisis context. Specifically, it examines their effects on perceived sincerity, response efficacy, and negative electronic word-of-mouth (eWOM), while also testing the moderating role of perceived severity.

BACKGROUND AND HYPOTHESES

Apologies are central strategies in crisis communication and image repair (Benoit, 1997; Coombs & Holladay, 2008). In their basic form, apologies consist of an explicit expression of remorse, often conveyed through an illocutionary force indicating device (IFID, e.g., "We are sorry") and an acknowledgment of responsibility (Scher & Darley, 1997; Lewicki et al., 2016). These two elements constitute the definitional minimum of an apology. More elaborate or "complete" apologies additionally include explanations, reparative actions, and a promise of forbearance (Lazare, 2005; Shao et al., 2022).

Expanding on these distinctions, we propose that a further differentiation between incomplete and fake apologies is essential. While incomplete apologies omit certain components – such as corrective action or a promise to improve – fake apologies merely mimic the linguistic form of an apology without assuming genuine accountability or moral responsibility (Staubach & Wannow, 2024). Phrases such as "We are sorry if anyone felt offended" exemplify non-apologies that deflect blame and undermine sincerity. Within Image Repair Theory (Benoit, 1997), fake apologies can be defined as hybrid strategies that blend mortification (apology) with evasion of responsibility, thereby signaling contrition without accepting blame. From the perspective of the Situational Crisis Communication Theory (SCCT; Coombs, 2007), fake apologies can be seen as a hybrid crisis response: they outwardly align with "rebuild" strategies but functionally resemble "diminish" strategies. As such, fake apologies may backfire, reducing perceived sincerity and credibility while failing to achieve reputational repair.

Public apologies on social media often address incidents that differ in how severely they are perceived by audiences. Consequently, recipients' evaluation of an apology depends not only on the message itself but also on the perceived severity of the transgression (Cheng, 2018; Shao et al., 2022). Drawing on the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986), high-severity situations trigger central processing, where recipients pay closer attention to message quality and argument strength. Under such conditions, people are more sensitive to the presence or absence of key apology components (Lee et al., 2019) and "fake" expressions. In contrast, when severity is perceived as low, recipients are likely to rely on peripheral cues, processing the apology more superficially. In these cases, even minimal or formulaic expressions of regret may suffice to restore impressions, as recipients devote less cognitive effort to evaluating the message (Petty et al., 1983; Coombs & Holladay, 2008).

To develop a first understanding of the effects of different apology types was evaluated using three key outcome variable: Perceived sincerity, response efficacy, and negative electronic word-of-mouth (eWOM).

While all real apologies contain at least the minimal components of remorse and responsibility, more complete apologies also include elements such as explanation, reparation, and forbearance. These additional components signal greater accountability and effort to restore trust, suggesting that more comprehensive apologies should yield more favorable reactions:

H_1: A complete real apology leads to a) higher perceived sincerity, b) higher perceived response efficacy, and c) lower negative eWOM intentions than an incomplete real apology.

Beyond structural completeness, genuineness plays a central role in how apologies are perceived. Thus, real apologies should evoke stronger positive evaluations than fake ones:

H_2: Real apologies result in a) higher perceived sincerity, b) higher perceived response efficacy, and c) lower negative eWOM intentions than fake apologies.

From a crisis communication standpoint, silence is often interpreted as indifference or guilt, whereas any form of apology conveys concern and responsiveness. Hence, we assume that even incomplete or fake apologies are expected to outperform a complete lack of response:

H_3: All types of apologies (real and fake) elicit a) higher perceived sincerity, b) higher perceived response efficacy, and c) lower negative eWOM intentions than no response.

Finally, the level of perceived severity determines how deeply recipients process a message. Under high-severity conditions, audiences are likely to scrutinize message quality more closely and reward more complete and authentic apologies:

H_4: The perceived severity of the transgression moderates the effects of apology type on a) perceived sincerity, b) perceived response efficacy, and c) negative eWOM intentions.

METHODOLOGY

Qualitative Pre-Study and Pre-Test

To validate the concept of fake apologies and ensure a realistic crisis context, we conducted 30 exploratory interviews analyzing real brand crises and apologies. Participants consistently described fake apologies as insincere and responsibility-avoiding, for instance: *"A classic line is 'I'm sorry that people felt offended,' which is a non-apology because it shifts the blame to others."* Another participant criticized apologies that *"sound like the company only responds because people are overreacting, not because it truly understands what it did wrong"*. Typical corporate platitudes such as *"We value all feedback"* or disclaimers like *"It wasn't meant seriously"* were also mentioned as examples of fake or hollow apologies.

Based on the pre-study and established apology frameworks, we designed four distinct apology messages: two real (short = IFID + responsibility; long = five components) and two fake apologies with similar structure but lacking genuine accountability and empathy. For example, instead of acknowledging gender stereotypes, the fake version merely expressed regret that "many users found the post offensive." A pretest (n = 126) confirmed that participants clearly

distinguished real from fake apologies in terms of perceived sincerity and genuineness ($p < .001$ for both), while still recognizing the formal similarity in structure. All messages were perceived as realistic ($M \geq 5.5$). The insights from the qualitative pre-study also guided the development of a plausible scenario involving moderate severity, where both genuine and fake responses would be credible. The scenario featured the fictitious brand *CleanGlo* and a stereotypical gender portrayal (a woman cleaning while a man relaxed on the sofa). The pretest ($n = 126$) confirmed that the incident was perceived as moderately severe ($M = 4.10$, on a 7-point scale).

Experimental Study

The main experiment was conducted on Prolific with U.S. participants. Screening criteria included regular social media use and a human intelligence task (HIT) approval rate of at least 95%. After excluding participants who failed the attention check, the final sample comprised 334 respondents (*52.6% female; M* = 38.6, *SD* = 10.51). Gender distribution did not differ across the five conditions ($\chi^2(4) = 3.07$, $p = .55$); likewise, a one-way ANOVA revealed no significant differences in pre-brand attitude across conditions ($F(4, 333) = 0.67$, $p = .61$).

Participants read a short introduction to the fictitious brand *CleanGlo*, viewed the controversial Instagram post, and were randomly assigned to one of five response conditions (two real apologies, two fake apologies, and a no-response control). The "no-response" scenario served as the control condition, informing participants that the company had not reacted to the post.

Perceived sincerity was measured with one item (*"I perceive the reaction of CleanGlo to be sincere."*). Response efficacy was assessed with two items adapted from Kirchhoff et al. (2012) and Lewicki et al. (2016) ($\alpha = .92$). Negative eWOM was assessed with three items based on Ruvio et al. (2020), adapted to the social media context ($\alpha = .93$). Perceived severity of the transgression was measured with the item *"In my opinion, the post is a serious affront."* Responses were recorded on a seven-point Likert scale (1 = strongly disagree, 7 = strongly agree). For subsequent analyses, a median split was applied (*Mdn* = 4.33) to differentiate between low- and high-severity conditions.

RESULTS AND DISCUSSION

To test the hypotheses, first, a series of one-way ANOVAs – one for each dependent variable – including appropriate post-hoc tests was conducted to examine the main effects of response type (no response, incomplete real apology, complete real apology, incomplete fake apology, complete fake apology). For perceived sincerity, the effect was highly significant, $F(4, 329) = 19.92$, $p < .001$, $\eta^2 = .20$, indicating that the type of response substantially influenced how sincere the company's statement was perceived to be. A similar pattern emerged for response efficacy ($F(4, 329) = 13.41$, $p < .001$, $\eta^2 = .14$). For negative eWOM, a smaller but still significant effect was found ($F(4, 329) = 3.69$, $p = .006$, $\eta^2 = .04$). showing that response type influenced participants' willingness to share negative comments about the company.

Post-hoc comparisons (Tukey's HSD and Games-Howell, depending on homogeneity assumptions) revealed no significant difference between the complete real apology, and the incomplete real apology on perceived sincerity ($p = .25$) and negative eWOM ($p = .16$); however, there was a marginally significant difference in terms of response efficacy ($p = .09$). H_{1a} and H_{1c} must be rejected, while H_{1b} can be marginally supported. Further post-hoc analyses indicated that only the complete real apology was perceived as more sincere than the short fake

apology (p = .030) and resulted in higher efficacy than both fake conditions (p < .001 and p = .053). No significant differences emerged for negative eWOM. Therefore, H_{2a} and H_{2b} are (partially) supported; H_{2c} is rejected. Comparing all apologies with no response, post-hoc tests revealed that all real and fake apologies were rated significantly more sincere than no response (all p < .001). For response efficacy, the results were comparable (all p < .05). Regarding negative eWOM, only the complete real apology significantly outperformed no response (p < .01). Thus, H_{3a} and H_{3b} are fully supported, while H_{3c} is rejected. Table 1 includes all mean values showing a clear pattern despite only partially significant results.

	Mean Values		
	Perceived Sincerity	Response Efficacy	Negative eWOM
No response	2,43	2,98	2,89
Incomplete fake (2 components)	3,88	3,86	2,31
Complete fake (5 components)	4,26	4,27	2,38
Incomplete real (2 components)	4,12	4,39	2,54
Compete real (5 components)	4,69	5,05	2,00

Table 1. Mean values of the dependent variables across response types

To examine whether perceived severity moderated the effects of response type, a two-way ANOVA (response type × perceived severity) was conducted for each dependent variable. The interaction effect reached significance only for negative eWOM, $F(4, 321) = 3.09$, p = .016, $\eta^2 = 21$, indicating that the impact of response type on consumers' negative word-of-mouth intentions depended on the perceived severity of the transgression. Follow-up analyses using separate one-way ANOVAs for low- and high-severity groups revealed that response type had no significant effect under low-severity conditions (p > .10). However, under high perceived severity, differences between responses were significant (p < .001): only the complete real apology effectively reduced negative eWOM intentions, whereas all other responses failed to mitigate negative reactions supporting H_{4c}.

CONCLUSIONS AND IMPLICATIONS FOR THEORY AND PRACTICE

Theoretically, this research extends Image Repair Theory and SCCT by introducing fake apologies as hybrid strategies that mimic "rebuild" responses while functionally aligning with "diminish" or "deflect" tactics. Practically, the results emphasize that genuine accountability and empathy are indispensable for successful crisis response. Companies should avoid pseudo-apologies like "We're sorry that this was perceived the wrong way" and instead communicate transparently, acknowledge wrongdoing, and outline concrete corrective steps to rebuild credibility and trust.

The findings demonstrate that apology authenticity and completeness substantially shape consumer responses to brand crises. From a managerial perspective, companies should make a deliberate effort to craft complete and genuinely authentic apologies, especially in response to serious incidents. Such statements, which combine sincerity with explicit responsibility-taking and corrective intent, are more likely to restore credibility and rebuild trust than brief or perfunctory expressions of regret.

References Available Upon Request