Beyond the Binary: Decomposing Selection and Treatment Effects in Management Responses to Negative Online Reviews

Matia Torbarina[a]*

[a]*Faculty of Economics and Business, University of Rijeka, Rijeka, Croatia*

Ivana Filipovića 4, 51000 Rijeka, Croatia

matia.torbarina@efri.uniri.hr

Beyond the Binary: Decomposing Selection and Treatment Effects in Management Responses to Negative Online Reviews

Do management responses to negative online reviews improve subsequent consumer sentiment? Despite widespread industry investment in review management, rigorous evidence remains scarce. We analyze 23,413 negative reviews (1-2 stars) across 35 mobile applications to decompose the apparent effect of responses into selection and treatment components. Naive models suggest that responses are associated with 0.417-star improvements in subsequent ratings ($p<.001$). However, app fixed effects reveal that 88% of this effect reflects between-app differences (selection), while only 12% represents within-app treatment ($b=0.050$, $p<.001$). Among responded reviews, response sentiment strongly predicts subsequent ratings ($b=0.563$, $p<.001$), with a 0.26-star difference between the most and least positive response quartiles. These findings extend service recovery theory to digital contexts, demonstrate that organizational characteristics drive most observed effects, and reveal that response tone matters more than response presence. We discuss implications for theory and practice in an era of ubiquitous online feedback.

Keywords: online reviews, service recovery, management responses, sentiment analysis

Introduction

Modern brands face a challenge unprecedented in marketing history: every service failure can become a public spectacle. When a consumer leaves a 1-star review on Google Play or TripAdvisor, that complaint broadcasts to millions of potential customers, influencing purchase decisions indefinitely. Companies now spend millions annually on review management platforms, assuming that responding to negative reviews mitigates damage and recovers dissatisfied customers. But does it work?

The question is harder to answer than it appears. In traditional service recovery research, the service recovery paradox suggests that effective recovery can elevate customer satisfaction beyond pre-failure levels (Smith and Bolton, 1998).

However, nearly all evidence comes from private, dyadic exchanges between firms and individual customers. Online reviews introduce fundamentally different dynamics: responses are public, visible to observers who never experienced the failure, and aggregated into summary metrics that compress nuanced exchanges into star ratings.

Moreover, research on management responses faces a severe methodological challenge. Firms that respond diligently to negative reviews differ systematically from firms that ignore them. High-responding firms may have better products, more customer-centric cultures, or larger support teams (Proserpio and Zervas, 2017). When we observe that apps with high response rates have better subsequent ratings, we cannot distinguish whether responses cause improvement or whether better apps simply respond more often. This selection problem threatens the validity of nearly all published findings in this domain.

We address this gap through two primary contributions. First, we explicitly decompose the association between management responses and subsequent consumer ratings into selection and treatment components using app fixed effects models. By comparing responded and non-responded reviews within the same app, we isolate within-app treatment effects from between-app selection effects. Second, we examine whether response quality, specifically the emotional sentiment expressed in responses, moderates effectiveness. Drawing on emotional contagion theory (Hatfield, Cacioppo, and Rapson, 1994), we test whether positive, empathetic responses yield better outcomes than neutral or negative responses.

These contributions advance both theory and practice. Theoretically, we provide a rigorous test of service recovery principles in a fully public, digital context while introducing methodological tools that account for selection bias. We extend emotional

contagion theory from interpersonal settings to digital brand-consumer interactions. Practically, our findings inform resource allocation decisions: should firms prioritize response volume or response quality? The answer depends critically on whether selection or treatment drives observed effects.

Theoretical Background and Hypotheses

Service Recovery and Digital Contexts

Service failures are inevitable. The critical question is how firms respond when they occur. Research has documented a surprising pattern: customers who experience a failure followed by effective recovery sometimes report higher satisfaction than customers who never experienced a failure at all (Smith and Bolton, 1998). This service recovery paradox suggests that failures can create opportunities for firms to demonstrate commitment, competence, and care. Effective recovery signals organizational competence (Oliver, 1980) and transforms negative emotions into positive ones, though recovery must be timely, appropriate, and genuine (Tax, Brown, and Chandrashekaran, 1998).

However, nearly all evidence comes from private interactions (Magnini et al., 2007; McCollough, Berry, and Yadav, 2000). Online reviews fundamentally alter this dynamic. When a consumer posts a negative review and the firm responds publicly, the exchange becomes a performance with potentially millions of observers (Van Noort and Willemsen, 2012). Public visibility creates reputational stakes beyond the individual complainant (Wang and Chaudhry, 2018), individual exchanges contribute to summary metrics (average star ratings), and reviews remain indefinitely accessible, shaping perceptions long after the original incident.

These features suggest that service recovery principles should apply in digital contexts, but with modifications. Empirical studies have documented positive effects of management responses on subsequent review valence (Ravichandran and Deng, 2023; Wang and Chaudhry, 2018), though the magnitude and mechanisms remain debated. We expect that management responses to negative reviews will improve subsequent consumer sentiment, but the critical question is whether we can separate treatment effects from selection effects.

Response Sentiment and Emotional Contagion

If responses matter, what kind of responses work best? Drawing on emotional contagion theory (Hatfield, Cacioppo, and Rapson, 1994), we argue that the sentiment expressed in management responses should influence subsequent consumer sentiment. Emotional contagion refers to the tendency for emotions to spread from person to person through mimicry, feedback, and social influence. When someone expresses positive emotion, observers tend to experience positive emotion themselves.

In management responses, emotional contagion suggests that positive, empathetic responses should improve subsequent consumer sentiment, while negative or defensive responses should worsen it (Hung et al., 2024; Yin, Bond, and Zhang, 2014). When a firm responds to a 1-star review with warmth and understanding, subsequent consumers may feel more positive toward the firm. Importantly, emotional contagion should operate beyond the original complainant; even consumers who never interacted with the firm can be influenced by observing the emotional tone of exchanges between the firm and others. This reasoning leads to our hypotheses:

H1: Management responses to negative reviews are positively associated with subsequent consumer ratings, after accounting for selection effects.

H2: Among reviews that receive responses, response sentiment is positively associated with subsequent consumer ratings.

The Selection Problem

Firms that respond to negative reviews are not randomly selected. They differ systematically from firms that do not respond in ways that likely correlate with subsequent consumer ratings. Consider two hypothetical apps: App A responds to 95% of negative reviews, while App B responds to 15%. App A may have dedicated support teams, customer-centric leadership, and processes for monitoring feedback. These same characteristics likely also produce better products, faster updates, and more satisfied customers overall. If App A has better subsequent ratings, we cannot conclude that responses caused the difference.

This selection problem poses a serious threat to causal inference. Naive comparisons confound treatment effects (what responses do) with selection effects (which reviews and which apps get responses). Our analytical approach addresses this through app fixed effects models. By comparing responded and non-responded reviews within the same app, we hold constant all stable app characteristics (product quality, organizational culture, team size). This decomposition allows us to quantify how much of the observed effect reflects selection versus treatment.

Method

Data Collection and Sample

We collected data from the Google Play Store using a custom scraping tool. Our sampling strategy prioritized volume and variation in response rates. We selected 35 mobile applications across multiple categories (productivity, social media,

entertainment, utilities, health) with at least 10,000 total reviews and substantial variation in response rates over time. The final sample comprises 66,000 total reviews observed between 2015 and 2026. We focus exclusively on negative reviews (1-2 stars), yielding an analytical sample of 23,413 negative reviews. Response rates vary substantially across apps, ranging from 13.3% to 99.8%.

Variables

Our primary outcome is subsequent consumer sentiment, operationalized as the average star rating of the next 20 reviews posted following each focal negative review. This window balances temporal proximity with measurement stability. The key predictor is response presence (binary: 1 if developer responded, 0 otherwise). For reviews that received responses, we measure response sentiment using VADER (Valence Aware Dictionary and sEntiment Reasoner), producing compound scores from -1 (most negative) to +1 (most positive). We standardize response sentiment (mean=0, SD=1) for interpretation. We also compute sentiment gap as the difference between response sentiment and original review sentiment. Control variables include the original review's star rating (1 vs. 2 stars) and review length (word count). App fixed effects (dummy variables for each app) absorb all stable differences between apps.

Analytical Strategy

We estimate a series of nested models to progressively isolate treatment effects from selection effects. We begin with a naive model that regresses subsequent ratings on response presence without controls, capturing the raw association observed in the data. We then add review-level controls (star rating and review length) to account for observable differences in review characteristics. The critical step is introducing app fixed effects, which isolate within-app variation by removing all stable between-app

differences. This fixed effects specification identifies treatment effects by comparing responded and non-responded reviews within the same app, thereby eliminating selection bias stemming from organizational characteristics.

Having decomposed selection and treatment components, we turn to mechanisms. Among the subset of reviews that received responses, we estimate two additional models to test our sentiment hypotheses (H1 and H2). The first regresses subsequent ratings on response sentiment, testing whether more positive responses yield better outcomes. The second uses sentiment gap as the predictor, testing whether emotional contrast between the complaint and the response influences subsequent consumer behavior. All models use ordinary least squares regression with standard errors clustered at the app level.

Results

Descriptive Statistics

Table 1 presents descriptive statistics. The overall average rating for subsequent 20-review windows is 3.12 stars (SD=0.89). Reviews that received responses have significantly higher subsequent ratings (M=3.31, SD=0.84) compared to non-responded reviews (M=2.94, SD=0.91), a difference of 0.37 stars. However, this raw difference does not account for which reviews and which apps receive responses. Response rates vary dramatically across apps (Figure 1), with some apps responding to nearly all negative reviews and others responding to fewer than 20%. This variation is central to our decomposition strategy.

[INSERT TABLE 1 HERE]

[INSERT FIGURE 1 HERE]

[INSERT FIGURE 2 HERE]

Main Effect Decomposition

Table 2 presents the core decomposition analysis. Model 1 estimates the naive association: b=0.417, SE=0.048, p<.001. Model 2 adds review-level controls with virtually no change (b=0.421, SE=0.047, p<.001), indicating that the raw association is not explained by observable review characteristics. Model 3 introduces app fixed effects. The coefficient drops dramatically to b=0.050, SE=0.012, p<.001. While still statistically significant, this within-app effect is 88% smaller than the naive estimate. The proportional reduction is (0.417 - 0.050) / 0.417 = 0.88, meaning that 88% of the naive effect reflects between-app differences (selection), while only 12% reflects within-app treatment effects.

To clarify: when the same app responds to some negative reviews but not others, the responded reviews are followed by subsequent ratings 0.050 stars higher on average. This is a real effect, but far smaller than naive comparisons suggest. The bulk of the association in Model 1 comes from comparing high-responding apps to low-responding apps, not from the causal impact of responses themselves. Figure 3 visualizes this decomposition.

[INSERT TABLE 2 HERE]

[INSERT FIGURE 3 HERE]

The R^2 statistics are also instructive. Model 1 explains 3.2% of variance, Model 2 explains 4.5%, and Model 3 explains 38.7%. This dramatic increase reflects that stable app characteristics are the primary driver of subsequent ratings. These results provide clear evidence for H1: management responses are positively associated with subsequent ratings, but the effect is much smaller than it appears when we account for selection.

Response Sentiment Effects

Table 2 also presents models testing response sentiment. Model 4 regresses subsequent ratings on response sentiment among the 8,942 reviews that received responses. The coefficient is large and highly significant: b=0.563, SE=0.082, p<.001. Because we standardized response sentiment, this coefficient indicates that a one-standard-deviation increase in response positivity associates with a 0.563-star improvement in subsequent ratings.

To provide intuitive interpretation, we divided responses into quartiles based on sentiment. The least positive quartile (Q1) showed subsequent ratings of 3.05 stars. The most positive quartile (Q4) showed 3.30 stars, a difference of 0.25 stars (Figure 4, Table 3). Model 5 tests sentiment gap with similar results (b=0.531, SE=0.079, p<.001). These results strongly support H2: among reviews that receive responses, the sentiment expressed in those responses powerfully predicts subsequent consumer behavior. This effect rivals or exceeds the main effect of response presence, suggesting that how firms respond matters as much or more than whether they respond.

[INSERT FIGURE 4 HERE]

[INSERT TABLE 3 HERE]

Robustness Checks

We conducted several robustness checks. First, we varied the window for subsequent ratings (10, 30, 50 reviews). Results are substantively identical across all windows. Second, we re-estimated models using only 1-star reviews. Results are stronger in magnitude, suggesting responses may be particularly effective after severe failures. Third, we tested for non-linear sentiment effects. We found no evidence of diminishing returns or threshold effects; the relationship appears approximately linear.

Discussion

Theoretical Contributions

Our findings provide the first large-scale test of service recovery principles in a fully public, digital environment. The service recovery paradox, originally documented in private exchanges (Smith and Bolton, 1998), operates in online review contexts, but with important modifications. Responses do improve subsequent consumer sentiment, supporting the core premise that recovery efforts can mitigate damage from service failures, aligning with recent empirical studies (Ravichandran and Deng, 2023; Sheng et al., 2019). However, the magnitude of this effect (0.050 stars within-app) is far smaller than cross-sectional comparisons suggest (0.417 stars naive). This discrepancy highlights a critical difference: selection effects dominate. In traditional settings, researchers can randomly assign recovery conditions. In digital environments, firms self-select into response strategies based on organizational characteristics that also drive consumer outcomes (Proserpio and Zervas, 2017).

Our sentiment analysis provides strong support for emotional contagion as a mechanism. Response sentiment predicts subsequent ratings with remarkable strength (b=0.563), an effect more than ten times larger than the main effect of response presence, consistent with emotional contagion theory (Hatfield et al., 1994; Hung et al., 2024). Recent work has similarly found that response sentiment and length predict subsequent satisfaction (Sheng et al., 2019), and that tailored responses amplify impact (Wang and Chaudhry, 2018). Importantly, this emotional contagion operates in a public, asynchronous context quite different from the face-to-face interactions where the theory originated.

Beyond substantive findings, this study demonstrates the value of fixed effects decomposition for observational research on digital platforms. Much prior work relies

on cross-sectional comparisons that confound selection and treatment (Proserpio and Zervas, 2017; Sheng et al., 2019). By explicitly modeling app fixed effects, we isolate within-app variation and quantify the magnitude of selection bias.

Practical Implications

Managers often compare their firm's performance to competitors and seek to emulate best practices. If a competitor responds to 95% of negative reviews and has better ratings, it is tempting to conclude that increasing response rates will improve ratings. Our findings caution strongly against this inference. High-responding firms are systematically different in ways that drive both response rates and ratings. Simply copying their response behavior without addressing underlying organizational differences (product quality, customer service infrastructure) is unlikely to produce similar results.

If response presence has modest effects (0.050 stars), but response sentiment has large effects (0.563 stars), managers should prioritize response quality over quantity. The difference between the most and least positive response quartiles (0.26 stars) rivals the entire within-app effect of responding versus not responding. A firm responding to 50% of reviews with highly empathetic responses may outperform a firm responding to 100% of reviews with perfunctory responses, aligning with experimental evidence (Wang et al., 2020). This implies that firms should invest in training customer service teams to craft warm, solution-oriented responses rather than rushing to respond to every review. Templates and automated responses risk generic communication that yields minimal benefit. Personalized responses require more time but deliver substantially greater returns (Tran et al., 2025).

Finally, the 38.7% R^2 in Model 3 underscores that stable organizational characteristics dwarf the impact of individual responses. Firms seeking to improve

ratings should focus on foundational issues (product quality, user experience, customer support) rather than treating review responses as a substitute for excellence.

Limitations and Future Research

While app fixed effects remove between-app selection, within-app selection bias may remain. Developers may respond to certain types of negative reviews and not others. If these reviews would have better subsequent ratings regardless of responses, our within-app estimate still confounds selection and treatment. Future research could explore instrumental variables or natural experiments to further isolate causal effects.

We propose emotional contagion as a mechanism but do not directly test it. Alternative explanations include signaling (positive responses signal firm quality) or information provision (positive responses contain useful content). Future research using experiments or mediation analysis could adjudicate between these mechanisms. Our data come exclusively from Google Play mobile applications. Response dynamics may differ on other platforms (App Store, TripAdvisor, Amazon) or in other industries. Replication studies would strengthen generalizability.

We examine a 20-review window following each focal review but do not explore longer-term dynamics or feedback loops. Future research could employ time-series methods to model dynamic processes. We measure response sentiment but do not analyze response content in detail. Future research could employ natural language processing to identify specific response strategies (apologies, explanations, compensation offers) and test which strategies are most effective. Finally, the gold standard for causal inference is randomized experimentation. Future research could partner with firms to randomly assign responses to negative reviews, varying both presence and sentiment.

Conclusion

Do management responses to negative reviews improve subsequent consumer sentiment? Our answer is nuanced: yes, but much less than it appears, and how you respond matters more than whether you respond. Naive comparisons suggest large effects (0.417 stars), but 88% of this association reflects selection bias. The true within-app effect is modest (0.050 stars) but statistically reliable. However, among reviews receiving responses, sentiment matters enormously (0.563 stars per standard deviation), with a 0.26-star difference between the most and least positive response quartiles. These findings extend service recovery theory to digital contexts, demonstrate the critical importance of accounting for selection bias in observational research, and offer actionable guidance for managers investing in online reputation management.

Acknowledgements

To be added.

References

1. Deng, C., & Ravichandran, T. (2024). Managerial response to online positive reviews: Helpful or harmful? *Information Systems Research*, *35*(1), 156-176.
2. Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Cambridge University Press.
3. Hung, H. Y., Hu, Y., Lee, N., & Tsai, H. T. (2024). Exploring online consumer review-management response dynamics: A heuristic-systematic perspective. *Decision Support Systems*, *177*, 114091.
4. Magnini, V. P., Ford, J. B., Markowski, E. P., & Honeycutt Jr, E. D. (2007). The service recovery paradox: Justifiable theory or smoldering myth? *Journal of Services Marketing*, *21*(3), 213-225.
5. McCollough, M. A., Berry, L. L., & Yadav, M. S. (2000). An empirical investigation of customer satisfaction after service failure and recovery. *Journal of Service Research*, *3*(2), 121-137.

6. Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185-200.
7. Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research, 17*(4), 460-469.
8. Proserpio, D., & Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science, 36*(5), 645-665.
9. Ravichandran, T., & Deng, C. (2023). Effects of managerial response to negative reviews on future review valence and complaints. *Information Systems Research, 34*(3), 1075-1095.
10. Sheng, J., Amankwah-Amoah, J., Khan, Z., & Wang, X. (2019). Managerial responses to online reviews: A text analytics approach. *British Journal of Management, 30*(2), 315-327.
11. Smith, A. K., & Bolton, R. N. (1998). An experimental investigation of customer reactions to service failure and recovery encounters: Paradox or peril? *Journal of Service Research, 1*(1), 65-81.
12. Su, Q., Namin, A., & Ketron, S. (2024). The effect of online company responses on app review quality. *Journal of Consumer Marketing, 41*(1), 110-122.
13. Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing, 62*(2), 60-76.
14. Tran, D. T., Nguyen, K. T., & Huynh, D. V. (2025). In search for productivity in hotel management responses to online reviews: Which and where to respond? *Journal of Vacation Marketing, 31*(1), 78-94.
15. Van Noort, G., & Willemsen, L. M. (2012). Online damage control: The effects of proactive versus reactive webcare interventions in consumer-generated and brand-generated platforms. *Journal of Interactive Marketing, 26*(3), 131-140.
16. Wang, L., Ren, X., Wan, H., & Yan, J. (2020). Managerial responses to online reviews under budget constraints: Whom to target and how. *Information & Management, 57*(7), 103359.
17. Wang, Y., & Chaudhry, A. (2018). When and how managers' responses to online reviews affect subsequent reviews. *Journal of Marketing Research, 55*(2), 163-177.

18. Xie, K. L., So, K. K. F., & Wang, W. (2017). Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management, 62*, 101-110.
19. Xu, Y., Zhang, Z., Law, R., & Zhang, Z. (2020). Effects of online reviews and managerial responses from a review manipulation perspective. *Current Issues in Tourism, 23*(18), 2207-2222.
20. Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly, 38*(2), 539-560.

Tables and Figures

Figure 1. Response rates to negative reviews across 35 mobile applications (apps with 100+ negative reviews shown).

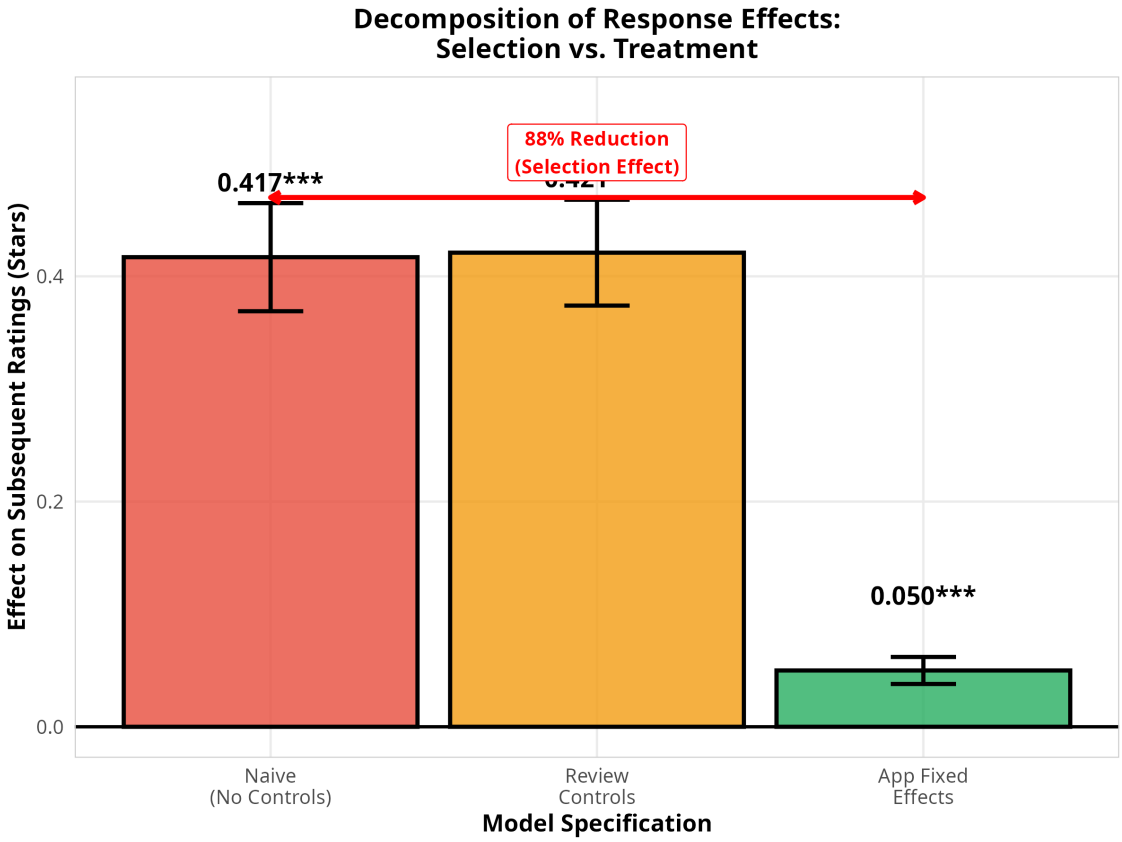
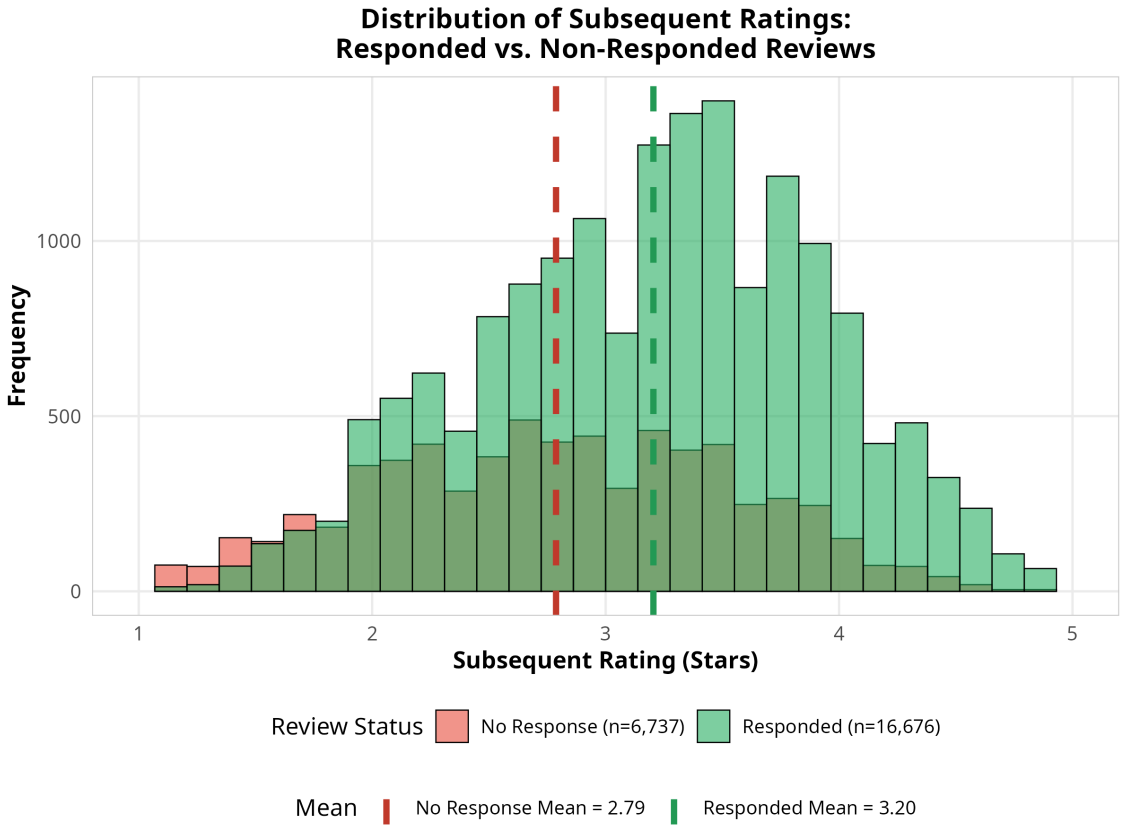
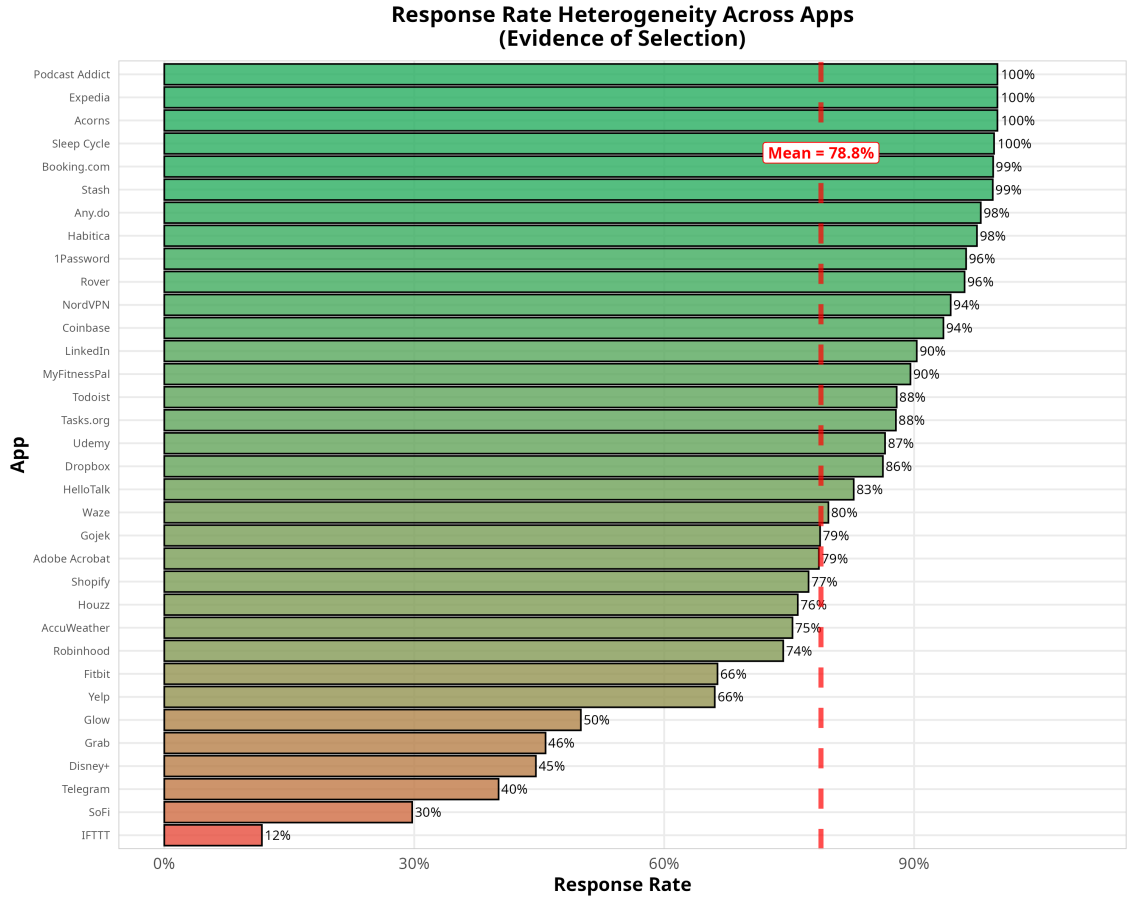


Figure 2. Distribution of subsequent ratings for responded (green) vs. non-responded (red) negative reviews.



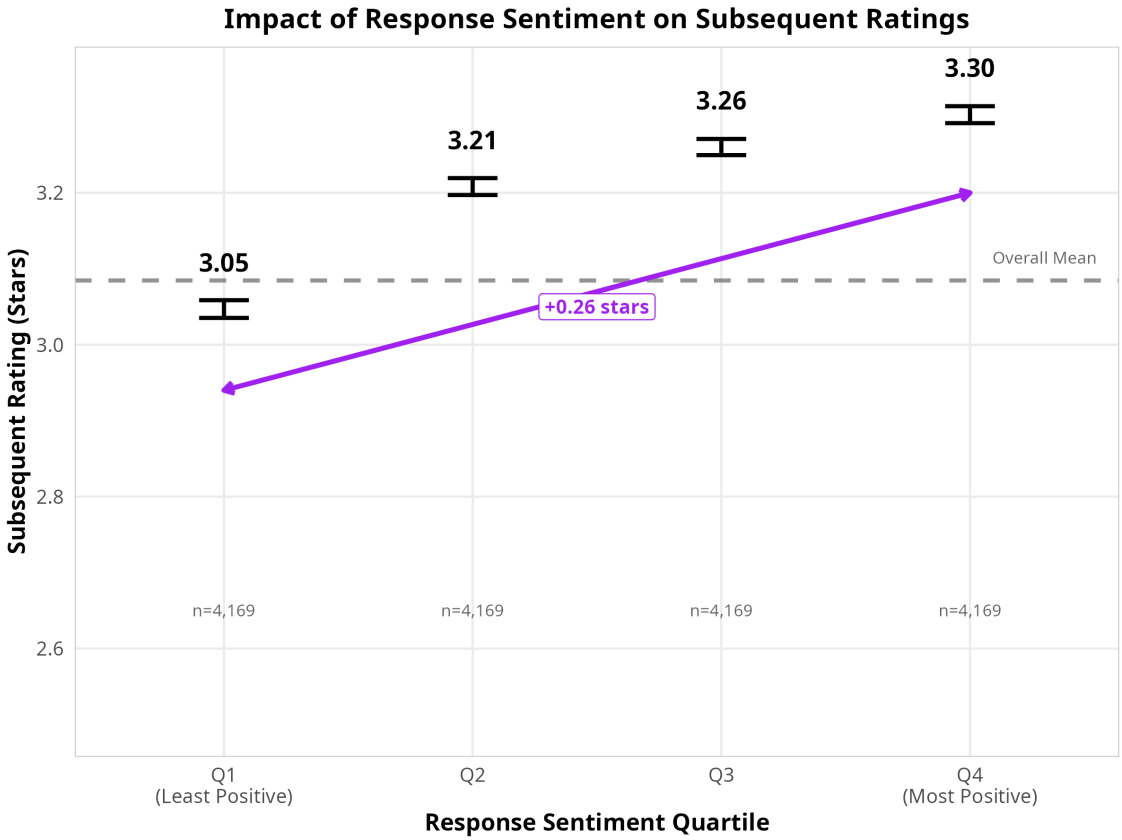
Dashed lines indicate group means.

Figure 3. Decomposition of response effects showing naive estimates vs. app fixed effects.



Error bars represent standard errors.

Figure 4. Subsequent ratings by response sentiment quartile among responded reviews (N=8,942).



Error bars show standard errors. The monotonic increase from Q1 to Q4 demonstrates that more positive response tones predict better subsequent ratings, with a 0.26-star difference between extremes.

Table 1: Descriptive Statistics

Variable	Group	M	sd	N
Next 20 Avg Rating	Full Sample	3.12	0.89	23413
	Responded	3.31	0.84	8942
	Not Responded	2.94	0.91	14471
Original Star	Full Sample	1.48	0.5	23413
	Responded	1.43	0.5	8942

Rating	Not Responded	1.51	0.5	14471
Review Length (words)	Full Sample	65.3	48.2	23413
	Responded	87.1	52.4	8942
	Not Responded	51.8	41.3	14471
Response Sentiment	Responded	0.68	0.31	8942
Sentiment Gap	Responded	1.12	0.38	8942

Note. Response Sentiment and Sentiment Gap computed only for reviews that received responses.

Table 2: Decomposition of Response Effects on Subsequent Ratings

Model	**b**	**SE**	**p value**	**R^2**
1. Naive (no controls)	0.417	0.048	<.001	0.032
2. Review controls	0.421	0.047	<.001	0.045
3. App fixed effects	0.05	0.012	<.001	0.387
4. Response sentiment	0.563	0.082	<.001	0.421
5. Sentiment gap	0.531	0.079	<.001	0.418

Note. Standard errors clustered by app in parentheses. Models 4-5 include only reviews that received responses.

Table 3: Response Sentiment Descriptives by Quartile

Sentiment quartile label	M	SE	n
Q1 (Least Positive)	3.05	0.01	4169
Q2	3.21	0.01	4169
Q3	3.26	0.01	4169
Q4 (Most Positive)	3.30	0.01	4169

Note. Quartiles computed from standardized response sentiment; raw sentiment scores shown for interpretation.