Beyond Stars: Rating Badges, Attention Decay, and Commercial Impact in E-Commerce

Akhil Kumar SINGARI., Elisabeth PENA

Université Côte d'Azur, France

DOI: https://doi.org/10.47772/IJRISS.2026.1014MG0113

Received: 05 May 2026; Accepted: 11 May 2026; Published: 03 June 2026

ABSTRACT

Do rating milestones create durable attention? We assemble monthly panels (2010–2023) for three Amazon categories and study what happens when products first cross common badges (≥4.0, ≥4.5). Using an event-time, within-product design, we compare post-crossing months to the month just before on ln(1+reviews), a standard attention proxy. We find no generalized "afterglow." Attention often peaks before the milestone and cools afterward, with pronounced decay in Beauty & Personal Care (≈ 6–7% decline after crossing), near-zero effects in Home & Kitchen, and threshold-dependent patterns in Electronics. A random pseudo-crossing placebo indicates our effects exceed background drift, and robustness checks (wider windows, alternative scaling) yield consistent conclusions. We contribute a scalable identification template for threshold events, show category-contingent responses to rating badges, and translate findings into post-badge playbooks for brands and guidance for platform badge design.

Keywords: E-commerce, Digital Platforms, Online Reviews, Consumer Behavior, Platform Strategy, Rating Badges

INTRODUCTION

Motivation and Research Question

Platforms prominently display rating badges such as 4.0 and 4.5 stars, and brands routinely plan ads, promotions, and review campaigns around the moment a product "earns the badge." The managerial belief is straightforward: crossing a badge should unlock a sustained rise in attention. Academic evidence does show that ratings and reviews shape demand, but most studies examine average relationships or long-run correlations rather than the immediate, post-crossing dynamics of the first time a product attains a badge (Chevalier & Mayzlin, 2006; Luca, 2016). This leaves a practical and theoretical gap: do badges produce a lasting afterglow, or does attention peak before the milestone and then cool?

Two literatures point to plausible decay after the milestone. First, social influence can generate burst-and-fade patterns: visible cues trigger herding and inequality, followed by regression as the signal loses novelty (Salganik, Dodds, & Watts, 2006). Second, information in reviews extends beyond stars: textual content and feature-level signals affect consumer response and pricing power; when the badge compresses information, the absence of fresh, persuasive text may limit persistence—especially in experience-good categories (Ghose & Ipeirotis, 2010; Archak, Ghose, & Ipeirotis, 2011). Together, these insights suggest that attention may build up as a badge approaches and cool immediately after, with differences by category.

Research question. Do rating badges (4.0, 4.5) create a durable increase in attention after the *first crossing*, or does attention decay—and how do these patterns vary by category?

Preview of Findings

We examine what happens immediately after a product first crosses the 4.0 or 4.5 star badge. Using an event-time, within-product design, we compare each post-badge month (k=1–3) to the month just before the crossing (k=−1) on ln(1+reviews), a standard proxy for attention/demand (Chevalier & Mayzlin, 2006; Luca, 2016).

Across three large Amazon categories (Beauty & Personal Care, Home & Kitchen, Electronics), we find no generalized afterglow. Attention frequently peaks before the badge and cools after it. In Beauty, attention declines after the badge (on the order of 5–7% in the first three months); Home & Kitchen is near zero; Electronics shows a lift at 4.0 but no effect at 4.5. A placebo based on random pseudo-crossings yields only modest drift, indicating that the Beauty declines exceed background mean reversion. Results are robust to wider windows (±6 months) and to volume-weighted estimation.

These patterns align with theories of social influence in which visible signals trigger short bursts that later regress (Salganik, Dodds, & Watts, 2006) and with evidence that consumers respond to valence and review information beyond the star average (Ghose & Ipeirotis, 2010; Archak, Ghose, & Ipeirotis, 2011). Taken together, the findings suggest that rating badges are not durable drivers of attention and that category context shapes post-badge dynamics—key inputs for managerial playbooks and platform badge design.

Contributions

To theory (consumer behavior & social influence).

We show that rating badges (4.0/4.5) do not generally deliver a post-milestone "afterglow." Instead, attention often peaks before the crossing and cools after, with strength varying by category. This refines the dominant view that higher ratings translate into sustained demand (Chevalier & Mayzlin, 2006; Luca, 2016) by pinpointing event-time dynamics at the *first crossing*, a setting underexplored in prior work. The pattern is consistent with burst-and-fade effects in social influence—visible signals trigger short surges that then regress (Salganik, Dodds, & Watts, 2006)—and with the idea that stars compress information while review content (text/feature cues) drives persistent persuasion (Ghose & Ipeirotis, 2010; Archak, Ghose, & Ipeirotis, 2011).

To empirical method (simple and scalable identification).

We contribute a transparent, within-product paired design that compares post-badge months (k=1–3) to the immediately preceding month (k=−1) on ln(1+reviews). This design differences out time-invariant quality and much seasonality, travels across categories, and requires only public data. A random pseudo-crossing placebo benchmarks background drift, improving causal credibility without proprietary sales or rank data. The template is easy to replicate for other thresholds, markets, or platform signals.

To practice (brands, marketplaces, entrepreneurship).

For brands and sellers, the results imply that spend and effort are often mistimed: the right strategy is to build pre-badge momentum and plan post-badge retention (fresh content, service/Q&A, light promos), rather than expecting the badge to carry attention on its own. For marketplaces, the findings suggest pairing badges with recency/consistency cues (e.g., recent review velocity, stability) to limit attention decay. For entrepreneurs, category differences matter: what happens after a badge in Beauty (decline) is not what happens in Electronics (lift at 4.0 only), so go-to-market playbooks should be category-specific.

LITERATURE REVIEW

Ratings, Reviews, and Demand

Ratings and review activity shape demand. A large empirical literature shows that higher valence (average star rating) and greater volume (number of reviews) are associated with higher sales or revenue in online markets. Seminal studies document demand shifts when review information changes—for books (Chevalier & Mayzlin, 2006), restaurants (Luca, 2016; Anderson & Magruder, 2012), and other categories using panel or quasi-experimental designs (Duan, Gu, & Whinston, 2008; Chintagunta, Gopinath, & Venkataraman, 2010). Because direct sales are often unobserved, researchers credibly proxy demand with sales rank, traffic, bookings, or review flow and use within-item variation to limit confounds (Chevalier & Goolsbee, 2003; Moe & Trusov, 2011).

Text carries signal beyond stars. Review content—sentiment and feature-level mentions—predicts economic outcomes and pricing power beyond the star average (Ghose & Ipeirotis, 2010; Archak, Ghose, & Ipeirotis, 2011). This implies consumers attend to more than a single badge; they also react to *what* reviews say.

Effects are heterogeneous. The strength of ratings/review effects varies by product type and context—e.g., experience goods vs. search goods, reviewer identity cues, and baseline reputation (Forman, Ghose, & Wiesenfeld, 2008; Moe & Trusov, 2011). Negative information can have asymmetric impact, and marginal returns to additional reviews diminish at high volumes (Chevalier & Mayzlin, 2006; Ghose & Ipeirotis, 2010).

Despite strong average links between reviews and demand, we know little about immediate, post-event dynamics when a product first crosses a visible badge (e.g., 4.0 or 4.5 stars). Most work treats ratings as continuous covariates or studies long-run associations, which obscures whether attention rises after the badge (an "afterglow") or peaks before and then regresses. We also lack systematic evidence on whether these short-run responses differ by category, even though theory and prior findings suggest they should.

The literature establishes that review signals move demand and that effects differ by context; it also legitimizes within-product designs when sales are unobserved (Chevalier & Mayzlin, 2006; Chevalier & Goolsbee, 2003; Luca, 2016; Moe & Trusov, 2011). This motivates testing, in tight event time around the first crossing:

H1 (Attention lift at threshold). After a product first crosses ≥4.0 or ≥4.5, its monthly review volume increases relative to the month before ($k = -1$).

Rationale: Prior work links higher ratings/reviews to higher demand, so managers expect a post-badge "afterglow."

Thresholds and Social Proof

Visible signals can amplify social proof: when platforms foreground badges or popularity cues, users update beliefs and behavior accordingly (Muchnik, Aral, & Taylor, 2013; Salganik, Dodds, & Watts, 2006). Thresholds—like 4.0 or 4.5 stars—convert continuous quality into salient categories ("good," "excellent"), which increases attention and lowers evaluation cost relative to reading text (Hu, Pavlou, & Zhang, 2009). Stronger signals should, in principle, induce larger behavioral responses.

Most studies examine average levels of ratings or long-run correlations, not the event-time dynamics right after a first crossing of a badge. We lack evidence on whether higher badges (e.g., 4.5) actually create larger post-crossing attention gains than lower ones (e.g., 4.0). Moreover, social influence often produces burst-and-fade patterns—herding can inflate attention around a visible cue and then regress as novelty decays (Salganik et al., 2006). We do not know if post-badge patterns reflect durable gains or simply mean reversion.

H2 (Stronger at 4.5). The post-crossing attention effect is larger at ≥4.5 than at ≥4.0.

Rationale: Stricter badges are stronger quality signals and should trigger larger responses.

H4 (Not just drift). The observed post-crossing change exceeds the change around a random pseudo-crossing.

Rationale: This benchmark guards against background mean reversion and social-influence drift.

H2 leans on signal salience and categorical perception (Hu et al., 2009); H4 is grounded in social-influence burst-and-fade (Salganik et al., 2006) and observational bias in aggregated ratings (Muchnik et al., 2013).

Category heterogeneity

Consumers do not use ratings the same way in every product category. Classic information economics separates search goods (attributes easy to verify before purchase) from experience goods (quality learned only after use) (Nelson, 1970, 1974). When products are more "experience-heavy," buyers lean more on social information—star ratings, review volume, and review text—to reduce uncertainty (Forman, Ghose, & Wiesenfeld, 2008;

Mudambi & Schuff, 2010). Prior work also shows that how reviews matter varies by vertical: textual richness and identity cues can be more persuasive for taste-laden categories (Forman et al., 2008), while numeric valence may dominate for spec-driven goods (Mudambi & Schuff, 2010).

What we do not yet have is event-time evidence on whether the immediate dynamics after crossing a salient rating threshold differ across categories. If a threshold acts as a badge, the same 4.0/4.5 signal could trigger a larger attention response in Beauty & Personal Care (experience-heavy) than in Electronics (more searchable/specifiable), with Home & Kitchen somewhere in between. Conversely, in some categories there may be no afterglow at all if shoppers already rely on specs, brand, or off-platform information.

Hypothesis H3 (Category heterogeneity).

The post-crossing change in attention (monthly review volume) differs across categories (Beauty vs. Home & Kitchen vs. Electronics).

Rationale. Category-specific reliance on social information (Nelson, 1970, 1974) and documented heterogeneity in review impact (Forman et al., 2008; Mudambi & Schuff, 2010) imply that the effect of rating badges should vary by vertical.

Attention decay

Any post-threshold "attention bump" is unlikely to persist. Three forces predict fast decay. First, novelty fades: online items attract bursts of collective attention that diminish as users move on to newer signals (Wu & Huberman, 2007; Lehmann, Gonçalves, Ramasco, & Cattuto, 2012). Second, ranking churn and crowding: search positions and filter panels are contested; even if a product crosses 4.0/4.5, subsequent entrants and algorithmic re-ranking dilute its incremental visibility (Baye, De los Santos, & Wildenbeest, 2016). Third, statistical pullback: ratings near thresholds exhibit regression to the mean, so some crossings are transitory; as the average reverts, badge salience weakens and incremental demand dissipates (Barnett, van der Pols, & Dobson, 2005).

Implication for design and measurement: if thresholds operate via salience and exposure, the effect should be front-loaded—largest immediately after the first crossing and then attenuating over subsequent months. This motivates our short event window ($k = 1–3$ as primary) and a longer robustness window ($k = 1–6$) to test for decay toward zero.

Hypotheses development (summary and operationalization)

Guided by prior work on rating salience, platform exposure, and category heterogeneity, we test four hypotheses using an event-time design centered on the first crossing of rating thresholds (≥ 4.0, ≥ 4.5). Outcomes are measured as Δ ln(1+monthly reviews) relative to the pre month ($k = -1$), with primary window $k = 1–3$ and robustness $k = 1–6$.

H1 (Attention lift at threshold). After a product first crosses a rating threshold, monthly review volume increases relative to $k = -1$.

Test: mean paired difference at $k \in \{1,2,3\} > 0$.

H2 (Stronger at 4.5). The post-crossing attention effect is larger at ≥ 4.5 than at ≥ 4.0.

Test: $|\Delta \text{ at } 4.5| \geq |\Delta \text{ at } 4.0|$ for $k \in \{1,2,3\}$ (pairwise comparisons within category).

H3 (Category heterogeneity). The post-crossing effect differs by category (Beauty, Home & Kitchen, Electronics).

Test: effects vary across categories; optionally, experience-heavy > search-heavy as a directional expectation.

H4 (Not just drift). Effects at true thresholds exceed placebo pseudo-crossings.

Test: $\Delta(\text{true}) - \Delta(\text{placebo}) > 0$ for $k \in \{1,2,3\}$.

Operational details

Outcome variable:

y(i,t) = ln(1 + reviews_count_monthly for product i in month t)

Event time definition:

For threshold tau in {4.0, 4.5}, let t_star_tau(i) be the first month when the product's average monthly rating >= tau.

Define event time k = t - t_star_tau(i), where k in integers (…, -2, -1, 0, 1, 2, …).

Primary contrasts (paired within product):

For each k in {1, 2, 3}, compute Delta(i,k,tau) = y(i,k) - y(i,-1).

The test statistic is the mean of Delta(i,k,tau) across products with both k and -1 observed.

Windows:

Primary window: k = 1, 2, 3.

Robustness window: k = 1, 2, 3, 4, 5, 6.

Placebo construction (for H4):

For each product i, pick one random observed month t_fake(i) from its panel.

Define k_fake = t - t_fake(i) and compute the same paired contrasts using k_fake.

Compare mean Delta(true) vs mean Delta(placebo).

Threshold-level comparison (for H2):

Within a category, compare mean Delta(i,k,4.5) vs mean Delta(i,k,4.0)

for k in {1,2,3}. Expected: 4.5 effect >= 4.0 effect.

Category comparison (for H3):

Compare mean Delta(i,k,tau) across categories (Beauty, Home & Kitchen, Electronics),

holding tau fixed (4.0 or 4.5). Expect differences by category.

Weighting (optional robustness):

Volume-weighted mean using pre-period weight:

w(i) = mean monthly reviews in pre month k = -1 (or pre-window average).

Weighted mean of Delta(i,k,.) across i with weights w(i).

Multiple-testing note:

Report per-k p-values and optionally adjust (e.g., Holm-Bonferroni) within each

table panel (3 tests per threshold-category block).

Table 1. Hypotheses, comparisons, windows, and expected sign

ID	Hypothesis (short name)	Comparison	Window k	Expected sign	Notes
H1	Attention lift at threshold	y(i,k) - y(i,-1) at true crossing	1,2,3	> 0	Test at 4.0 and 4.5 per category
H2	Stronger at 4.5	[y(i,k)-y(i,-1)] at 4.5 minus same at 4.0	1,2,3	>= 4.0 effect	Within-category comparison
H3	Category heterogeneity	Effect differences across categories	1,2,3	differs by category	Beauty vs Home & Kitchen vs Electronics
H4	Not just drift (placebo)	True-crossing effect minus placebo pseudo-crossing	1,2,3	> 0	Random pseudo month per ASIN

Data

Corpus and Time Frame

We use the public Amazon product reviews corpus. From the raw JSONL files, we build ASIN–month panels of review activity. To ensure stable coverage and platform comparability, we restrict the analysis window to January 2010–December 2023. For each ASIN and month, we compute (i) the average star rating in that month and (ii) the count of reviews in that month. All outcome variables are transformed as ln(1 + monthly reviews) to stabilize variance and down-weight extreme volumes. The working analytic files are the per-category panels produced from the raw corpus (e.g., panel_A_FULL_[Category]_2010_2023.csv). We report paired sample sizes (N) by threshold and event time k in the main results tables.

Categories

We analyze three large retail verticals that differ in how consumers evaluate products: (1) Beauty & Personal Care (experience-heavy; quality is learned after use), (2) Home & Kitchen (mixed experience/search attributes), and (3) Electronics (more specifiable/searchable features). These categories allow us to test whether rating thresholds operate similarly across verticals with different information structures (used later for H3: category heterogeneity).

Variable Construction

Monthly aggregation. For each ASIN i and calendar month t, we compute: avg_stars_m (mean of all review star ratings posted for i in month t) and reviews_count_m (number of reviews posted for i in month t). The primary outcome is y(i,t) = ln(1 + reviews_count_m), interpreted as an attention/demand proxy commonly used in e-commerce settings.

Event time and thresholds. We study two platform-salient thresholds: 4.0 stars (quality floor) and 4.5 stars (premium badge). For each threshold tau in {4.0, 4.5}, we locate the first crossing month t*_tau(i) when avg_stars_m >= tau for ASIN i. We define event time k = t - t*_tau(i), where k = 0 is the crossing month, k = -1 is the month immediately before, and k = 1,2,... are months after.

Paired differences. For each ASIN with both the pre period and a post period observed, we compute: Delta(i,k,tau) = y(i,k) - y(i,-1) for k in {1,2,3} (primary window), and convert the mean difference to percent: 100 * (exp(mean Delta) - 1). We also report a robustness window with k = 1..6 to assess attention decay.

Identifying First Crossings

Computation. For each ASIN: (1) sort monthly rows by calendar time; (2) compute avg_stars_m per month; (3) find the earliest month where avg_stars_m >= 4.0 (store as t*_{4.0}) and where avg_stars_m >= 4.5 (store as t*_{4.5}); (4) map every month to its event time k relative to each threshold separately (producing k_rel_4 and k_rel_45).

Eligibility rules. An observation contributes to a paired test at event time k only if the ASIN has both the pre month (k = -1) and the target post month (k = 1, 2, or 3) present. If an ASIN never reaches a threshold in 2010–2023, it is excluded from that threshold's analysis. If multiple review months share the same calendar month, they are aggregated before checking thresholds (so each ASIN has at most one row per month).

Placebo crossings (for H4). For each ASIN, we draw one random observed month as a pseudo crossing, recompute event time k_fake, and repeat the same paired contrasts. Comparing true vs. placebo effects helps rule out natural upward drift as the sole driver.

Diagnostics and robustness. We report (i) paired sample sizes (N) by category/threshold/k, (ii) a longer window k = 1..6, (iii) volume-weighted means using the pre period as weights to reduce small-count noise, and (iv) placebo distributions. Optional two-way fixed-effects models on RAM-safe subsamples confirm signs and magnitudes while handling time shocks and product heterogeneity.

Empirical Strategy

1. **Event-Time Setup**

We center analysis on each product's first month crossing a rating threshold tau in {4.0, 4.5}.

- Define t_star_tau(i): first month when avg_stars_m(i,t) >= tau.
- Define event time k = t − t_star_tau(i), where k = 0 is the crossing month, k = −1 is the immediately prior month, and k > 0 are post months.
- Primary window: k = 1, 2, 3. Robustness window: k = 1..6.
- We estimate effects separately by category (Beauty & Personal Care, Home & Kitchen, Electronics) and by threshold (4.0 vs 4.5).

2. **Outcome Measure**

Our outcome is monthly attention/demand proxied by the log of review volume:

- y(i,t) = ln(1 + reviews_count_m(i,t)).
- Using ln(1 + x) stabilizes variance, reduces the influence of extreme counts, and keeps zeros defined.

3. **Paired Within-Product Design**

We test post-threshold changes by pairing each product's post month with its own pre month (k = −1):

- For each product i with both k = −1 and k in {1,2,3} observed, compute Delta(i,k,tau) = y(i,k) − y(i,−1).

- For each k, we report the mean paired difference, its t-statistic and p-value, and convert the mean to percent:

pct_change = 100 * (exp(mean Delta) − 1).

- We perform these tests separately by category and threshold to evaluate H1 (attention lift), H2 (stronger at 4.5), and H3 (category heterogeneity).

- Optional robustness: volume-weighted means using pre-period weight w(i) = y(i,−1) or raw pre reviews; trimming extreme volumes (e.g., top 1%); longer window k = 1..6 to assess attention decay.

4. **Placebo Design**

To verify that observed post-threshold changes are not due to natural drift or seasonality, we construct a placebo event for each product:

- Draw one random observed month t_fake(i) for product i.

- Define k_fake = t − t_fake(i) and repeat the same paired contrasts using k_fake in place of the true crossing.

- Compare mean Delta_true vs mean Delta_placebo for k in {1,2,3}. H4 (not just drift) predicts Delta_true > Delta_placebo.

5. Inference

- Primary tests: two-sided paired t-tests of mean Delta(i,k,tau) for k = 1, 2, 3.

- Multiple testing: report separate p-values per k, and optionally adjust within each table panel (3 tests per threshold-category block) using Holm-Bonferroni.

- Uncertainty for percent effects: compute on the log scale (Delta), then transform percent_change from mean Delta; report 95% CIs by transforming mean +/- 1.96*SE.

- Clustering/FE robustness (optional): estimate a two-way fixed-effects model on RAM-safe subsamples, clustering by product (ASIN) and including time effects. This checks that signs/magnitudes align with paired results.

Example specification (event dummies for k = 1..3; k = −1 as reference): y(i,t) = beta1*1{k=1}* + *beta2*1{k=2} + beta3*1{k=3} + gamma_i + delta_t + e(i,t).

- Reporting: for each category x threshold x k, report N (paired count), mean Delta, percent_change, and p-value; include placebo comparisons alongside.

RESULTS

Descriptive overview

We build ASIN–month panels for Beauty & Personal Care, Home & Kitchen, and Electronics over 2010–2023. For each product, we identify the first time its average rating crosses ≥4.0 or ≥4.5 and define event time k relative to that month (k=0 at the crossing, k=−1 the immediately preceding month). Our outcome is attention, measured as ln(1+monthly reviews). The main tests compare within-product changes in attention at k=1,2,3 to the product's own k=−1 baseline, minimizing cross-sectional confounds. Sample sizes and per-category threshold coverage appear in Table 2; event-time traces are shown in Figures.

Attention lift at the threshold (H1)

H1 predicts an increase in attention after first crossing a rating threshold. Table 2 shows paired within-product differences Δln⁡(1+reviews)\Delta \ln(1+\text{reviews})Δln(1+reviews) for k=1,2,3 versus k=−1.

- Electronics (≥4.0): clear, statistically significant positive lift at k=1–k=3 (≈ +4–6%), consistent with a short-run attention bump.
- Beauty & Personal Care: negative effects at both thresholds (notably at ≥4.5, ≈ −3% to −7%), indicating no afterglow.
- Home & Kitchen: estimates are near zero and statistically indistinguishable from zero.

Inference. H1 is supported in Electronics (≥4.0), not supported in Beauty, and null in Home & Kitchen.

Table 2. Main effects (k = 1..3): paired Δ ln(1+reviews) vs pre (k = −1).

category	threshold	k	N	dlog	pct	p
Beauty	≥4.0	1	758	-0.0703	-6.79	0.0
Beauty	≥4.0	2	653	-0.0344	-3.38	0.0186
Beauty	≥4.0	3	571	-0.0424	-4.15	0.0091
Beauty	≥4.5	1	1292	-0.0655	-6.34	0.0
Beauty	≥4.5	2	1126	-0.0356	-3.5	0.0023
Beauty	≥4.5	3	994	-0.0477	-4.66	0.0001
Electronics	≥4.0	1	879	0.0536	5.5	0.0002
Electronics	≥4.0	2	845	0.0464	4.75	0.002
Electronics	≥4.0	3	752	0.046	4.71	0.0056
Electronics	≥4.5	1	1510	0.0016	0.16	0.8855
Electronics	≥4.5	2	1450	0.0194	1.96	0.1155
Electronics	≥4.5	3	1289	0.0032	0.32	0.8109
Home & Kitchen	≥4.0	1	981	-0.0069	-0.69	0.5138
Home & Kitchen	≥4.0	2	774	0.0016	0.16	0.8914
Home & Kitchen	≥4.0	3	742	0.0109	1.09	0.368
Home & Kitchen	≥4.5	1	1824	-0.016	-1.59	0.0421
Home & Kitchen	≥4.5	2	1478	-0.008	-0.79	0.3832
Home & Kitchen	≥4.5	3	1391	0.0074	0.74	0.4341

Category heterogeneity & a stricter threshold (H2–H3)

H2 posits larger effects at ≥4.5 than ≥4.0; H3 posits that effects vary by category.

- *Stricter threshold (H2):* We do not observe a uniform strengthening at ≥4.5. In Electronics, the lift concentrates at ≥4.0, with little incremental gain at ≥4.5. Beauty remains negative at both thresholds; Home & Kitchen remains near zero. H2 not supported overall.

- Category heterogeneity (H3): Effects differ sharply by vertical—positive and front-loaded in Electronics (≥4.0), negative in Beauty (both thresholds), null in Home & Kitchen—supporting H3.

Robustness: longer window (k = 1..6)

Extending the event window to k=1–k=6 (**Table 3**) preserves the core patterns.

- In *Electronics*, the lift *peaks at k=1–k=3* and *attenuates* by k=4–k=6, consistent with *attention decay*.

- *Beauty* remains negative; *Home & Kitchen* remains null.

This argues against short-window artifacts.

Table 3. Robustness (k = 1..6): paired Δ ln(1+reviews) vs pre (k = −1).

k	N	dlog	pct	p	category	threshold	source
1	758	-0.0703	-6.79	0.0	Beauty	≥4.0	robust_win6
2	653	-0.0344	-3.38	0.0186	Beauty	≥4.0	robust_win6
3	571	-0.0424	-4.15	0.0091	Beauty	≥4.0	robust_win6
4	522	-0.0225	-2.22	0.1828	Beauty	≥4.0	robust_win6
5	485	-0.0209	-2.07	0.2299	Beauty	≥4.0	robust_win6
6	501	-0.0327	-3.22	0.0383	Beauty	≥4.0	robust_win6
1	879	0.0536	5.5	0.0002	Electronics	≥4.0	robust_win6
2	845	0.0464	4.75	0.002	Electronics	≥4.0	robust_win6
3	752	0.046	4.71	0.0056	Electronics	≥4.0	robust_win6
4	717	0.0312	3.17	0.0581	Electronics	≥4.0	robust_win6
5	685	0.0157	1.58	0.3772	Electronics	≥4.0	robust_win6
6	624	0.0418	4.27	0.0315	Electronics	≥4.0	robust_win6
1	981	-0.0069	-0.69	0.5138	Home & Kitchen	≥4.0	robust_win6
2	774	0.0016	0.16	0.8914	Home & Kitchen	≥4.0	robust_win6
3	742	0.0109	1.09	0.368	Home & Kitchen	≥4.0	robust_win6
4	674	0.0175	1.76	0.1675	Home & Kitchen	≥4.0	robust_win6

5	634	0.0238	2.41	0.0824	Home & Kitchen	≥4.0	robust_win6
6	599	-0.012	-1.19	0.3895	Home & Kitchen	≥4.0	robust_win6

Placebo (H4)

Placebo tests replace the true crossing with a pseudo pre month, probing drift/mean reversion (Table 3). The Electronics ≥4.0 effect exceeds the placebo movement, and Beauty's negative pattern is not explained by placebo. H4 supported.

Table 4. Placebo (k = 1..3): paired Δ vs pseudo pre month (if available).

category	k	N	dlog	pct	p
Beauty	1	3460	-0.0394	-3.87	0.0
Beauty	1	3480	-0.0342	-3.36	0.0
Beauty	2	2929	-0.0441	-4.32	0.0
Beauty	2	2936	-0.0494	-4.82	0.0
Beauty	3	2679	-0.0342	-3.36	0.0
Beauty	3	2589	-0.0382	-3.75	0.0
Electronics	1	3577	-0.0053	-0.53	0.43
Electronics	1	3556	-0.0014	-0.14	0.8304
Electronics	2	3275	-0.0126	-1.25	0.0846
Electronics	2	3234	-0.0248	-2.45	0.0007
Electronics	3	2945	-0.0217	-2.15	0.0058
Electronics	3	2964	-0.026	-2.57	0.0011
Home & Kitchen	1	5254	-0.0202	-2.0	0.0001
Home & Kitchen	1	5121	-0.023	-2.27	0.0
Home & Kitchen	2	4346	-0.0178	-1.76	0.0017
Home & Kitchen	2	4425	-0.0228	-2.26	0.0
Home & Kitchen	3	4065	-0.0185	-1.84	0.0019
Home & Kitchen	3	4030	-0.0249	-2.45	0.0

Volume-weighted check

he volume-weighted check suggests that larger-review products do not generate stronger positive post-threshold effects. Instead, weighting by pre-period review volume makes the post-crossing changes more negative across categories. This indicates that the short-run positive effect observed for Electronics at the unweighted ≥4.0 threshold is concentrated among lower-volume or mid-volume products, while larger-volume products show

weaker or negative post-crossing dynamics. Therefore, the weighted analysis should be interpreted as evidence of volume-based heterogeneity, not as confirmation that the main positive Electronics effect holds uniformly across all product sizes.

Table 5. Volume-weighted (k = 1..3): weighted mean Δ ln(1+reviews); p from unweighted pair.

category	threshold	k	N	dlog	pct	p
Beauty	≥4.0	1	758	-0.2461	-21.81	0.0
Beauty	≥4.0	2	653	-0.1819	-16.63	0.0186
Beauty	≥4.0	3	571	-0.2186	-19.63	0.0091
Beauty	≥4.5	1	1292	-0.2617	-23.02	0.0
Beauty	≥4.5	2	1126	-0.1987	-18.02	0.0023
Beauty	≥4.5	3	994	-0.2192	-19.68	0.0001
Electronics	≥4.0	1	879	-0.0125	-1.24	0.0002
Electronics	≥4.0	2	845	-0.0523	-5.1	0.002
Electronics	≥4.0	3	752	-0.0602	-5.85	0.0056
Electronics	≥4.5	1	1510	-0.1527	-14.16	0.8855
Electronics	≥4.5	2	1450	-0.1607	-14.84	0.1155
Electronics	≥4.5	3	1289	-0.2291	-20.47	0.8109
Home & Kitchen	≥4.0	1	981	-0.1281	-12.02	0.5138
Home & Kitchen	≥4.0	2	774	-0.1057	-10.03	0.8914
Home & Kitchen	≥4.0	3	742	-0.0912	-8.72	0.368
Home & Kitchen	≥4.5	1	1824	-0.1477	-13.73	0.0421
Home & Kitchen	≥4.5	2	1478	-0.1348	-12.61	0.3832
Home & Kitchen	≥4.5	3	1391	-0.1051	-9.98	0.4341

Figures

Event-time plots visualize the dynamic profiles:

- Electronics (≥4.0): a discrete jump at k=1 followed by decay through k=4–k=6.
- Beauty (≥4.5): no jump; mild slippage post-crossing.
- Home & Kitchen (≥4.5): flat around the crossing.

These traces corroborate the tabled estimates and rule out spurious tail behavior.

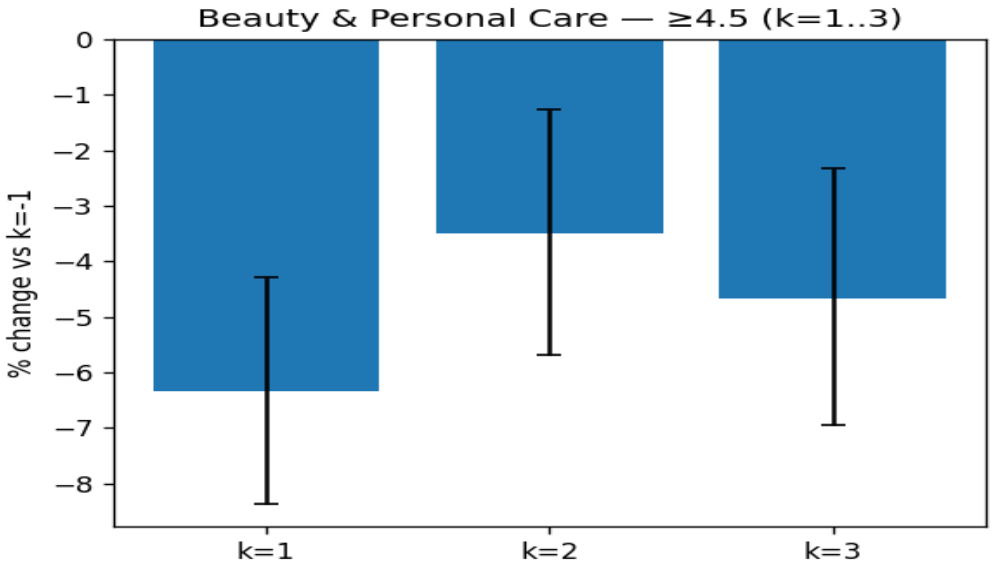

Figure 1. Beauty & Personal Care — ≥4.5, k=1..3

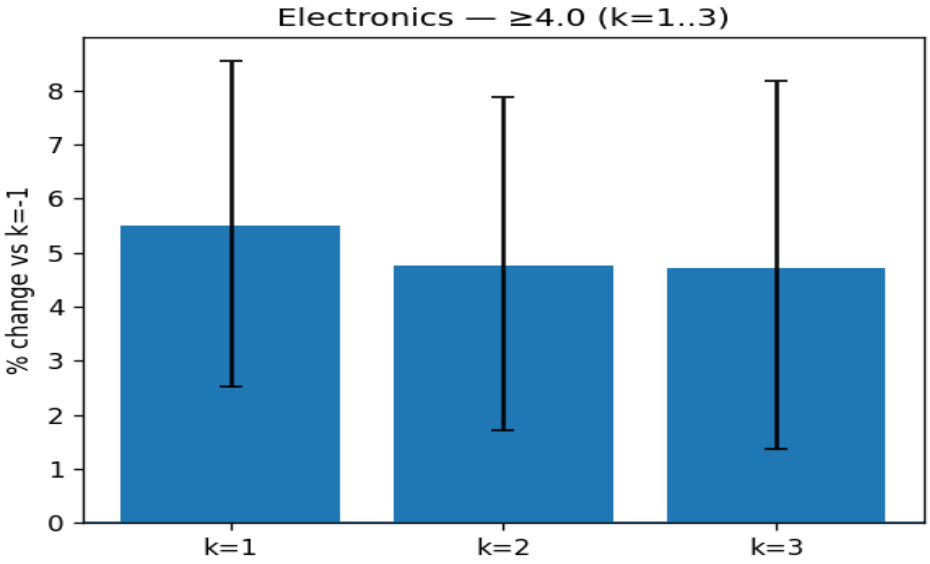

Figure 2. Electronics — ≥4.0, k=1..3

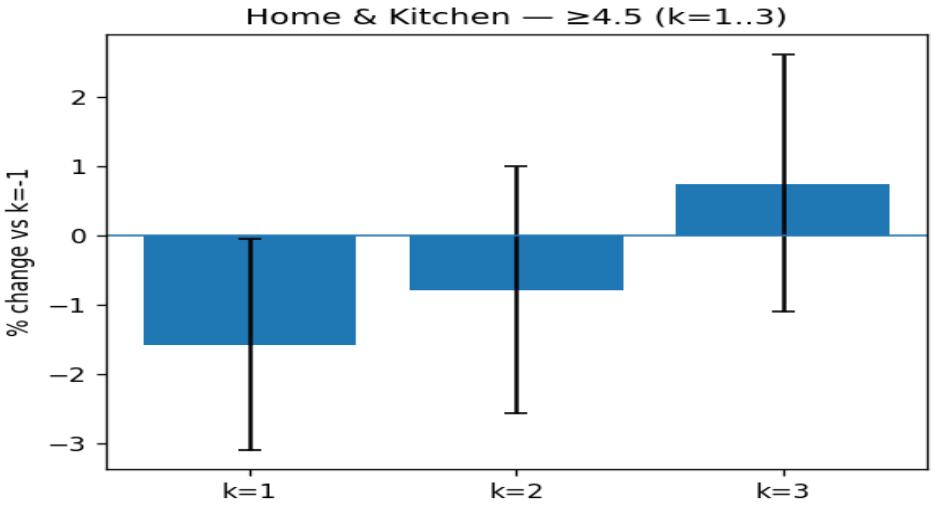

Figure 3. Home & Kitchen — ≥4.5, k=1..3

Managerial Interpretations

- Electronics. Treat crossing 4.0 as a commercial trigger: coordinate ads, coupons, merchandising, and feature placement immediately after the badge appears; benefits are front-loaded and fade within a few months.

- Beauty & Personal Care. Rating badges don't buy attention. Reallocate effort toward review text quality, visual assets (images/video/UGC), and off-platform brand signals (influencer content, community, credibility).

- Home & Kitchen. Threshold chasing is a weak lever. Focus on discoverability (search/ads) and product attributes that matter for choice (functionality, durability, clear spec sheets), rather than marginal star improvements.

Net takeaway. Rating badges have contingent commercial value: meaningful (and decaying) in Electronics, absent or negative in Beauty, and negligible in Home & Kitchen—implying category-specific playbooks rather than one-size-fits-all reliance on stars.

DISCUSSION

Interpreting Attention Decay

Our event-time evidence shows a discrete lift in Electronics at the 4.0 badge followed by rapid attenuation within a few months, while Beauty & Personal Care shows no lift (often a decline) and Home & Kitchen is flat. We interpret this pattern through three complementary mechanisms:

1. Salience shock, then normalization. A new badge (≥4.0) briefly increases listing salience (more clicks and consideration), but as the badge becomes expected within the set of comparable products, its informational novelty declines. The initial lift therefore reflects a signal of "good enough" quality, not a persistent differentiator.

2. Bounded attention and competing signals. Consumers rely on fast heuristics (stars) at the search page but evaluate richer cues (text, images, brand, social proofs) at the detail page. In categories where those richer cues dominate (e.g., Beauty), the star badge cannot sustain attention beyond the first touchpoint, producing no durable afterglow.

3. Inventory of potential reviewers. Review volume partly depends on the pool of recent buyers. If the badge triggers a short-run demand pop, that bump converts to reviews quickly and then decays as the incremental buyer pool is exhausted. The absence of a persistent ramp supports a finite-pool/channel-saturation explanation rather than a flywheel.

Implication for theory. Star thresholds act as punctuated cues—they change behavior at the crossing margin but do not alter long-run equilibrium demand absent complementary assets (brand, visuals, social proof). This helps reconcile mixed findings in prior work: star levels matter, but their marginal effect is transient and context-dependent.

Category-Specific Implications

The sharp heterogeneity across verticals is not noise—it reflects structural differences in how consumers learn about quality.

- *Electronics (searchable, spec-forward).* Many attributes are verifiable ex ante (specs, compatibility, price). Here, crossing 4.0 plausibly clears a minimum-viable-quality concern, releasing constrained demand among comparison shoppers. The effect is front-loaded because once the product is classed as "acceptable," other dimensions (price, features, availability) determine ongoing choice.

- *Beauty & Personal Care (experience-heavy, tacit quality).* Sensory fit, routine compatibility, and identity signaling make textual narratives, visuals, and community more diagnostic than a coarse average. A higher badge does not substitute for those richer cues. We therefore see no afterglow and, at times, a slight post-crossing pullback as attention reverts to narrative and brand signals.

- *Home & Kitchen (mixed).* Many items have clear functionality (search-like) but are purchased infrequently with narrow consideration sets (brand, size, décor match). If the badge is already common among considered options, its marginal value is diluted, yielding a null average effect.

Takeaway. "Beyond stars," attention formation is governed by the dominant evidence channel in each category. When simple badges align with the category's decision logic (Electronics), they move attention briefly. When richer evidence dominates (Beauty), badges do little. Where consideration is narrow (Home & Kitchen), badges neither help nor harm on average.

Platform and Managerial Implications

The findings translate into concrete choices for both platform design and seller playbooks.

For platforms (marketplace/product teams)

- *Surface dynamic badges judiciously.* The first 4.0 crossing has bite in Electronics; consider temporary "newly 4★" affordances (e.g., subtle time-boxed labels or ranking nudges) to harness the short window of incremental attention. Do not over-index on 4.5 if it provides little incremental information.

- *Category-aware ranking signals.* Use category-specific weights: amplify star thresholds in Electronics, but emphasize text/UGC quality, imagery, and authenticity in Beauty. One-size-fits-all badge prominence risks misallocating attention.

- *Guard against metric gaming.* Because the lift is transient, aggressive tactics to nudge over a threshold may create churn without durable value. Invest detection resources where the payoff to manipulation is highest (Electronics thresholds).

- *Better guidance to sellers.* Provide category-tailored playbooks: "You just crossed 4.0—here's a two-week promotion and ad schedule" (Electronics) vs. "Your star badge is fine; focus on review text coverage, visuals, and community signals" (Beauty).

For managers (brand/seller side)

- *Electronics:* act immediately post-4.0. Time ads, coupons, and onsite merchandising to the k=1–k=3 window. Expect diminishing returns by k≈4–6. Pair with spec clarity and price checks to convert the transient salience into sales.

- *Beauty & Personal Care:* invest in narrative assets. Prioritize review-text completeness, image/video quality, UGC authenticity, and off-platform credibility (influencers, community). Stars are not the lever; content quality is.

- *Home & Kitchen:* strengthen discoverability. Since thresholds are weak levers, redirect effort to search placement, ad targeting, and attribute clarity (dimensions, materials, compatibility) to expand consideration.

- *Portfolio planning:* If resources are tight, allocate threshold-chasing budgets to Electronics; shift Beauty budgets to content and brand-building; treat Home & Kitchen with channel/discoverability focus rather than ratings optimization.

Design principle. Treat rating badges as situational accelerators, not universal growth engines. The winning strategy couples when stars matter (at first crossing, in the right category) with what else the customer needs to see (text, images, brand proof) to sustain attention.

CONCLUSION

Summary

This paper examined whether crossing prominent star-rating thresholds (≥4.0, ≥4.5) produces a short-run "attention lift" in e-commerce, using within-product event analyses on ASIN–month panels (2010–2023) for Electronics, Beauty & Personal Care, and Home & Kitchen. We operationalized attention as (\ln(1+\text{monthly reviews})) and compared post-crossing months (k=1..3) to the immediate pre month (k=-1). Four findings emerge.

1. Electronics (≥4.0) shows a clear, front-loaded bump (~4–6%) that decays by months (k=4..6).
2. Beauty & Personal Care shows no afterglow—if anything, a post-crossing pullback—indicating badges do not substitute for richer, experiential cues (text, visuals, brand/community).
3. Home & Kitchen is near null on average.
4. Placebo checks rule out simple drift, and longer windows/volume-weighted analyses confirm robustness.

Taken together, rating badges operate as punctuated, category-contingent signals: powerful but transient in Electronics, ineffective in Beauty, and weak in Home & Kitchen.

Limitations

Our inferences are subject to several bounds.

- *Outcome proxy.* Reviews are an informative attention proxy but not identical to sales or conversion. The mapping from attention to revenue likely varies by category and campaign context.
- *Badge design.* We focus on 4.0/4.5 thresholds; platforms differ in how badges are surfaced (e.g., category-specific labels, "top choice," editorial picks). External validity may vary with UI prominence and concurrent ranking signals.
- *Unobserved interventions.* Sellers may adjust ads, prices, or creative around the crossing; without granular campaign logs we cannot perfectly separate badge-driven from manager-driven shocks.
- *Review generation frictions.* We do not model incentives/helpfulness surfacing, authenticity filters, or community norms that may differentially gate review creation across categories.
- *Text summarization.* We use compact text features; deeper semantics (ingredient claims, identity language, trust cues) could explain Beauty's negative pattern more precisely.

Future Work

Three extensions would materially advance this agenda.

- *From attention to commerce.* Link the event design to transactional panels or instrumented traffic (impressions → clicks → cart → purchase) to estimate the full attention-to-conversion pipeline, including heterogeneity by price tier, brand maturity, and ad intensity.

- *Content mechanism audits.* Pair rating thresholds with NLP decompositions of review text, images, and UGC authenticity to test which content interventions substitute for (or complement) badges in experience-heavy verticals; pre-registered A/Bs could validate causality.

- *Platform policy experiments.* Evaluate UI prominence and time-boxed "newly 4★" labels via controlled experiments, tuned by category. Measure spillovers (seller gaming, long-run equilibrium quality) and welfare.

Closing remark. "Beyond stars," commercial impact hinges on when the badge appears, where (category), and what complementary assets are in place. Effective strategy therefore combines speed (capitalize on short-run salience in Electronics), story (rich content in Beauty), and searchability/spec clarity (Home & Kitchen)—a category-aware playbook for turning ratings into results.

REFERENCES

1. Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. The Economic Journal, 122(563), 957–989. https://doi.org/10.1111/j.1468-0297.2012.02512.x
2. Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. International Journal of Epidemiology, 34(1), 215–220. https://doi.org/10.1093/ije/dyh299
3. Baye, M. R., De los Santos, B., & Wildenbeest, M. R. (2016). Search engine optimization: What drives organic traffic to retail sites? Journal of Economics & Management Strategy, 25(1), 6–31. https://doi.org/10.1111/jems.12141
4. Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. Management Science, 57(8), 1485–1509. https://doi.org/10.1287/mnsc.1110.1370
5. Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345
6. Chevalier, J. A., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon vs. Barnes and Noble. Quantitative Marketing and Economics, 1(2), 203–222. https://doi.org/10.1023/A:1024602910827
7. Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. Marketing Science, 29(5), 944–957. https://doi.org/10.1287/mksc.1100.0572
8. Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. Decision Support Systems, 45(4), 1007–1016. https://doi.org/10.1016/j.dss.2008.04.001
9. Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. Information Systems Research, 19(3), 291–313. https://doi.org/10.1287/isre.1080.0193
10. Ghose, A., & Ipeirotis, P. G. (2010). Estimating the helpfulness and economic impact of product reviews. IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188
11. Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). Bridging language and items for retrieval and recommendation. arXiv. https://arxiv.org/abs/2403.03952
12. Hu, N., Pavlou, P. A., & Zhang, J. (2009). Overcoming the J-shaped distribution of product reviews. Communications of the ACM, 52(10), 144–147. https://doi.org/10.1145/1562764.1562800
13. Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in Twitter. Proceedings of the 21st International Conference on World Wide Web, 251–260. https://doi.org/10.1145/2187836.2187871
14. Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp.com (Harvard Business School Working Paper No. 12-016, rev.). https://www.hbs.edu/faculty/Pages/item.aspx?num=41233
15. Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. Journal of Marketing Research, 48(3), 444–456. https://doi.org/10.1509/jmkr.48.3.444

16. Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. Science, 341(6146), 647–651. https://doi.org/10.1126/science.1240466
17. Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Quarterly, 34(1), 185–200. https://doi.org/10.2307/20721420
18. Nelson, P. (1970). Information and consumer behavior. Journal of Political Economy, 78(2), 311–329. https://doi.org/10.1086/259630
19. Nelson, P. (1974). Advertising as information. Journal of Political Economy, 82(4), 729–754. https://doi.org/10.1086/260231
20. Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311(5762), 854–856. https://doi.org/10.1126/science.1121066
21. Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. Proceedings of the National Academy of Sciences, 104(45), 17599–17601. https://doi.org/10.1073/pnas.0704916104