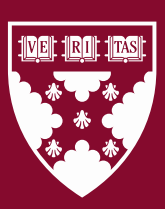
Working Paper 19-087

# A Preference for Revision Absent Improvement

Ximena Garcia-Rada
Leslie K. John
Ed O'Brien
Michael I. Norton



Harvard Business School

# A Preference for Revision Absent Improvement

Ximena Garcia-Rada
Texas A&M University

Leslie K. John
Harvard Business School

Ed O'Brien
University of Chicago

Michael I. Norton
Harvard Business School

**Working Paper 19-087**



To collect data for study 1, the authors worked closely with the team in a professional development office at a business school and are grateful for their support. Please address correspondence to Ximena Garcia-Rada.

Funding for this research was provided in part by Harvard Business School.

# A Preference for Revision Absent Improvement

XIMENA GARCIA-RADA

LESLIE K. JOHN

ED O'BRIEN

MICHAEL I. NORTON

**Author Note**

Ximena Garcia-Rada (xgarciarada@mays.tamu.edu) is an Assistant Professor of Marketing at Mays Business School, Texas A&M University; Leslie K. John (ljohn@hbs.edu) is the James E. Burke Professor of Business Administration at the Harvard Business School, Harvard University; Ed O'Brien (eob@chicagobooth.edu) is an Associate Professor of Behavioral Science at the University of Chicago Booth School of Business; Michael I. Norton (mnorton@hbs.edu) is the Harold M. Brierly Professor at the Harvard Business School, Harvard University. Please address correspondence to Ximena Garcia-Rada.

**Acknowledgments**

The authors thank Derek Koehler and Len Berry for helpful feedback, Xiang Ao for statistical advice, and Anne Marie Green, Holly Howe, Elinora Pentcheva, Shannon Sciarappa, and Shun Wang for research assistance. To collect data for study 1, the authors worked closely with the team in a professional development office at a business school and are grateful for their support.

## ABSTRACT

People regularly encounter revised stimuli (e.g., revised versions of products, new editions of books, tweaked recipes, and technological updates). In principle, a world of constant revision should benefit people by affording them the most up-to-date offerings. In practice, however, the current research reveals a potential cost: People cannot easily tell the difference between genuine improvement from merely superficial change and instead appear to uniformly assume all revised stimuli are better stimuli—even when nothing has improved in reality. Six studies document this effect and its psychological underpinnings, suggesting people rely on revision labels as an overgeneralized heuristic into stimulus quality. For example, participants became more likely to choose a stimulus merely when it was labeled as revised—and even when the product was indeed revised but made worse. Accordingly, this effect was attenuated when it was easier for participants to evaluate stimulus quality.

## RESEARCH TRANSPARENCY STATEMENT

**General Disclosures**
**Conflicts of interest:** All authors declare no conflicts of interest.
**Funding:** No funding to report.
**Artificial intelligence:** No artificial intelligence assisted technologies were used in this research or the creation of this article.
**Ethics:** This research was approved by the Institutional Review Boards at the authors' institutions; informed consent was received from all participants.
**Preregistration:** We preregistered all studies on AsPredicted.org except for Study 1, which we ran before we adopted preregistration as standard practice. We do not deviate from our preregistered plans.

**Study 1**
**Preregistration:** No aspects of the study were preregistered.
**Materials:** Questionnaires used in Studies 1A-1C has been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK). We did not post the stimuli of Studies 1B-1C because the university with which we collaborated did not allow us to do so (i.e., resumes could include identifiers or details that could reveal creators' identities)
**Data:** Raw and processed data of Studies 1B and 1C has been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK). We did not post the study 1A data because the university with which we collaborated did not allow us to do so.
**Analysis scripts:** SPSS syntax has been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK).

**Study 2A**
**Preregistration:** The hypotheses, methods, and the analysis plan were preregistered (https://aspredicted.org/8VF_HC9) prior to data collection.
**Open materials, data, and code:** Study materials, data (raw and processed), and SPSS syntax have been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK).

**Study 2B**
**Preregistration:** The hypotheses, methods, and the analysis plan were preregistered (https://aspredicted.org/541_ZR4) prior to data collection.
**Open materials, data, and code:** Study materials, data (raw and processed), and SPSS syntax have been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK).

**Study 3**
**Preregistration:** The hypotheses, methods, and the analysis plan were preregistered (https://aspredicted.org/blind.php?x=VCS_CXT) prior to data collection.
**Open materials, data, and code:** Study materials, data (raw and processed), and SPSS syntax have been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK).

**Study 4**
**Preregistration:** The hypotheses, methods, and the analysis plan were preregistered (https://aspredicted.org/1DJ_GYX) prior to data collection.
**Open materials, data, and code:** Study materials, data (raw and processed), and SPSS syntax have been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK).

**Study 5**
**Preregistration:** The hypotheses, methods, and the analysis plan were preregistered (https://aspredicted.org/TBV_RH4) prior to data collection.
**Open materials, data, and code:** Study materials, data (raw and processed), and SPSS syntax have been made available in a Research Box folder (https://researchbox.org/4113&PEER_REVIEW_passcode=GEYBKK).

## A Preference for Revision Absent Improvement

Things change. Things also *get* changed—often. For example, people regularly encounter revised versions of products, new editions of books, tweaked recipes, and technological updates. The current research asks: Do people actually prefer revised stimuli—and should they? People indeed likely prefer revised stimuli, because revisions *should* be better stimuli; after all, the creator presumably revised the stimulus to improve it. However, the current research documents that even when revised stimuli are no better in reality—and even when they are worse—people prefer them anyway, because they *assume* a quality improvement. In a public demonstration hinting at this effect (Perrigo, 2017), city passersby were invited for a sneak-peak to play with the next new iPhone, and raved about its sleek new feel and performance; unbeknownst to them, they actually had been handed a well-worn older model.

People may be right to prefer revisions on average. But as this idea suggests, people may also be at risk of *overgeneralizing* this preference to cases in which nothing has been improved. People may struggle to separate genuine improvement from merely superficial change—and thus pay costs (e.g., paying up for a revised product simply because it is labeled as revised) that leave them no better off (e.g., if the revision is not any better than its predecessor).

Why might people prefer revised stimuli (over their prior or unrevised counterparts), even in cases when there is no actual improvement? Decades of research demonstrates that people rely on heuristics to infer stimulus quality. Assessing stimulus quality can be hard to do (Payne et al., 1992; Tversky & Kahneman, 1974), and thus people draw on easier-to-evaluate peripheral cues for help. For example, people infer quality merely from the price of a product (Gneezy et al., 2014; Rao & Monroe, 1989), the duration of an experience (Yeung & Soman, 2007), the effort exerted by the creator (Kruger et al., 2004), and the visual presentation of the

content (Hagen, 2021). Doing so may be reasonable on average but can also lead people astray. For example, while healthier food is more expensive than less healthy food on average, this is not always the case. When unhealthy food is expensive, people mistakenly choose it (as they think they are making a healthy choice: Haws et al., 2017). Research on the self-fulfilling influence of expectations demonstrates that such cues not only affect people's initial choice of a stimulus; they bleed into how people's experience them. For example, drinking a nasty "vinegar" beer makes it taste worse and drinking a "pricey" wine makes it taste better, despite such labels changing nothing within the stimulus itself (for a review, see Lee et al., 2006).

Together, these prior literatures suggest that revision labels may operate the same way. When people encounter a stimulus that is labeled as revised, they may assume it *must* be improved—as they draw on their lay knowledge that revised offerings are indeed generally better in reality. Moreover, people may express this preference even after they experience the allegedly "revised" stimulus, as positive expectations may foster favorable construal and interactions (as in our earlier iPhone anecdote). Somewhat problematically, people may pay up for revised offerings that they merely believe are better.

To state our hypotheses more formally:

**H1:** Merely labeling a stimulus as "revised" will increase people's preference for it. Note that we will test for this effect while holding constant the stimulus itself (the "revision" is merely a label), or even when the revision is a step backward—i.e., when it degrades or worsens the product.

**H2:** The effect will be mediated by quality inferences.

**H3:** The effect will be moderated by the evaluability of stimulus quality. That is: If the effect indeed operates via the heuristic logic as proposed, then it should be attenuated when it is

easier for people evaluate stimulus quality (Gigerenzer & Gaissmaier, 2011; Payne et al., 1988; Tversky, 1972)—meaning, when people no longer seek to draw on the revision label for help.

We tested these hypotheses across six studies, spanning a diversity of stimuli, contexts, and methodologies. Table 1 provides a summary.

*TABLE 1: SUMMARY OF STUDIES*

| Study | Stimuli | Design | Primary Outcome(s) | Main Finding |
|---|---|---|---|---|
| 1A-1C | Resumes | 2 (version: original, revised) × 2 (label: accurate, inaccurate) | Appeal | • Participants rated a resume labeled "revised" as more appealing than one labeled "original"—regardless of label accuracy (H1). |
| 2A | Selfie sticks | 2 (label: control, revised) | Choice | • Participants preferred a shorter selfie stick when labeled "newer version" over a longer, superior counterpart (H1). |
| 2B | Guidebooks | 2 (label: control, revised) | Choice | • Participants preferred a useless guidebook when labeled "updated edition" over a superior counterpart (H1). |
| 3 | Video games | 2 (label: version 2, version 5) | Quality, enjoyment | • Participants identified fewer bugs in a video game labeled version 5 (vs. version 2); in turn, they enjoyed playing the game more (H2). |
| 4 | Product sets | 2 (label: unrevised, revised) × 2 (evaluability: easy, hard) | Accuracy (incentivized) | • Participants' reliance on this heuristic was moderated by how easy (vs. hard) it was to evaluate stimulus quality (H3). |
| 5 | Paper towels | 2 (revision label: absent, present) × 2 (evaluability: easy, hard) | Quality, preference | • Participants' reliance on this heuristic was moderated by how easy (vs. hard) it was to evaluate stimulus quality (H3). |

Across studies, we reported all manipulations, measures, and exclusions (if any). We predetermined our sample sizes before any data were collected using a rule of thumb to attain at least 200 participants per condition, minding financial considerations. Sensitivity analyses using *GPower* (Faul et al., 2007) to determine the smallest effect size given our sample sizes and statistical criteria (e.g., an alpha level of 0.05 and 80% statistical power) confirmed that our studies are adequately powered to detect small-to-moderate effects (Cohen, 1988). We conducted robustness checks excluding participants who failed the comprehension checks (see SOM-A).

## Study 1: Basic Effect

In Study 1, we tested for the basic effect in a naturalistic setting using a commonly revised stimulus to assess whether merely labeling a resume as "revised" may increase its appeal (H1).

**Method**

*Study 1A: Creators.* This study took place during a process to help MBA students at a Northeastern business school in the United States to prepare their resumes. We collaborated with the school's Career and Professional Development (CPD) team during their process to help students revise their resumes: students submit their initial resume to CPD; receive feedback over the course of two months from a resume coach; and then submit a revised resume to prospective employers. This process granted us the opportunity to analyze a natural set of revised stimuli.

Students (N = 302) requested feedback. After they had submitted their revised version, CPD asked whether they were willing to be contacted about the review process, and we obtained the email addresses of the 77.1% of students who agreed (233 out of 302). We emailed a short survey to these 233 students, hereafter referred to as "creators." A priori, we decided to send one follow-up email to non-responders and to close the survey once two days passed without any new survey responders, resulting in a response rate of 18.9% (44 out of 233).

In the survey, these 44 creators rated four items. First, they rated, "What percentage of the final version of your resume is different from your original resume?" on a 0–100 scale with endpoints labeled 0 (*the final version is exactly the same as the original*) and 100 (*the final version is completely different from the original*). Second, they rated, "Relative to my original resume, my final resume is…" using the options: -3 (*dramatically worse*), -2 (*moderately worse*), -1 (*a bit worse*), 0 (*about the same*), 1 (*a bit better*), 2 (*moderately better*), and 3 (*dramatically*

*better*). Third, they rated, "How satisfied are you with the final version of your resume?" on a scale from 1 (*not at all satisfied*) to 7 (*extremely satisfied*). Fourth, they rated, "How many times did you obtain feedback from a resume coach?" via a numeric text entry box.

This survey thus allowed us to assess whether creators indeed revised their resumes and viewed the revised resume as improved—which was their goal. At the end of the survey, we asked creators whether we could use their (anonymized) resumes for future research. Of these 44 creators, data from eleven creators were excluded: five creators did not grant us permission to use their resumes, and six creators did not answer the question "What percentage of the final version of your resume is different from your original resume?". This process resulted in data from 33 creators (i.e., 33 pairs of resumes) to be used in Studies 1B-1C.

*Study 1B: Observers evaluating resumes without labels.* Next, an independent sample of participants evaluated one of the resume pairs generated in Study 1A. These observers ($N$ = 204, 46.6% male*;* $M_{age}$ = 35.58 years, $SD$ = 13.83) were community members who participated via in-person laboratory sessions. Observers rated both versions of one randomly selected resume pair from Study 1A (i.e., one student's original resume and the same student's revised resume; the name of each creator in all resumes was replaced with the generic name "Alex Newman" with all other identifiers removed). However, resumes did not have labels of original and revised. Instead, observers only read: "On the next pages, you will see two resumes and you will be asked to rate the appeal of each resume." Each resume was presented on a different page on a computer screen and observers were asked to rate the "overall appeal" on a scale from 1 (*very low*) to 7 (*very high*). Between-subjects, we manipulated the order in which the resumes were presented. Some of these observers rated the original (though it was not labeled as original), followed by its corresponding revision (again, the revision was not labeled as such); others were presented with

the versions in the opposite order (i.e., the revision first, followed by the original version, though they were not labeled as such). Thus, we ended up with a pool of 66 resume pairs (i.e., we had a pool of 33 resumes pairs in Study 1A where we counterbalanced which version was judged first).

*Study 1C: Observers evaluating resumes with labels.* Finally, another independent sample of participants evaluated the same resume pairs. These observers ($N$ = 453, 50.1% male; $M_{age}$ = 30.27 years, $SD$ = 12.27) were community members as in Study 1B. Again, observers rated both versions of one randomly selected resume pair (i.e., one student's original resume and the same student's revised resume)—here, however, the resumes were indeed labeled. First, these observers were shown a version and were asked to "please rate the ORIGINAL DRAFT of the resume with respect to the following dimension: Overall appeal" on a 1 (*very low*) to 7 (*very high*) scale. Next, they were shown the other version in the pair and were asked to "please rate the REVISED DRAFT of the resume with respect to the following dimension: Overall appeal" on a 1 (*very low*) to 7 (*very high*) scale.

Between-subjects, we manipulated whether the resume that was labeled "original" was truly the original, and whether the resume that was labeled "revised" was truly the revision. That is, some of these observers simply rated the actual original, followed by its actual revision; we swapped these labels for other observers (unbeknownst to them). This setup produced 66 resume pairs (i.e., a pool of 33 control pairs, in which actual originals were paired with their actual revisions; and a pool of 33 experimental pairs, in which actual revisions labeled as "original" were paired with their corresponding originals labeled as "revision"). Thus, each observer was randomized to rate both versions of a randomly selected resume pair from either the control pool or the experimental pool. Our results collapse across the 33 resumes within each pool. We

obtained demographics from participants in Studies 1B-1C from a separate study completed during the same session.

**Results**

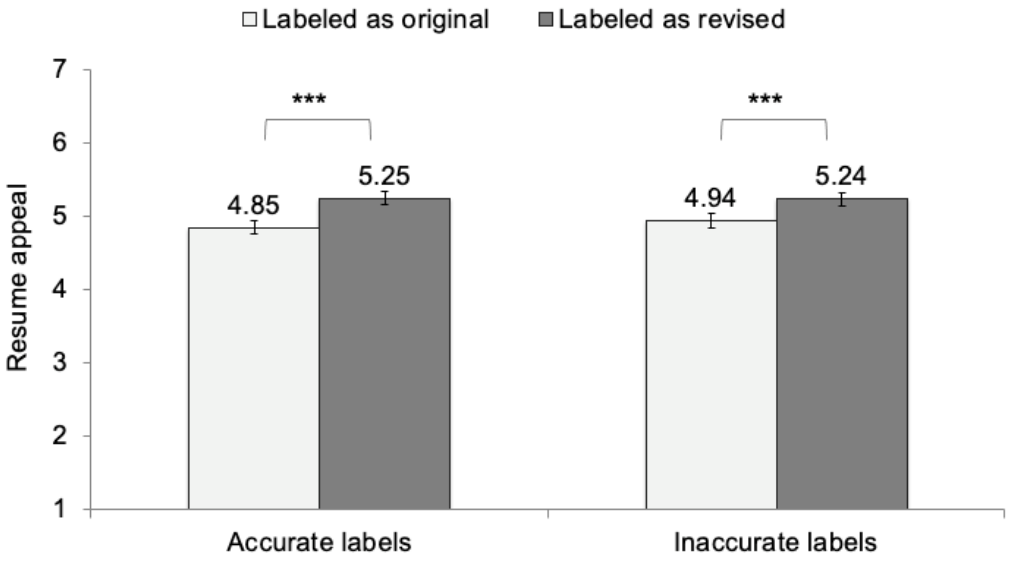
*Study 1A. Do creators think their revisions are improvements?* Creators deemed their revised resume to be significantly different from their original resume ($M$ = 36.34%, $SD$ = 26.33; one sample $t$-test against 0%, $t(37) = 8.51$, $p < .001$), and deemed their revised resume to be significantly better than their original resume ($M$ = 1.55, $SD$ = 0.90; one sample $t$-test against the scale midpoint, $t(43) = 11.38$, $p < .001$). Moreover, the more dissimilar their two versions were, the higher quality the creators perceived their revised resume ($r(38) = .61$, $p < .001$).

*Study 1B. Are the revisions really any better than the originals?* When the resumes were simply presented as "different" versions—i.e., without revision labels that may elicit our effect—observers judged both resumes in the pair as similarly appealing ($M_{revised}$ = 5.01, $SD$ = 1.43 vs. $M_{original}$ = 5.07, $SD$ = 1.47). A 2 × 2 mixed ANOVA revealed that resumes that had been revised by creators in Study 1A were judged as no better than the original resumes ($F(1, 202) = 0.99$, $p$ = .321, $\eta_p^2 < 0.01$). Resumes presented on the first screen were considered equally appealing as resumes presented on the second screen (no effect of presentation order: $F(1, 202) = 0.32$, $p$ = .572, $\eta_p^2 < 0.01$); the interaction was not significant ($F(1, 202) = 1.72$, $p$ = .191, $\eta_p^2 = 0.01$).

*Study 1C. Are resumes evaluated more positively when labeled revised?* Observers exhibited a preference for resumes labeled as "revisions," independent from the veracity of this label. Specifically, observers perceived resumes labeled as revisions more positively than those labeled as originals ($M_{revised}$ = 5.25, $SD$ = 1.37 vs. $M_{original}$ = 4.89, $SD$ = 1.36; $F(1, 451) = 51.61$, $p$ < .001, $\eta_p^2 = 0.10$); the main effect of accuracy and the interaction were not significant ($p$s >

.250). When resumes were labeled accurately, participants perceived revised resumes to be of higher quality than original ones ($M_{revised}$ = 5.25, $SD$ = 1.37 vs. $M_{original}$ = 4.85, $SD$ = 1.34; $F$(1, 451) = 33.86, $p$ < .001, $\eta_p^2$ = 0.07). Yet when resumes were labeled inaccurately, participants *still* perceived resumes labeled as revised to be of higher quality than resumes labeled as original ($M_{revised}$ = 5.24, $SD$ = 1.37 vs. $M_{original}$ = 4.94, $SD$ = 1.39; $F$(1, 451) = 18.83, $p$ < .001, $\eta_p^2$ = 0.04; see Figure 1).

*FIGURE 1: STUDY 1C RESULTS*




*NOTE.*—Error bars indicate ±1 SEM. *$p$ < .05, **$p$ < .01, ***$p$ < .001.

In sum, Studies 1A-1B-1C provide initial support for our primary hypothesis (H1): People judged the same resume more favorably merely when it was labeled as revised, irrespective of reality.

## Studies 2A-2B: Choice

In Studies 2A-2B, we tested whether, within an incentive-compatible design, people are more likely to choose a product merely when it is labeled as revised. Moreover, we tested whether this effect holds even when the revised product is actually the *worse* choice option.

**Study 2A: Method**

*Participants*. We recruited U.S.-based adults ($N$ = 602; 51.0% male; $M_{age}$ = 42.01 years, SD = 12.23) from Amazon Mechanical Turk (MTurk) to complete a study for a fixed payment.

*Procedure*. Participants chose between two selfie sticks ostensibly listed on Amazon.com. They were shown information on the two options, including a photo, brand, price, dimensions, and features (see stimuli in SOM-B). These products were sold by different fictitious companies (Utechnologies and TechLog), but otherwise were identical in price, color, and other features. To control for chronological newness, we described both products as being "Released 9/1/21."

The critical difference between the two selfie sticks was their extension length: one extended to 24 inches (the objectively better option) and the other extended to 16 inches (the objectively worse option). Via a pretest, we confirmed that participants in this population indeed deem longer selfie-sticks as better than shorter ones (see SOM-C). Between-subjects, we randomly assigned participants to one of two conditions: those in the *control* condition saw the two options without labels while those in the *revised* condition read that the inferior option had been revised. For ecological validity, we denoted revision in this study following Amazon's practices for communicating the existence of newer or older versions of the same item: participants in the *revised* condition read the inferior selfie stick was a "NEWER Version" and were informed about the existence of an older version, also available on Amazon.

Participants chose one of the two selfie stick options and their choice was made consequential using a lottery paradigm: prior to choosing, participants were informed that one participant would be selected at random and would receive a gift card to purchase their chosen selfie stick. Finally, participants answered a comprehension check asking them to identify how the selfie sticks were different. This and subsequent studies concluded with standard demographics. Once data collection was completed, we selected a winner and sent them an MTurk bonus of $24 (i.e., the price of the selfie stick).

**Study 2A: Results**

More participants chose the worse selfie stick (over the better one) when that worse option was labeled as revised (chosen by 30.9% of these participants) versus when it had no such label (chosen by 14.5% of these participants), $\chi^2(1, N = 602) = 23.14$, $p < .001$, $\phi = 0.20$.

**Study 2B: Method**

*Participants.* We recruited U.S-based adults ($N = 401$; 42.9% male; $M_{age} = 39.53$ years, SD = 13.76) from Prolific Academic to complete a study for a fixed payment. We used the platform's pre-screeners to exclude Massachusetts-based participants from being able to sign up, due to our stimulus (see below).

*Procedure*. Participants chose between two guidebooks featuring Boston-area restaurants that they would download and keep for a possible future trip. To help them choose, participants were shown, side-by-side, the cover of the two guidebooks, which included the list of the eight restaurants featured in the given guidebook. Each restaurant name was a clickable link; we encouraged participants to click them to get more information on the restaurants.

If a participant clicked a given restaurant's link, a new browser page opened with the results of the live search result for that specific restaurant. Critically, one guidebook featured restaurants that were currently open (the objectively superior option), while the other featured restaurants that had closed in the past couple of years (the objectively inferior option, as this guidebook would be useless for their future Boston trip). However, participants would only learn about the status of these restaurants if they clicked the links. Upon clicking a restaurant link in the superior guidebook, they would see a small map showing the restaurant's location and key information (including a link to the restaurant's website and menu). Upon clicking a restaurant link in the inferior guidebook, participants would see a red banner reading "Permanently closed" (see SOM-B for screenshots of what participants saw).

Between-subjects, we randomly assigned participants to either a control or a revised condition. Those in the control condition saw the guidebooks without labels; those in the revised condition saw the same two guidebooks but the inferior one was labeled "updated edition." Participants selected a guidebook, which they would receive at the end of the study. On the next pages, participants elaborated briefly on why they chose the given guidebook; indicated how many, if any, restaurant links they clicked on; and completed a comprehension check in which they indicated how the guidebooks differed.

**Study 2B: Results**

More participants chose the worse guidebook (over the better one) when that worse option was labeled as revised (chosen by 43.0% of these participants) versus when it had no such label (chosen by 21.9% of these participants), $\chi^2(1, N = 401) = 20.39$, $p < .001$, $\phi = 0.23$.

In an exploratory analysis, we also examined whether participants clicked on links to view information about the restaurants. A smaller proportion of participants in the revised condition indicated that they clicked on at least one restaurant link (revised = 64.5% vs. control = 76.6%, $\chi^2(1, N = 401) = 7.09$, $p = .008$, $\phi = 0.13$). This self-reported behavior is consistent with other exploratory results we gleaned from Qualtrics meta-data, which measured (a) the number of clicks participants made, and (b) the time participants spent evaluating the guidebooks: Participants in the revised condition made fewer clicks than participants in the control condition ($M_{revised} = 8.80$, SD = 8.63 vs. $M_{control} = 10.43$, SD = 9.32, $t(399) = 1.83$, $p = .069$, $d = 0.18$); they also spent less time evaluating the guidebooks ($M_{revised} = 1.49$ minutes, SD = 1.29 vs. $M_{control} = 2.02$ minutes, SD = 2.13, $t(328.66) = 3.03$, $p = .003$, $d = 0.30$). These findings suggest that participants indeed treated "revised" options in ways that appear consistent with heuristic processing, such that the label led them to evaluate the stimuli more quickly and less critically. Together, Studies 2A-B provide further support for our primary hypothesis (H1), here in contexts of real choice.

## Study 3: Mediation by Quality Inferences

In Study 3, we tested the proposed process (H2) such that a revision label may increase people's preferences because it may lead to more favorable quality inferences. We tested this idea even while allowing participants to directly interact with the allegedly revised stimulus.

### Method

*Participants*. We recruited U.S.-based adults ($N = 500$; 53.8% male; $M_{age} = 36.02$ years, SD = 11.14) from MTurk to complete a study for a fixed payment.

*Procedure*. The study was a two-condition between-subjects design in which we manipulated whether a video game was labeled as Version 2 or Version 5. To begin, all participants read that we—the requesters of the MTurk HIT—had been developing a game called *ART Time* and were shown a brief description of it. The game allows users to freely "paint" a blank canvas, with a selection of different tools that create different shapes and colors. We hired a developer to create the game for us, ensuring that all participants would have the same novel experience. Moreover, we instructed the developer to design the game with buggy features (e.g., the cursor would sometimes lag a split-second behind).

All participants read that, to date, we had released five updates to the game. Participants in the *revised* condition were informed that they had been randomized to play "Version 5 of 5 (i.e., our newest update)," whereas participants in the *control* condition were informed that they had been randomized to play "Version 2 of 5 (i.e., an older, pre-updated version)." In both conditions, we described the given version as having been developed in "early 2019."

Next, all participants played the game for two minutes, during which all other keyboard controls were disabled. When time expired, the page automatically continued to a survey screen where participants rated a block of enjoyment questions and a question about the number of bugs they experienced while playing (thus serving as a direct measure of perceived quality); the order of the two blocks was randomized, each presented on individual pages. In one block, participants rated five items each on a 1 (*not at all*) to 7 (*extremely*) scale: how much they liked the game, how fun, enjoyable, and cool it was, and how happy they were playing it (hereafter referred to as the enjoyment scale; $\alpha = 0.97$). In the other block, participants reported the number of "specific individual bugs" they had experienced when playing the game, from 0 to 20 (with a 21$^{st}$ option, "More than 20 [please type your number])." Afterwards, participants who reported noticing any

number of bugs greater than zero were asked to provide some examples, via an open-ended essay box. Finally, all participants reported any general confusion with the task and answered a comprehension check in which they identified which version of the game they had played.

**Results**

First, participants again preferred the same stimulus merely when it was labeled as more (vs. less) revised: participants enjoyed playing the same game more when informed they were playing Version 5 relative to Version 2 ($M_{revised}$ = 5.85, SD = 1.29 vs. $M_{control}$ = 5.26, SD = 1.50, *unequal variances*, $t(485.12) = -4.74$, $p < .001$, $d = -0.43$).

Second, participants noticed fewer bugs when informed they were playing Version 5 compared to Version 2 ($M_{revised}$ = 1.75, SD = 4.09 vs. $M_{control}$ = 2.85, SD = 4.13, $t(498) = 2.98$, $p = .003$, $d = 0.27$). Common descriptions of these alleged bugs included perceived speed issues (examples of actual quotes: "It lagged a bit overlaying other stroke patterns"; "It did not move where I wanted it to") and a perceived lack of features (examples of actual quotes: "unable to erase"; "no choice of color, limited control"). Interestingly, an exploratory analysis indicated that significantly more participants noticed zero bugs in the revised condition (65.1%) compared to those in the control condition (40.7%, $\chi^2$ (1, $N$ = 500) = 29.76, $p < .001$, $\phi = 0.24$).

Third, we tested whether the effect of revision label on enjoyment was mediated by number of bugs using PROCESS Model 4 with 5,000 bootstrapped samples, and 95% confidence intervals (Hayes, 2017). As hypothesized, there was a significant indirect effect via number of bugs noticed: $a \times b = 0.10$, $SE = 0.05$, 95% $CI_{Boot}$ [0.03, 0.21].

Study 3 again replicates the basic effect (H1) while also providing insight into process (H2): a revision label increased participants' experienced enjoyment for playing the same game,

because participants became less likely to encode ostensible bugs in the game while playing (i.e., they perceived it to be of higher quality).

## Study 4: Moderation by Evaluability

In Study 4, we tested moderation (H3): If the effect indeed operates via the heuristic logic as proposed, then it should be attenuated when it is easier for people to evaluate stimulus quality. Here, we introduced a financial cost for choosing a revised but inferior option.

### Method

*Participants*. We recruited U.S.-based adults ($N$ = 1008; 47.4% male; $M_{age}$ = 41.53 years, SD = 11.98) through MTurk to complete this study for a fixed payment. Participants additionally read that they could earn a $0.50 bonus based on their study performance (which was true).

*Procedure*. The study was a 2x2 between-participants design in which we manipulated evaluability (easy versus hard) and revision status (as in other studies, for half of the participants, the sets were labeled as revised). All participants evaluated ten product sets, presented one set at a time in randomized order: Plate set; Cup set; Lego set; Nail set; Crayon set; Book set; Spice set; Puzzle pieces; Stamp collection; Playing cards. For each set, they were told how many pieces were supposed to be in it; critically, they also saw an image of each set that showed the number of pieces actually in it (and whether any pieces were missing). By design, we crafted all the stimuli to be of objectively low quality, such that some pieces were always missing.

Between-subjects, we manipulated how easy or hard it was for participants to quickly visually assess this. Specifically, in the *hard* conditions, the sets were large, ranging from 32 to 108 pieces; whereas in the *easy* conditions, the sets were smaller (about a quarter of the size),

ranging from 8 to 27 pieces. For example, in the hard conditions, participants read that the plate set "should have 32 plates," and they saw 30 plates; by contrast, in the easy conditions, participants read that the plate "should have 8 plates," and they saw 7 plates (see examples in SOM-B).

In addition to manipulating evaluability, we also manipulated revision label. For those in the revised conditions, each set bore the label "revised + updated;" while for those in the unrevised conditions, each set bore the label "not-yet-revised + not-yet-updated." Critically, these same labels were used across all conditions; thus, even if there are any incidental effects of the phrasing of the labels, this cannot account for our moderation hypothesis across conditions.

After they viewed each set, we asked participants whether it was complete or incomplete (via forced-choice) and told them that they would receive a 5-cent bonus for each correct answer. This was true; thus, participants could earn up to 50 cents in bonus payment. Because all sets were in fact incomplete, our primary dependent measure—accuracy—was simply the number of sets (out of ten sets total) that each participant marked as incomplete.

After this count-for-pay task, participants predicted their scores (i.e., they indicated how many of the ten sets they had accurately indicated to be complete versus incomplete); answered a comprehension check assessing whether they noticed the product revision label manipulation; and rated the task difficulty (1 = not at all difficult; 7 = very difficult). Finally, participants were informed of their score and received the corresponding performance bonus as we had advertised.

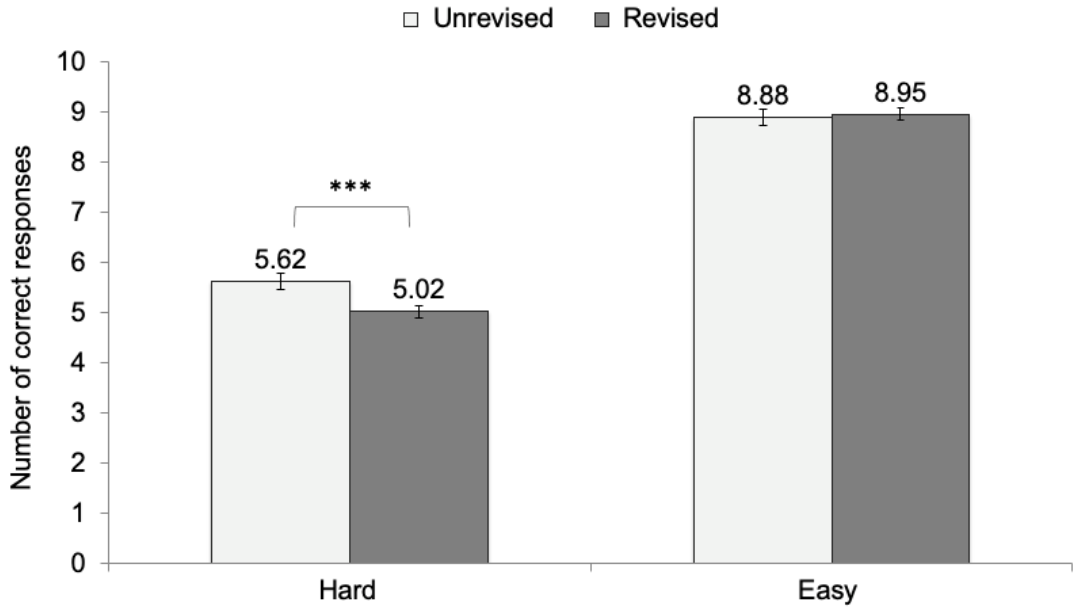
**Results**

*Evaluability.* As intended, participants in the hard conditions reported that the task was more difficult than those in the easy conditions ($M_{hard}$ = 4.78, SD = 1.43 vs. $M_{easy}$ = 2.47, SD =

1.70, $t(976.71) = 23.24$, $p < .001$, $d = 1.47$). Accordingly, participants in the hard condition spent more time evaluating the ten sets than those in the easy conditions ($M_{hard} = 3.66$ minutes, SD = 2.85 vs. $M_{easy} = 2.00$ minutes, SD = 1.01, $t(627.92) = 12.31$, $p < .001$, $d = 0.78$).

*Performance.* Unsurprisingly, there was a main effect of evaluability: Participants in the easy conditions were more accurate than those in the hard conditions and thus earned more bonus money ($M_{hard}$= 5.32, SD = 2.56 vs. $M_{easy} = 8.91$, SD = 1.97; $F(1, 1004) = 629.27$, $p < .001$, $\eta_p^2 = 0.39$). There was also a marginally significant main effect of revision label such that participants in the revised conditions were less accurate than those in the unrevised conditions ($M_{revised} = 6.96$, SD = 3.00 vs. $M_{unrevised} = 7.26$, SD = 2.80; $F(1, 1004) = 3.47$, $p = .063$, $\eta_p^2 < 0.01$). Importantly however, these main effects were qualified by a significant interaction ($F(1, 1004) = 5.52$, $p = .019$, $\eta_p^2 = 0.01$; see Figure 2).

*FIGURE 2: STUDY 4 RESULTS*




*NOTE.*— Error bars indicate ±1 SEM. *$p < .05$, **$p < .01$, ***$p < .001$.

For the hard conditions, and as hypothesized, participants in the revised condition performed worse than those in the unrevised condition ($M_{revised}$ = 5.02, SD = 2.54 vs. $M_{unrevised}$ = 5.62, SD = 2.54; $F(1, 1004) = 8.88$, $p = .003$, $\eta_p^2 = 0.01$); this difference translated into an 11% reduction in bonus money earned. However, this effect was eliminated among participants in the easy conditions, who performed similarly well regardless of labeling ($M_{revised}$ = 8.95, $SD$ = 1.95 vs. $M_{unrevised}$ = 8.88, SD = 1.99; $F(1, 1004) = 0.12$, $p = .731$; $\eta_p^2 < 0.01$).

We also explored differences in overconfidence. We computed overconfidence scores by subtracting each participant's actual score from their predicted score. Thus, positive scores indicate overconfidence; negative scores indicate underconfidence. There was a main effect of both evaluability ($F(1, 1004) = 30.12$, $p < .001$, $\eta_p^2 = 0.03$) and revision label ($F(1, 1004) = 4.56$, $p = .033$, $\eta_p^2 = 0.01$). Importantly however, and mirroring the performance results, these main effects were qualified by a significant interaction ($F(1, 1004) = 4.69$, $p = .031$, $\eta_p^2 = 0.01$). For the hard conditions, participants in the revised condition—who performed worse than those in the unrevised condition—were more overconfident ($M_{revised}$ = 0.84, SD = 3.00 vs. $M_{unrevised}$= 0.15, SD = 2.77; $F(1, 1004) = 9.25$, $p = .002$, $\eta_p^2 = 0.01$). However, within the easy conditions, there were no differences in overconfidence ($M_{revised}$ = -0.39, SD = 2.18 vs. $M_{unrevised}$ = -0.39, SD = 2.14; $F(1, 1004) = 0.00$, $p = .983$, $\eta_p^2 < 0.01$).

Study 4 finds evidence for moderation (H3), here in a context of material consequence. Participants' preference for revised stimuli led them to earn less money in the study—but they were less prone to this effect when stimulus quality was easier to evaluate (despite the label).

## Study 5: Moderated Mediation

Finally, in Study 5, we again tested moderation (H3). We did so in a context whereby participants evaluated revised stimuli that have truly been made better (rather than made worse).

**Method**

*Participants*. We recruited U.S.-based adults ($N$ = 1,003; 49.2% male; $M_{age}$ = 43.04 years, $SD$ = 12.10) from MTurk to complete a study for a fixed payment.

*Procedure.* The study was a 2x2 between-participants design in which we manipulated evaluability (easy vs. hard) and revision label (absent vs. present). All participants imagined they had spilled a drink on their work desk and had two options of similar paper towels to clean the mess, option A and option B. Between-subjects, we manipulated option B along two dimensions: evaluability (hard-to-evaluate vs. easy-to-evaluate) and revision label (no-revision vs. yes-revision). Thus, participants were randomly assigned to one of four possible conditions, where the only difference between option A and option B was in an extra phrase we provided with option B: No extra phrase (hard-to-evaluate, no-revision), "thicker and more absorbent" (easy-to-evaluate, no-revision), "revised" (hard-to-evaluate, yes-revision), and "revised – thicker and more absorbent" (easy-to-evaluate, yes-revised; see stimuli in SOM-B). In randomized order, participants rated their preference (1 = "I would definitely choose Option A"; 4 = "neutral preference"; 7 = "I would definitely choose Option B"), as well as their quality inferences (1 = "Option A is of higher quality"; 4 = "similar quality"; 7 = "Option B is of higher quality").

**Results**

*Preference.* A 2x2 ANOVA revealed no main effect of revision label ($F$(1, 999) = 2.14, $p$ = .144, $\eta_p^2 < 0.01$), and a significant effect main effect of evaluability ($F$(1, 999) = 366.72, $p$ <

.001, $\eta_p^2 = 0.27$). As hypothesized, however, this effect was qualified by a significant interaction ($F(1, 999) = 5.67$, $p = .017$, $\eta_p^2 = 0.01$). When evaluability was hard, we indeed replicated the basic effect: Participants preferred option B more when it was labeled as revised ($M_{\text{yes-revised}} = 3.71$, SD = 1.81 vs. $M_{\text{no-revised}} = 3.31$, SD = 1.61; $F(1, 999) = 7.41$, $p = .007$, $\eta_p^2 = 0.01$). Yet when evaluability was easy, the revision label no longer wielded an effect: Participants preferred option B to a similarly high degree regardless of whether or not it was labeled as revised ($M_{\text{yes-revised}} = 5.43$, SD = 1.58 vs. $M_{\text{no-revised}} = 5.52$, SD = 1.48; $F(1, 999) = 0.42$, $p = .517$, $\eta_p^2 < 0.01$).

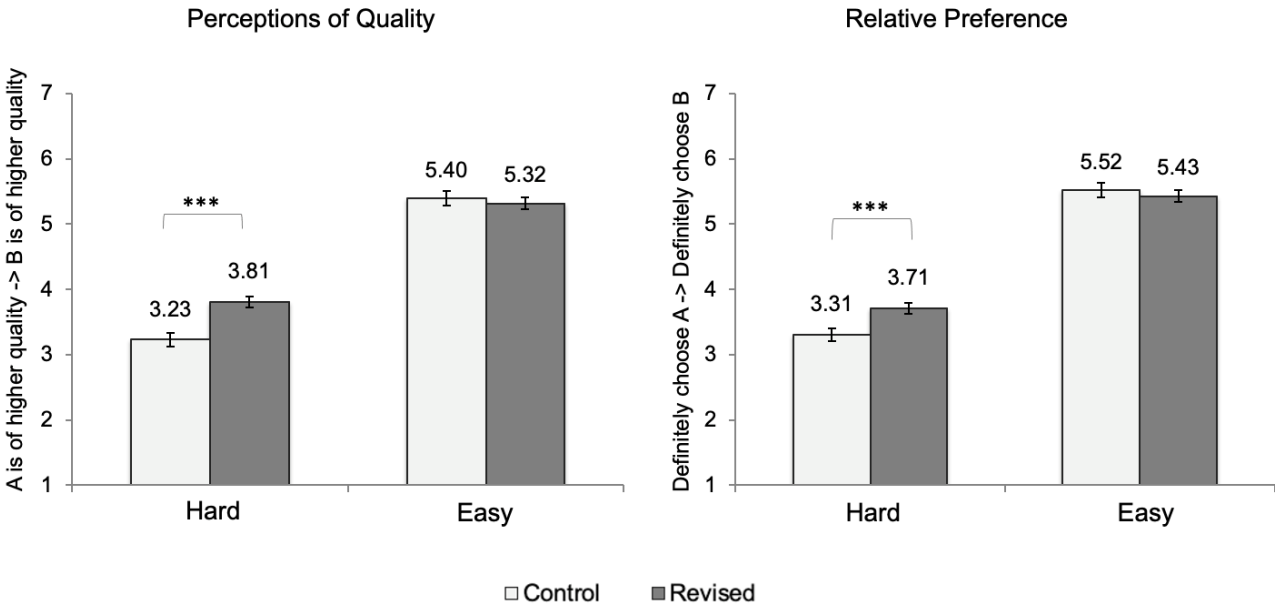
*Quality inferences.* A 2x2 ANOVA revealed a significant main effect of evaluability ($F(1, 999) = 350.53$, $p < .001$, $\eta_p^2 = 0.26$), and a significant main effect of revision label ($F(1, 999) = 6.44$, $p = .011$, $\eta_p^2 = 0.01$). Again, however, there was the critical interaction ($F(1, 999) = 11.22$, $p < .001$, $\eta_p^2 = 0.01$). When evaluability was hard, we again replicated the basic effect: Participants perceived option B as higher quality when it was labeled as revised ($M_{\text{yes-revised}} = 3.81$, SD = 1.78 vs. $M_{\text{no-revised}} = 3.23$, SD = 1.58; $F(1, 999) = 17.35$, $p < .001$, $\eta_p^2 = 0.02$). Yet when evaluability was easy, participants perceived option B as similarly high quality regardless of whether or not it was labeled as revised ($M_{\text{revised}} = 5.32$, SD = 1.43 vs. $M_{\text{control}} = 5.40$, SD = 1.38; $F(1, 999) = 0.33$, $p = .567$, $\eta_p^2 < 0.01$; see Figure 3).

*Moderated mediation.* Providing a further test of mechanism, we tested for moderated mediation using model 8 of PROCESS Macro (Hayes 2017), 5000 Bootstrapped samples, and 95% confidence interval. The index of moderated mediation was significant (95% $CI_{\text{Boot}}$ = [-0.93, -0.24]) such that the indirect effect of revision label on preference, via quality inferences, was significant when evaluability was hard ($a \times b = 0.51$, SE = 0.14, 95% $CI_{\text{Boot}}$ = [0.24, 0.78]) but *not* when evaluability was easy ($a \times b = -0.07$, SE = 0.11, 95% $CI_{\text{Boot}}$ = [-0.29, 0.15]).

Study 5 finds further evidence for moderation (H3), while also providing further process

insights: Participants were more likely to infer stimulus quality from the revision label when they had no other information regarding stimulus quality, consistent with a heuristics-based account.

*FIGURE 3: STUDY 5 RESULTS*




*NOTE*.—Error bars indicate ±1 SEM. *$p$ < .05, **$p$ < .01, ***$p$ < .001.

## General Discussion

To "revise and resubmit" is far more than an academic exercise; indeed, people regularly encounter revised stimuli across everyday life. In principle, a world of constant revision should benefit people by affording them the most up-to-date offerings. In practice, however, the current research reveals a potential cost. The results from six studies suggest people cannot easily tell the difference between genuine improvement and mere superficial change. Instead, people appear to uniformly assume revised stimuli are better stimuli—even when nothing has actually improved. This preference for revision is surely often justified, at least on average. People likely learn from everyday life that most revision efforts are made with the intention to make things better, not

worse—making people right to assume as much in many contexts. But as is the case with all heuristics, drawing on them can lead people astray in critical cases—cases in which an offering has allegedly been revised but was not improved in reality.

This effect thus raises practical implications. Things that have been unchanged or degraded during the revision process may nonetheless be adopted, so long as people *believe* they possess a "revised" version. This may happen innocently (e.g., co-authors may be prone to accepting the revised work of whoever leaves off, independent of actual progress), but also intentionally (e.g., companies that release annual updates simply to release annual updates). We found that people exhibit a preference for revision even after they directly interact with the allegedly revised stimulus (e.g., playing a game in Study 3), and even when holding constant other factors that may lead people to prefer the revision, such as mere chronological newness (Jie & Li, 2022).

If people cannot easily trust their bottom-up experience to draw more informed conclusions, then they could be influenced by creators who are motivated to make revisions for reasons beyond the desire to make things better. We suspect such cases of unfaithful revision efforts are at least somewhat common in everyday life—meaning that people may be at risk of costly deception (Bulow, 1986; Gershoff et al., 2012). Indeed, almost 50% of working professionals we surveyed (including marketing and brand managers), indicated that they have worked for a company that launched a revised product that was *not* an improvement (see manager study details in SOM-D).

By the same token, the current research suggests things that are improved in the revision process may go underappreciated if the revisions are not explicitly advertised. Consider that the final version of a person's resume or a restaurant's menu almost certainly underwent extensive

tweaking behind the scenes—yet consumers of such offerings may not appreciate this fact on their own, absent a label. Faithfully providing revision labels (as in our moderation studies) may help avoid this problem, while also protecting people from costly deception.

The current research also makes theoretical contributions. We identify a novel heuristic—people's reliance on revision labels—which may be especially timely in today's ever-changing marketplace as it suggests people may get caught seeking ever-newer offerings that are no better in reality. Our studies extend research on the link between expectations and experience, addressing an issue raised by Lee et al. (2006) in their review of outstanding questions on such links: "A third question concerns how specific perceptual, attentional, and cognitive mechanisms mediate the effect of expectations on experience (or reported experience)" (p. 1057). We find evidence for *how* expectations may change one's experience, whereby revision labels led participants to essentially pay less attention and notice fewer problems in a flawed stimulus.

An overgeneralized preference for revision also speaks to bridging an emerging collection of related phenomena that are susceptible to sequencing effects. For example, building on the tendency for the first item in an array to be preferred (Bulow, 1986; Gershoff et al., 2012), research has shown that "phantom firsts"—merely framing something as "first"—increases its appeal (Leboeuf et al., 2014). Interestingly, these phenomena suggest that in the absence of explicit revision, stimuli that are framed to have occurred earlier in a sequence are preferred (Smith et al., 2016). At the same time, other research points to the notion that items that have been in existence for longer are preferred, as with the "longevity bias" (Eidelman et al., 2010).

While our findings cannot be explained by such effects (e.g., because we conducted studies that held chronological newness constant), they invite exciting avenues for studying when and why different temporal sequences are preferred. For example, future research is

needed to examine when people desire older (or "pre-revised") experiences—as when experiencing nostalgia (Wildschut et al., 2006), rediscovering the joy of past experiences (O'Brien, 2019), or valuing original renditions of collectibles over later renditions (Newman & Bloom, 2012). Our framework predicts people may dislike revisions that they can blatantly see are worse, because they no longer need to recruit heuristic help (as perhaps was the case in, for example, the backfiring of "New Coke": Klein, 2015).

Finally, all our participants were U.S-based adults, leaving avenues for studying cultural differences. For example, because people from more collectivistic cultures tend to be more open to the possibility of a current change reverting to its original (or an even worse, yet-unknown) state (Ji et al., 2001), perhaps such participants are less trusting of revision labels.

In sum, the current research highlights dual implications for today's endless revisions wielding influence when perhaps they should not and failing to wield influence when perhaps they should. Revised stimuli are often, but not always, better stimuli—despite beliefs otherwise.

## References


Bulow, J. (1986). An Economic Theory of Planned Obsolescence. *The Quarterly Journal of Economics*, *101*(4), 729–749. https://doi.org/10.2307/1884176

Carney, D. R., & Banaji, M. R. (2012). First Is Best. *PLOS ONE*, *7*(6), e35088. https://doi.org/10.1371/journal.pone.0035088

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J. : L. Erlbaum Associates. http://archive.org/details/statisticalpower0000cohe_j0l3

Eidelman, S., Pattershall, J., & Crandall, C. S. (2010). Longer is Better. *Journal of Experimental Social Psychology*, *46*(6), 993–998. https://doi.org/10.1016/j.jesp.2010.07.008

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Gershoff, A. D., Kivetz, R., & Keinan, A. (2012). Consumer Response to Versioning: How Brands' Production Methods Affect Perceptions of Unfairness. *Journal of Consumer Research*, *39*(2), 382–398. https://doi.org/10.1086/663777

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gneezy, A., Gneezy, U., & Lauga, D. O. (2014). A Reference-Dependent Model of the Price–Quality Heuristic. *Journal of Marketing Research*, *51*(2), 153–164. https://doi.org/10.1509/jmr.12.0407

Hagen, L. (2021). Pretty Healthy Food: How and When Aesthetics Enhance Perceived Healthiness. *Journal of Marketing*, *85*(2), 129–145. https://doi.org/10.1177/0022242920944384

Haws, K. L., Reczek, R. W., & Sample, K. L. (2017). Healthy Diets Make Empty Wallets: The Healthy = Expensive Intuition. *Journal of Consumer Research*, *43*(6), 992–1007. https://doi.org/10.1093/jcr/ucw078

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach*. Guilford Publications.

Ji, L.-J., Nisbett, R. E., & Su, Y. (2001). Culture, Change, and Prediction. *Psychological Science*, *12*(6), 450–456. https://doi.org/10.1111/1467-9280.00384

Jie, Y., & Li, Y. (2022). Chronological Cues and Consumers' Preference for Mere Newness. *Journal of Retailing*, *98*(3), 527–541. https://doi.org/10.1016/j.jretai.2021.11.003

Klein, C. (2015). *Why Coca-Cola's 'New Coke' Flopped*. https://www.history.com/articles/why-coca-cola-new-coke-flopped

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The Effort Heuristic. *Journal of Experimental Social Psychology*, *40*(1), 91–98. https://doi.org/10.1016/S0022-1031(03)00065-9

Leboeuf, R. A., Williams, E. F., & Brenner, L. A. (2014). Forceful Phantom Firsts: Framing Experiences as Firsts Amplifies Their Influence on Judgment. *Journal of Marketing Research*, *51*(4), 420–432. https://doi.org/10.1509/jmr.11.0298

Lee, L., Frederick, S., & Ariely, D. (2006). Try It, You'll Like It: The Influence of Expectation, Consumption, and Revelation on Preferences for Beer. *Psychological Science*, *17*(12), 1054–1058. https://doi.org/10.1111/j.1467-9280.2006.01829.x

Newman, G. E., & Bloom, P. (2012). Art and Authenticity: The Importance of Originals in Judgments of Value. *Journal of Experimental Psychology: General*, *141*(3), 558–569. https://doi.org/10.1037/a0026035

O'Brien, E. (2019). Enjoy It Again: Repeat Experiences Are Less Repetitive Than People Think. *Journal of Personality and Social Psychology*, *116*(4), 519–540. https://doi.org/10.1037/pspa0000147

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive Strategy Selection in Decision Making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral Decision Research: A Constructive Processing Perspective. *Annual Review of Psychology*, *43*(1), 87–131.

Perrigo, B. (2017). *Jimmy Kimmel Pranks Clueless People Into Thinking the iPhone 4 Is the iPhone X*. https://time.com/5014899/jimmy-kimmel-prank-iphone-4-iphone-x/

Rao, A. R., & Monroe, K. B. (1989). The Effect of Price, Brand Name, and Store Name on Buyers' Perceptions of Product Quality: An Integrative Review. *Journal of Marketing Research*, *26*(3), 351–357. https://doi.org/10.1177/002224378902600309

Smith, R. K., Newman, G. E., & Dhar, R. (2016). Closer to the Creator: Temporal Contagion Explains the Preference for Earlier Serial Numbers. *Journal of Consumer Research*, *42*(5), 653–668. https://doi.org/10.1093/jcr/ucv054

Tversky, A. (1972). Elimination by aspects: A Theory of Choice. *Psychological Review*, *79*(4), 281–299. https://doi.org/10.1037/h0032955

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wildschut, T., Sedikides, C., Arndt, J., & Routledge, C. (2006). Nostalgia: Content, Triggers, Functions. *Journal of Personality and Social Psychology*, *91*(5), 975–993. https://doi.org/10.1037/0022-3514.91.5.975

Yeung, C. W. M., & Soman, D. (2007). The Duration Heuristic. *Journal of Consumer Research*, *34*(3), 315–326. https://doi.org/10.1086/519500

# SUPPLEMENTAL ONLINE MATERIALS

A Preference for Revision Absent Improvement

## A. STUDIES 1-4: ROBUSTNESS CHECKS

### Study 1C (resumes)

This study was included in two bundles of unrelated studies: the first bundle was administered in December 2016 and the second one in January 2017. Because it was possible for people to participate in both bundles, we reran our analysis excluding 80 data points from the second bundle that correspond to individuals who participated in the first bundle. The results reported in the main text hold when excluding individuals who participated in both bundles: a 2x2 mixed ANOVA revealed a significant effect of revision label ($F(1, 371) = 41.19$, $p < .001$).

### Study 2A (Selfie sticks)

The results reported in the main text hold when excluding participants who answered incorrectly the comprehension check (6.5% of the sample). A greater proportion of participants chose the *worse* selfie stick over the objectively better option merely when the worse one was labeled as revised (chosen by 25.6% of these participants) vs. when it had no such label (chosen by 13.1% of these participants, $\chi^2(1, N = 563) = 14.09$, $p < .001$, $\phi = 0.16$).

### Study 2B (Restaurant guidebooks)

The results reported in the main text hold when excluding participants who answered incorrectly the comprehension check (2.2% of the sample). More participants in the revised condition preferred the useless guidebook (43.0%) than participants in the control condition (22.4%; $\chi2(1, N = 392) = 18.84$, $p < .001$, $\phi = 0.22$).

### Study 3 (Video game)

The results reported in the main text hold when excluding participants who answered incorrectly the comprehension check (5.2% of the sample). Participants enjoyed their experience significantly more when they read they were playing Version 5 relative to Version 2 ($M_{revised} = 5.87$, SD = 1.28 vs. $M_{control} = 5.25$, SD = 1.52, $t(472) = 4.85$, $p < .001$, $d = 0.44$). Also, participants noticed fewer bugs when they read they were playing Version 5 compared to Version 2 ($M_{revised} = 1.53$, SD = 3.68 vs. $M_{control} = 2.67$, SD = 3.86, $t(472) = -3.29$, $p = .001$, $d = -0.30$). Finally, the indirect effect of version on enjoyment through number of bugs also remained significant: $a \times b = 0.12$, $SE = 0.05$, 95% $CI_{Boot}$ [0.04, 0.24].

### Study 4 (Product sets)

The interaction effect between revision label and evaluability on accuracy in the count-for-pay task remained significant when excluding participants who answered incorrectly the comprehension check (4.9% of the sample): $F(1, 955) = 4.15$, $p = .042$, $\eta_p^2 < .001$.

## SB. STUDIES 2-6: IMAGE STIMULI

FIGURE W1. STUDY 2A: SELFIE STICKS

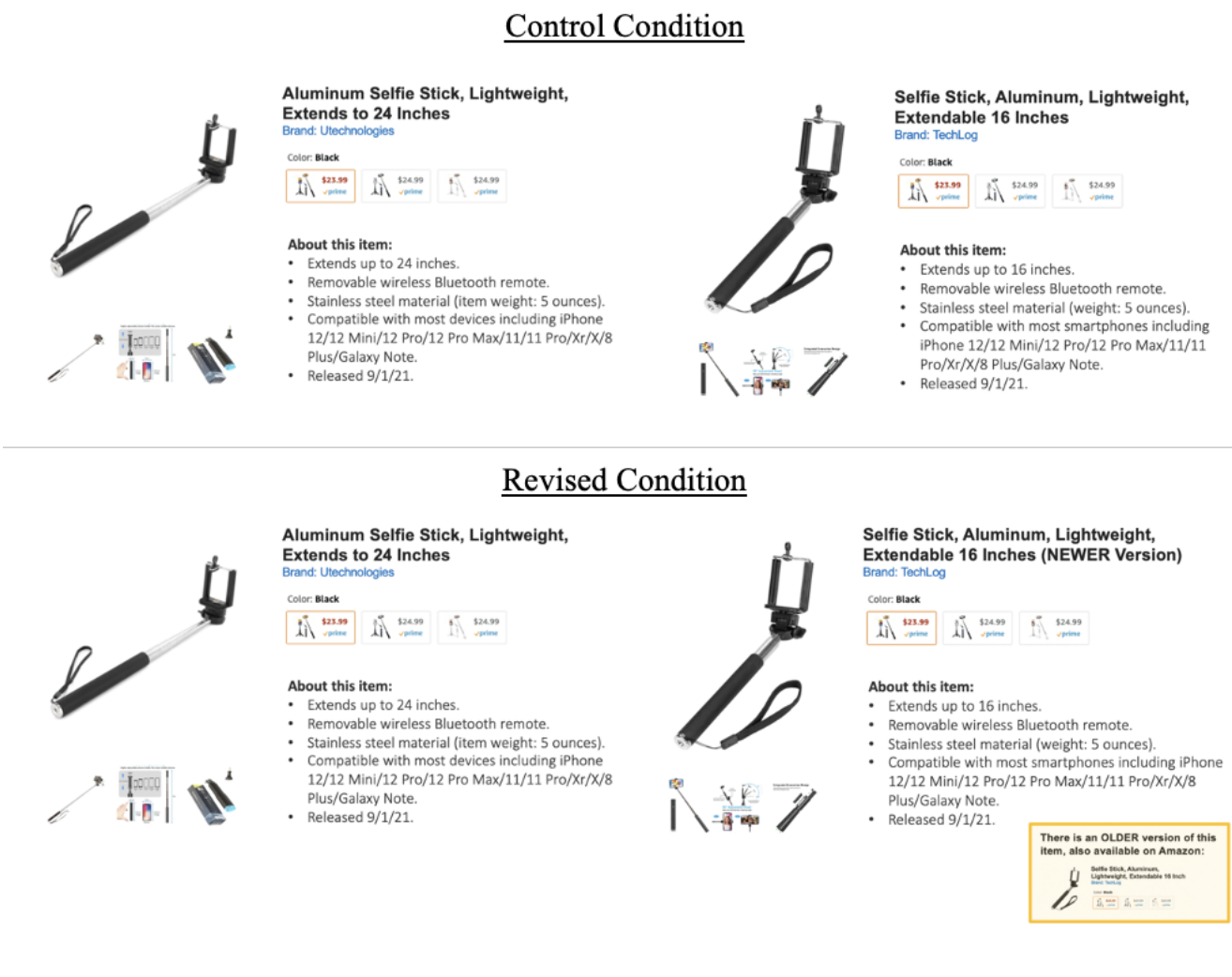


FIGURE W2. STUDY 2B: RESTAURANT GUIDEBOOKS

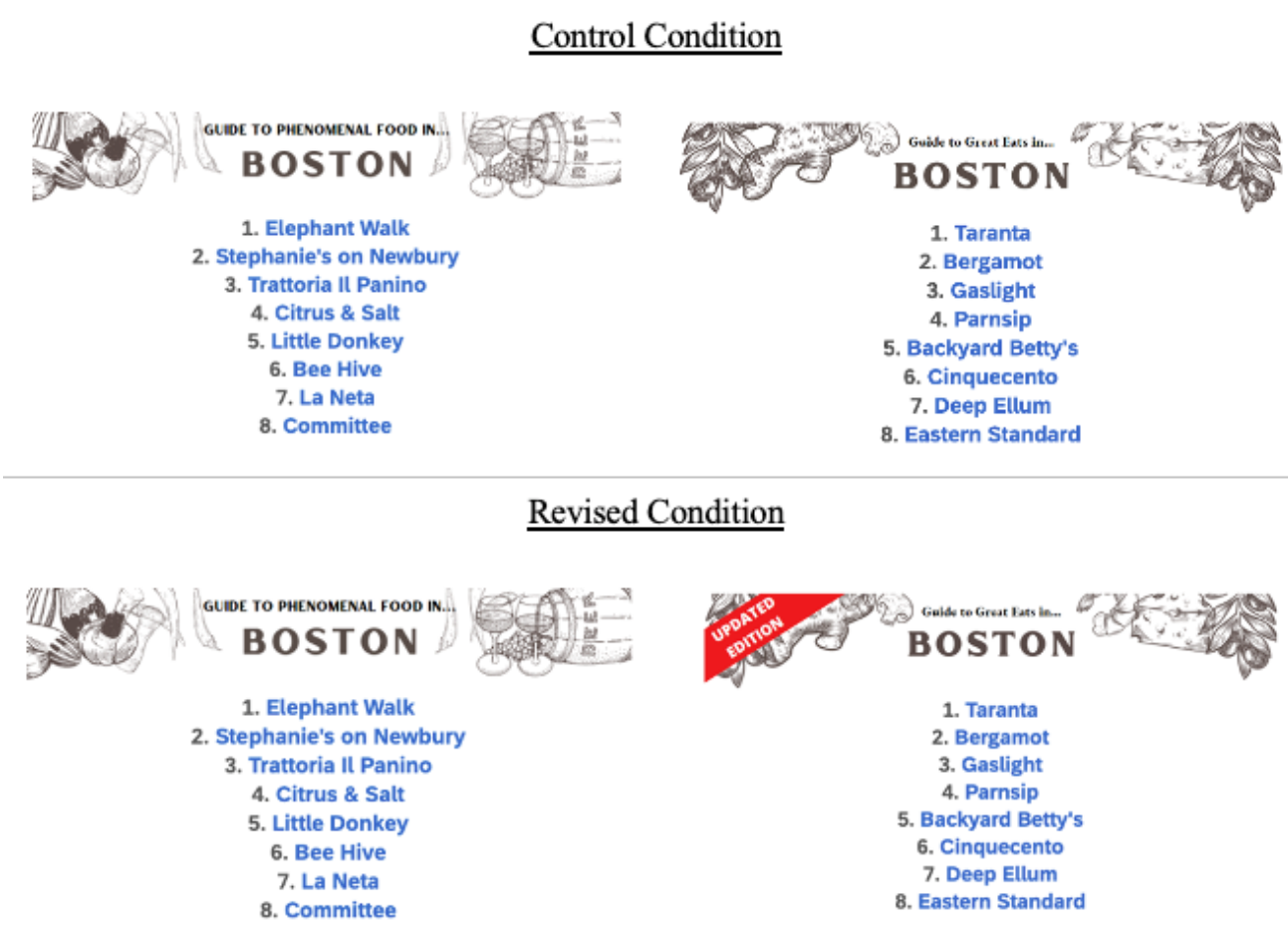


FIGURE W3. STUDY 2B: EXAMPLE OF WHAT PARTICIPANTS SAW IF THEY CLICKED ON A LINK FROM THE GUIDEBOOK FEATURING OPEN RESTAURANTS

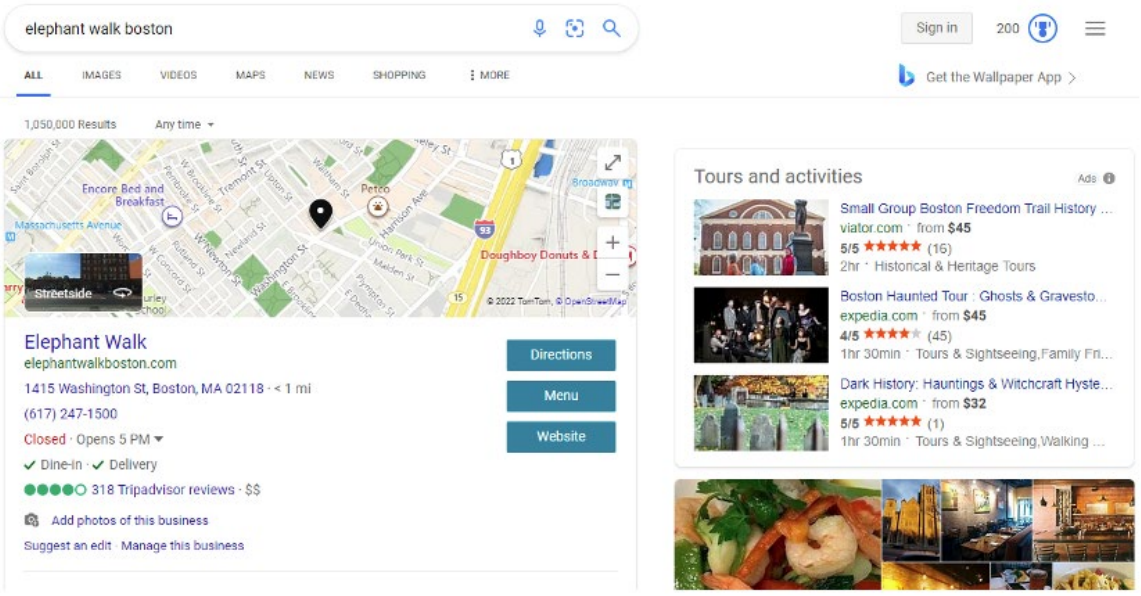



FIGURE W4. STUDY 2B: EXAMPLE OF WHAT PARTICIPANTS SAW IF THEY CLICKED ON A LINK FROM THE GUIDEBOOK FEATURING CLOSED RESTAURANTS

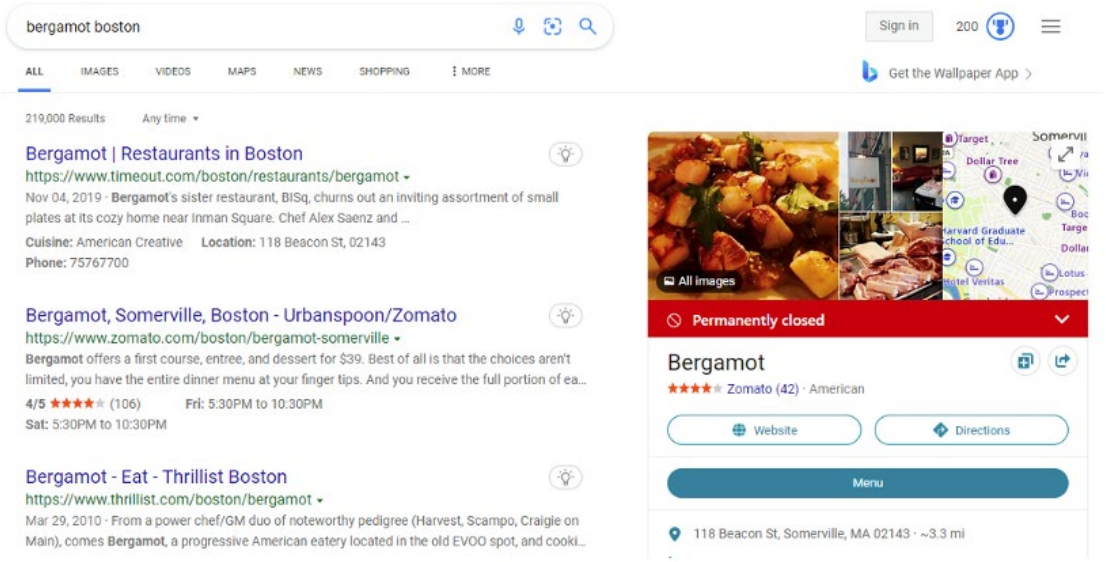


FIGURE W5. STUDY 3: VIDEO GAME

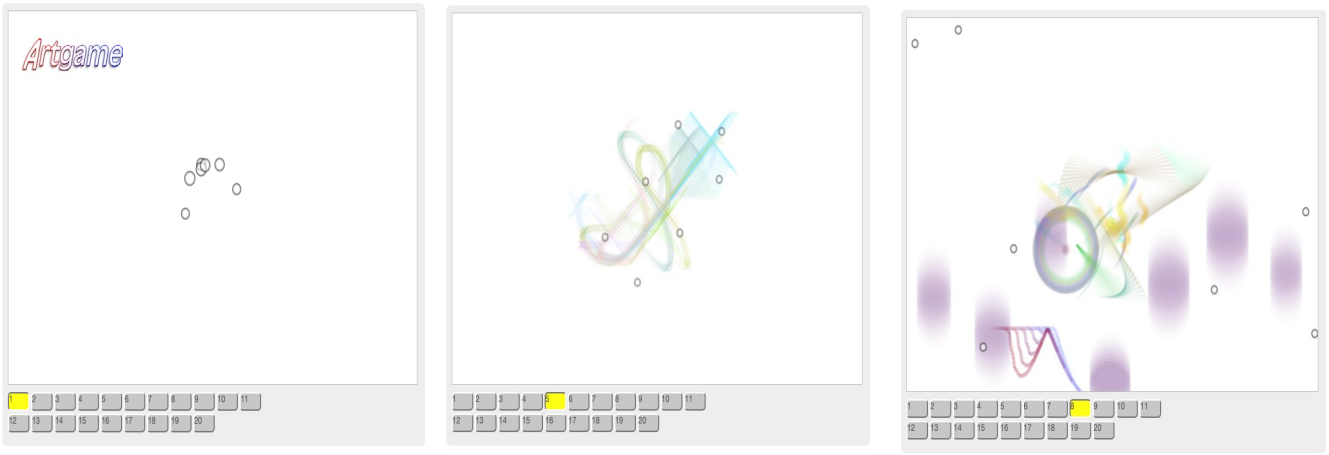
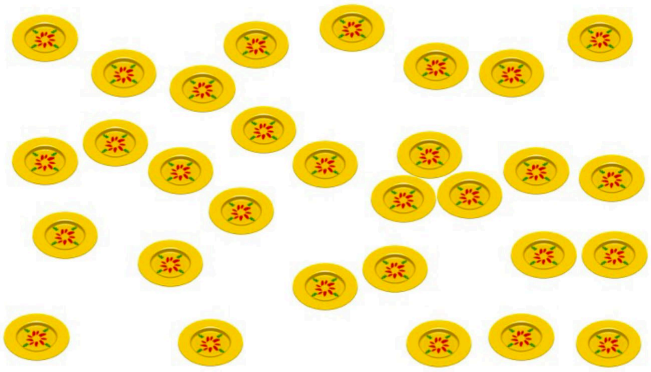


FIGURE W6. STUDY 4: EXAMPLES OF PRODUCT SETS

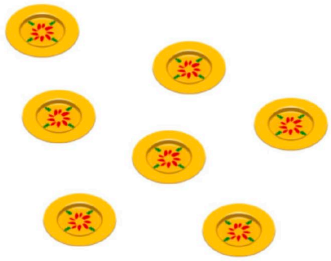
**HARD**

This plate set should have 32 plates:



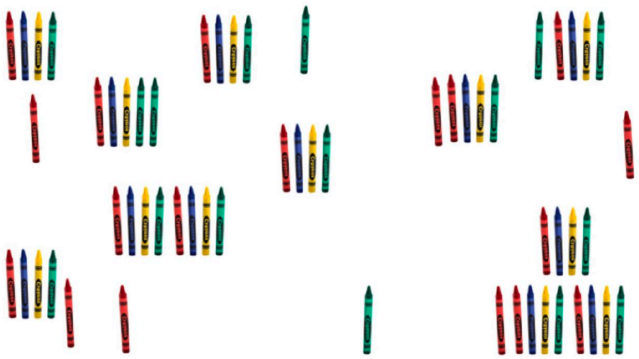
**EASY**

This plate set should have 8 plates:



**HARD**

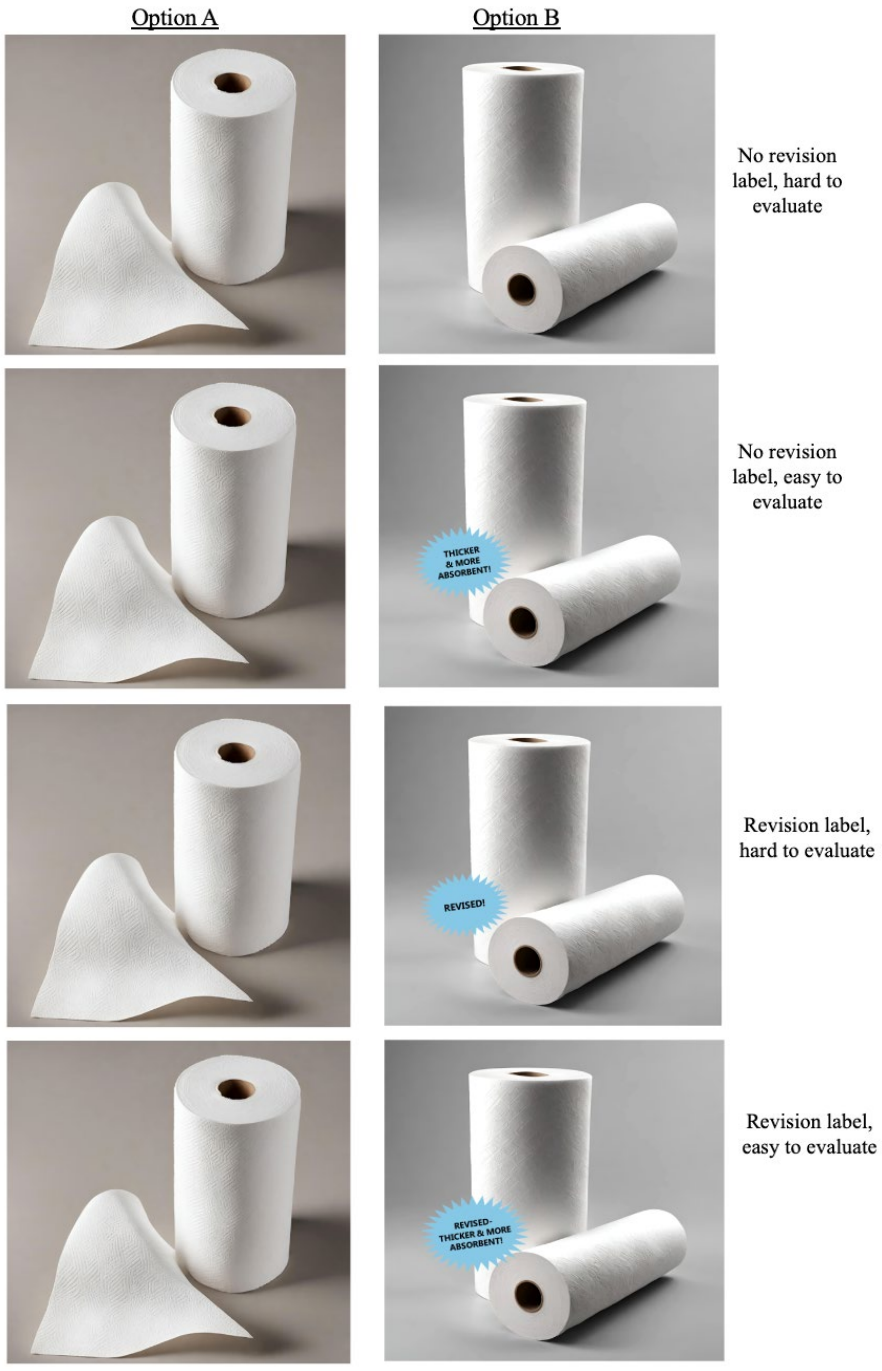
This crayon set should have 60 crayons:



**EASY**

This crayon set should have 15 crayons:

FIGURE W7. STUDY 5: PAPER TOWELS

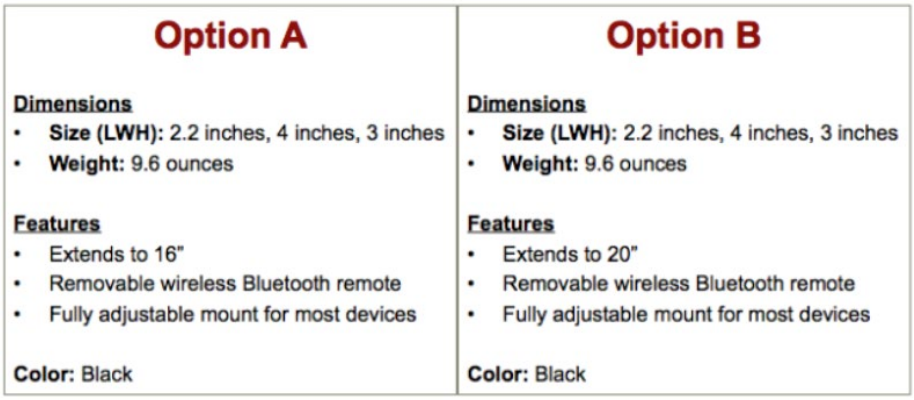


## C. STUDY 2A: SELFIE-STICK PRETEST

The goal of a selfie stick is to help users take pictures of themselves from a distance, so their *length* is central to their quality: All else equal, longer selfie sticks are better than shorter ones. To confirm this, we first conducted a pretest that assessed whether consumers view longer selfie sticks as objectively better than shorter ones.

Participants ($N$ = 403 MTurkers; 56.3% male; $M_{age}$ = 34.41 years, SD = 10.41) were shown two selfie sticks. The selfie sticks were identical in terms of color, weight, etc. and only varied in length: one option was 16-inches long and the other one was longer (see Figure W8). Between-subjects, we manipulated the length of the longer stick to be either 20-inch, 24-inch, 28-inch, or 32-inch. Participants rated the overall quality of the two selfie sticks—the 16-inch stick and the longer one—on a 7-point scale from "very low" to "very high." Consistent with the pretest, longer selfie sticks (regardless of the actual length) were deemed higher quality compared to a shorter 16-inch one (all $p$s < .001).

FIGURE W8. PRETEST STIMULI EXAMPLE



## D. MANAGER SURVEY

We surveyed a sample of individuals with professional work experience (including brand managers and marketers) and asked them to indicate firms' motivations for releasing revised versions of a product; and whether those revisions are *always* improvements over prior versions.

### Method

*Participants.* We recruited MBA students ($N$ = 126; 56.3% male; $M_{age}$ = 28.60 years, SD = 2.46) from a Midwestern US university to complete an online survey, which was one of several studies in a battery of unrelated studies. This survey was sent to 130 students enrolled in an MBA elective class; despite participation being optional and that no compensation was provided, only four students did not take the study. On average, participants had 5.13 years of professional experience (SD = 1.92). In addition, 25.0% of our sample reported having previous experience in marketing, branding, or product management, with an average of 2.69 years of professional experience in marketing.

*Procedure.* First, we assessed whether participants thought that, generally speaking, revised products are improvements over their predecessors. Participants indicated the extent to which they agreed with the statement: "Every revision of a product that companies launch is an improvement" by selecting an option on a 7-point scale from "Strongly disagree" to "Strongly agree". Next, participants were asked: "Why do you think companies release revised versions of products?" Here, participants rated five reasons on 7-point scales with endpoints labeled as "1 = Not at all" and "7 = Very much": (1) to improve the product, (2) to generate word-of-mouth, (3) to increase sales, (4) to renew the brand image, and (5) to grow their portfolio; all options were presented in random order. Immediately below, participants had the option of describing other reasons not listed (open-ended question).

In the second section of the survey, we assessed participants' perceptions of specific products *released by the companies they had worked for*. Participants first answered the question "Have you/the companies you worked for ever released a revised version of a product that is/was not really an improvement over the previous version?" by selecting a yes or no response option. Those who selected "yes" were then asked to describe a situation in which their company had released a revised version of a product that was not really an improvement over the previous version (open-ended question); those who selected "no" moved on to the next question. Next, all participants indicated what percentage of the company's revised products were not really an improvement over the previous version, on an 11-point scale from 0% to 100%.

In the third section, we assessed participants' perceptions of specific products *released by other companies*. Similar to the previous section, participants first answered the question "Can you think of any revised product that is/was not really an improvement over its previous version? This can be any product on the market" by selecting a yes or no response option. Those who selected "yes" were asked to describe the product and why it did not represent an improvement over its previous version (i.e., they answered two open-ended questions); those who selected "no" skipped those two questions. Finally, similar to the previous section, all participants indicated what percentage of the revised versions of products in the marketplace they thought did not represent an improvement over a previous version, on an 11-point scale from 0% to 100%.

The final section included standard demographics: participants reported how many years of work experience they had before starting the MBA program and if they ever worked in marketing, branding, or product management; those indicating that they had marketing

experience indicated how many years they worked in a marketing role, as well as the title of their last position in marketing. We obtained participants' gender and age from a different survey which participants had taken at the beginning of the semester.

**Results**

*Generally speaking, do working professionals think every product revision is an improvement?* On average, participants slightly disagreed with the statement "Every revision of a product that companies launch is an improvement" ($M$ = 3.21, SD = 1.35; on a 1-7 scale—significantly below the scale midpoint, $t(125) = -6.54$, $p < .001$). These results hold when we restrict the sample to professionals with marketing experience ($n$ = 32), who also disagree with the statement that every revision is an improvement ($M$ = 2.91, SD = 1.23; on a 1-7 scale—significantly below the scale midpoint, $t(125) = -5.04$, $p < .001$).

*Why do companies revise their products?* Here, we provided five possible reasons for why companies might revise their products. Notably, the top scoring reason was "to increase sales" ($M$ = 6.05, SD = 0.99), rated above "to improve the product" ($M$ = 5.39, SD = 1.06; $t(125) = -5.81$, $p < .001$). However, attesting to how companies do sometimes seek to improve their products through revisions, "to improve the product" was the second-highest rated response; it was followed by "to renew their brand image" ($M$ = 5.07, SD = 1.19), "to grow their portfolio" ($M$ = 4.74, $SD$ = 1.32), and "to generate word-of-mouth" ($M$ = 4.49, SD = 1.42). We find that 65.1% of the participants listed other reasons such as "to reduce costs" (12.7%), "to stay competitive" (11.9%) and "to fix bugs" (6.3%).

*Are specific products – launched by the firms that participants worked for – improvements over their predecessors?* Almost half of participants (46.0%) indicated that they have worked for a company that launched a revised product that was not an improvement over its predecessor. On average, participants indicated that 28.7% (SD = 24.49) of the revised products released by a company they worked for in the past did not represent improvements over previous versions. These results were more striking when restricting the analyses to participants who had experience in marketing, branding, product management, or a related area ($n$ = 32): 75.0% of them indicated that they have worked for a company that launched a revised product that was not an improvement over its predecessor; consistently, these participants indicated that 36.6% (SD = 29.14) of the revised products released by a company they worked for in the past did not represent improvements over previous versions.

*Are specific products – launched by other companies – improvements over their predecessors?* Almost all participants (92.9%) gave an example of a product that was not actually an improvement over its predecessor. On average, participants indicated that 36.2% (SD = 18.97) of the revised products in the marketplace do not represent improvements over previous versions.

## E. NOTE ABOUT BLANK OR UNFINISHED RESPONSES

**Study 1A.** Six participants did not answer the question "What percentage of the final version of your resume is different from your original resume?" This explains why some analyses reported in the main text include a smaller sample size ($N = 38$ as opposed to $N = 44$).

**Study 1B.** N/A

**Study 1C.** The original merged dataset included 457 observations. However, we identified four blank responses (i.e., four observations did not have ratings of an original and revised resume). We note that we are not sure if these observations correspond to a research assistant testing the studies or to actual participants who skipped the two questions asking them to rate the original and revised resumes. Therefore, the final dataset of Study 1C included 453 participants.

**Studies 2-5.** We utilized only complete responses (defined as "Finished = 1" in Qualtrics) and deleted unfinished responses (defined as "Finished = 0" in Qualtrics) before data was analyzed. The raw data posted in a Research Box folder includes unfinished responses; the processed data only includes finished ones.