Research Article

Psychological Science
2022, Vol. 33(8) 1278–1299

DOI: 10.1177/09567976221075302
www.psychologicalscience.org/PS

Losing Sight of Piecemeal Progress: People Lump and Dismiss Improvement Efforts That Fall Short of Categorical Change—Despite Improving

Ed O'Brien
Booth School of Business, The University of Chicago

Abstract

Fourteen experiments (N = 10,556 adult participants, including more than 20,000 observed choices across 25 issues) documented how people perceive and respond to relative progress out in the world, revealing a robust "negative-lumping" effect. As problematic entities worked to better their ways, participants shifted to dismiss them if they fell short of categorical reform—despite distinctions in improvement. This increased dismissal of relative gains as "all the same" was driven by the belief that falling short signals an eschewal of doing the bare minimum and lacking serious intent to change, making these gains seem less deserving of recognition. Critically, participants then "checked out": They underrewarded and underinvested in efforts toward "merely" incremental improvement. Finally, in all experiments, participants lumped together absolute failures but not absolute successes, highlighting a unique blindness to gradations of badness. When attempts to eradicate a problem fail, people might dismiss smaller but critical steps that were and can still be made.

Keywords

progress, change, tipping points, valence, improvement, comparison, open data, open materials, preregistration

Received 4/18/21; Revision accepted 12/26/21

> Faith, as you say, there's small choice in rotten apples.
>
> —William Shakespeare, *The Taming of the Shrew* (1593/2003, Act 1, Scene 1, Lines 136–137)

The world is filled with problems needing fixing. From alarming rates of climate change (Cook et al., 2016) to systemic disparities across race, gender, sexuality, and class (Ferguson, 2020), we face an urgent need of improvement efforts that might solve such problems for good.

When working for change, like working toward any goal, people often establish dividing lines that help keep track of progress—lines that often demarcate success and failure. Calls to combat climate change, for example, refer to a 1.5 °C "point of no return" (Aengenheyster et al., 2018)—the proposal that when carbon emissions raise Earth's average temperature by 1.5 °C, humanity will have irrevocably missed its window to prevent catastrophe. More broadly, people often set goals with, and perceive goal pursuit as containing, these kinds of tipping points of categorical change (Klein & O'Brien, 2016, 2018; O'Brien, 2020; O'Brien & Klein, 2017): People are prone to believing that losing X pounds or saving X dollars will officially bring some happiness, that increasing diversity by X percent or adopting X policy changes will officially bring some justice, and so forth ("all-or-nothing" goals; Soman & Cheema, 2004, p. 54).

Corresponding Author:
Ed O'Brien, The University of Chicago, Booth School of Business
Email: eob@chicagobooth.edu

On the one hand, calling for categorically defined change should have various motivational benefits. Normative models of rationality advise that people establish goal markers for accurately diagnosing success (Dawes et al., 1989; Fischoff, 1982; Simon, 1979). People are more likely to take action when solutions to problems are concretely specified than when they are debated in the abstract (Gollwitzer, 1999; Gollwitzer & Oettingen, 2011; Locke & Latham, 1990; Mento et al., 1987). On the other hand, no matter their motivation, people can (and often do) still fall short (Eskreis-Winkler & Fishbach, 2020; Kruglanski et al., 2002; Simonton, 2003)—raising an important question regarding how people respond to enjoying genuine progress, but not enough to transcend categorical failure.

I explored this question through the lens of observers: Rather than assessing effects of categorical thresholds on one's own motivation to pursue goals, I assessed how observers evaluate actors as they then work to hit these marks—and fall short (yet make relative progress). Understanding the observer's perspective is important—not just because little research (if any) has examined this perspective but also because solving problems entails group dynamics that observers affect, too. For example, if people believe that an organization has worked hard to reform, a rally of public support could bring additional resources, allowing them to further the cause; if people believe that they have lazily addressed the issue, a public firestorm could pressure collapse.

The current research tested whether observers become more likely to write off improvement efforts that fail to make categorical change (vs. how observers respond to that same degree of change but without categorical markers). For example, if neither Organization A nor Organization B succeeds in accomplishing the same central reform they were called to make, they may be similarly rebuffed in public opinion—even if one made many smaller reforms while the other did nothing. I refer to this possibility as "negative lumping": Observers may shift to lump and dismiss improvement outcomes as "all the same" when framed as absolute failures to change—despite distinctions in success.

Why? Generally speaking, framing outcomes as categorical failures likely elicits negative emotions (Heath et al., 1999; Kahneman & Tversky, 1979; Lewin et al., 1944) and attributions (Bandura & Simon, 1977; Eskreis-Winkler & Fishbach, 2019) that may inhibit observers' willingness to recognize distinctions between failed attempts. When people encode something as negative, it can be hard to appreciate its upsides (whereas it can be easy to appreciate the downsides of positive things; Baumeister et al., 2001; Ledgerwood & Boydstun, 2014; Rozin & Royzman, 2001). More specific to improvement contexts, categorical thresholds likely shift this reference for diagnosing success from "How much progress did they make?" to "Did they do what they were supposed to do?"—fostering expectations about basic norms of conduct and cooperation (Gavrilets & Richerson, 2017; Weiner, 1995). Accordingly, failing to do what one is supposed to do likely seems like an eschewal of doing the bare minimum that lacks serious intent to change—and people hesitate to credit (perceived) low-effort achievements (Klein & O'Brien, 2017; Kruger et al., 2004; Morales, 2005; Weiner, 1985).

These possibilities led to three hypotheses. First, people may exhibit a negative-lumping effect: People may shift to dismiss improvement efforts that fail to make categorical change—despite distinctions (Experiments 1–7). Second, this effect may be driven by observers' inferences about actors' intent to change (Experiments 8–12). Third, this effect may lead observers to underreward and underinvest in efforts toward "merely" relative progress (Experiments 13 and 14)—even when absolute progress requires step-by-step support.

I report all measures, manipulations, and exclusions (if any). All experiments were preregistered (including all sample sizes, measures, and hypotheses) for purposes of transparency and combatting selective reporting. I predetermined sample sizes of at least 100 adult participants per cell, or more whenever resources allowed. Procedures were reviewed and approved by The University of Chicago Institutional Review Board. All data files, full original study materials, and copies

Statement of Relevance

When people call for change, they often specify their goals. For example, climate activists might fight for corporations to reduce their carbon footprint "by 50% by decade's end," social activists might fight for "these five policy reforms," and so on. Decades of psychological science suggest that this is a wise strategy; calling for very concrete changes (as opposed to "just do better") should help motivate the target to actually do something about the problem. However, the present research revealed an unintended backfiring effect of this strategy: It reduces people's appreciation for critical improvements that were or could still be made but happen to fall short (e.g., people may dismiss a "mere" 40% reduction or a "mere" four reforms as inconsequential—and so settle for no changes instead). People's motivation to make a better world might be increased by helping them appreciate that relative progress is, in fact, progress.

of the preregistrations are publicly available at https://osf.io/q7vj9/.

Experiment 1: Negative Lumping

Method

Experiment 1 tested for the basic effect. Participants evaluated two problematic entities that worked to change their ways for the better over time, with one clearly making more improvements than the other by time's end. Participants' task was simple: to indicate which of the two they viewed more positively, relative to each other—plus a third option to dismiss them both as all the same. I hypothesized that participants would be more likely to choose "all the same" when I added threshold framing indicating that neither entity made *categorical* change—despite participants' task being relative, with one entity still outperforming the other by the same degree (i.e., regardless of the threshold). That is, the same relative progress that people might readily appreciate on its own may be more easily dismissed merely because it is framed as falling short.

I included comparison conditions in which still other participants evaluated entities that both passed these thresholds by the same degrees that others had failed.[1] These conditions rendered the tests especially interesting and informative. For example, perhaps negative lumping reflects any lumping; when people are made aware of any categorical boundary, entities that fall on the same side may suddenly seem more similar to each other. If so, then participants should conclude that passing entities are all the same at the same rate they do for failing entities. In contrast, I predicted asymmetrically strong negative lumping.

Participants. I requested 400 "Cloud Approved" participants from CloudResearch, yielding 401 individuals (age: $M = 39.43$ years, $SD = 12.38$; 43% women; 22% non-White) who participated for $1.00 each.

Procedure. Participants were randomly assigned to condition in a 2 (scores to compare: both low scores vs. both high scores; between-subjects) × 2 (presence of threshold: yes [present] vs. no [absent]; between-subjects) × 8 (domain: eight problems; within-subjects) design.

Participants were invited to complete a study on "social judgment" that involved evaluating pairs of entities and rating their views "about how these two targets compare to each other in terms of their change" (i.e., participants' task was described as explicitly relative right from the start). They then evaluated eight domains, one by one in random order, each corresponding to a different real-world problem. They evaluated (a) sustainability (two manufacturers working to combat their harmful environmental impact), (b) academics (two classrooms working to combat their low quality of learning), (c) health (two cities working to combat their poorly run public health infrastructures), (d) technology (two social media platforms working to combat their lack of inclusion and free speech), (e) habits (two students working to combat their bad habits), (f) athletics (two athletes working to combat their struggling performance), (g) happiness (two everyday people working to combat their unhappiness), and (h) personality (two mental health patients working to combat their antisocial behavior). I included eight domains simply for generalizability (e.g., they assess a mix of problems, across both organizations and individuals).

For each, participants read about two matched entities pursuing the same improvement goal. For example, for sustainability (for all domains, see Table A1 in the Appendix), participants read the following:

> Organization A and Organization B are both manufacturing companies. . . . At the start of last year, both organizations decided to try to improve their position in terms of issues related to sustainability and environmental impact. . . . They both start at the same point. Now, the year has come and gone. Here are their improvement scores.

I then inserted these scores, which I informed participants were determined by "an unbiased external rating system." The drawing of these scores is where the key manipulations took place.

Both-low participants learned that one entity earned a score between 1 and 20 "improvement points" and the other between 21 and 40 (randomly drawn); for both-high participants, one entity earned between 61 and 80 improvement points and the other between 81 and 100 (randomly drawn). I randomized which entity scored higher and their presentation order (these features were rerandomized for each domain, for each participant). Using this scoring system for the design yielded two major benefits. First, I maximized generalizability by drawing from many outcomes rather than limiting the test to one specific number. Second, I objectively defined and matched all relative degrees of progress; one entity always outperformed the other by a full 20 improvement points on average.

I then crossed this score-pair manipulation with the threshold manipulation. For each domain, no-threshold participants proceeded right to the dependent variable—"So: Given this information, how do you view these two organizations?"—and indicated their choice via one of

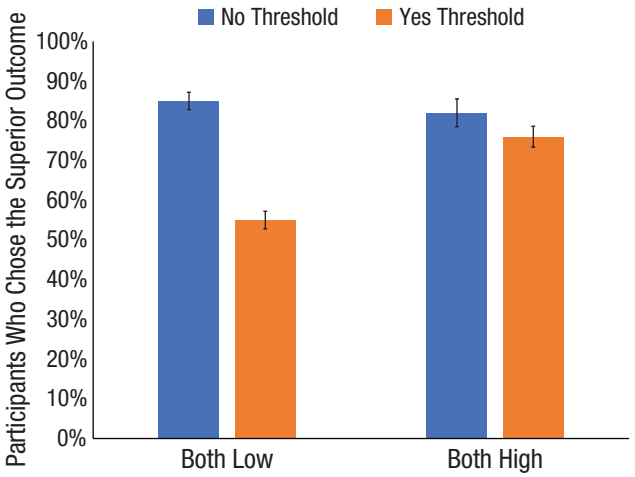


Fig. 1. Experiment 1: percentage of participants who chose the superior outcome in a pair as the superior one (vs. choosing "all the same"), separately for those who compared scores that were both high and both low in the presence or absence of a categorical threshold. Values are estimated across eight domains (collapsed): sustainability, academics, health, technology, habits, athletics, happiness, and personality. Error bars show ±1 *SE*.

three options (forced choice, shown in random order): "I view Organization A as superior," "I view Organization B as superior," and "Ultimately, I view them as no different from each other" (the word "organization" varied by domain). Hence, the answer should be obvious: Whichever entity scored higher should be deemed the superior one in the pair, because this is factually true.

I compared these choices of no-threshold participants with those of yes-threshold participants, who read one additional piece of information about the external rating system:

> This external rating system also sets a clear cut-off. This cut-off is a score of X. Thus, **any** score **at all** below X is considered a "Fail" in terms of improving [themselves], and **any** score **at all** above X is considered a "Pass" in terms of improving [themselves].

X was a randomly drawn number from 41 to 60 (rerandomized for each domain, for each participant; moreover, the word "themselves" varied to match each domain). These yes-threshold participants then completed the same dependent variable that no-threshold participants did.

After making their eight choices, all participants reported demographic information and rated how confusing they found the study, whether they vividly imagined the prompts, and how confident they felt in their responses (each from 1, *not very*, to 7, *very*). Last, they completed an attention check for whether they were shown threshold information (forced choice: "yes" vs. "no").

Results

Main results. As preregistered, participants' responses were recoded as binary: if they chose "all the same" or not (i.e., if they lumped or discriminated, respectively).[2] I then conducted a repeated measures binary logistic regression analysis using the generalizing estimating equations (GEE) procedure in SPSS, entering participant as a subject variable, domain as a within-subjects variable, score pair and threshold as between-subjects variables, and this binary choice as the dependent variable.

There was a main effect of threshold (Wald = 37.11, $df = 1$, $p < .001$)—critically, this was qualified by a two-way interaction with score pair (Wald = 14.22, $df = 1$, $p < .001$). Further, there was no three-way interaction with domain (Wald = 8.17, $df = 7$, $p = .318$), yielding a robust effect (see Fig. 1; for all other output, which is incidental to my hypothesis, see the Supplemental Material available online).

Next, I unpacked this interaction. First, when examining the pairwise comparisons for how participants evaluated the low-scoring pairs, I found that they indeed exhibited negative lumping: Among no-threshold participants ($n = 104$), the vast majority simply chose the higher scoring entity as the superior one in the pair (on average, across all eight domains: 85% of participants in this condition—roughly 88 of 104—made this choice); critically, however, significantly fewer of their yes-threshold counterparts ($n = 98$) shared this view (on average, across all eight domains: only 55% of participants in this condition—roughly 54 of 98—made this choice), thus leaving the rest to dismiss the two entities as all the same, Wald = 43.75, $p < .001$.

Second, and just as critical, this lumping effect did not emerge for equivalent successes. When comparing how participants evaluated the high-scoring pairs, I found a similar rate of no-threshold participants ($n = 101$; average estimate across domains: 82%, or roughly 83 of 101) and yes-threshold participants ($n = 98$; average estimate across domains: 76%, or roughly 74 of 98) simply chose the higher scoring entity as the superior one in the pair, Wald = 2.09, $p = .148$.

Results for each domain individually. Next, although there was no three-way interaction with domain, I was curious to explore domain-level fluctuations (for the full individual figures, see the Supplemental Material). From strongest to weakest effect size, the critical effect (negative lumping—reflected in the two-way interaction between threshold and score pair) was as follows: sustainability

(Wald = 11.59, p = .001), technology (Wald = 6.11, p = .013), health (Wald = 5.99, p = .014), athletics (Wald = 5.54, p = .019), personality (Wald = 5.22, p = .022), academics (Wald = 4.44, p = .035), happiness (Wald = 3.85, p = .050), and habits (Wald = 0.40, p = .526).

Other variables. Finally, most participants passed the attention check (98%; 394 of 401). Study confusion was low (overall: M = 1.68, SD = 1.25), engagement was high (overall: M = 4.59, SD = 1.81), and participants felt confident in their responses (overall: M = 5.75, SD = 1.26). All patterns held when analyses were rerun excluding attention-check failures and controlling for confusion, engagement, confidence, and demographics (for all of these results, see the Supplemental Material).

Discussion

Experiment 1 revealed initial evidence for negative lumping. Undertakings to make change were more likely to be dismissed as all the same merely when framed as failing to accomplish absolute reform—despite one remaining just as superior to the other. Participants did not lump equivalent passing outcomes, highlighting a unique effect of falling short of (vs. surpassing) categorical change.

Experiment 2: Lumping (and Not Lumping) the Same Pairs of Outcomes

Method

Next, I sought to replicate these effects while inverting the manipulation. Rather than manipulating the placement of the scores (holding the threshold constant), I manipulated the placement of the threshold (holding the scores constant). I hypothesized that the same literal outcomes may be more likely to be lumped together if framed as absolute failures than if framed as absolute successes.

Participants. I requested 300 participants from Amazon's Mechanical Turk (MTurk), yielding 301 individuals (age: M = 35.82 years, SD = 10.47; 33% women; 23% non-White) who participated for $1.00 each.

Procedure. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 8 (domain: eight problems; within-subjects) design. Procedures were essential identical to those in Experiment 1. Participants evaluated the same eight domains (see Table A1), and for each, they chose which of two entities they viewed as the superior one in the pair (or categorized them as "all the same"). Here, however, I drew different scores.

First, participants in all conditions always compared one entity that scored 26 to 50 improvement points with another that scored between 51 and 75 improvement points (all randomly drawn and randomized as in Experiment 1). Note that, if anything, the answer should have been even more obvious in this experiment (on average, one entity outperforms the other by 25 points).

Next, each participant was randomly assigned to one of three conditions. Control participants simply made their choices—which, again, for all participants, were explicitly relative (holding all else equal between the options). Both-fail and both-pass participants, however, first saw threshold information. Following the same prompt in Experiment 1, both-fail participants learned that any score below 76 to 100 was a fail (and any score above was a pass), whereas both-pass participants learned that any score above 1 to 25 was a pass (and any score below was a fail); all numbers were randomly drawn from these ranges. Then they made their choices.

Last, after making their eight choices, all participants completed the same end-of-study items as in Experiment 1 (demographics; confusion, engagement, and confidence; and attention check).

Results

Main results. I conducted a repeated measures binary logistic regression analysis using the GEE procedure in SPSS, entering participant as a subject variable, domain as a within-subjects variable, framing as a between-subjects variable, and choice as the dependent variable.

I again observed the key main effect of framing (Wald = 17.93, df = 2, p < .001), which was again not qualified by an interaction with domain (Wald = 18.57, df = 14, p = .182; see Fig. 2; for all other output, which is incidental to my hypothesis, see the Supplemental Material).

Pairwise comparisons confirmed that this main effect was driven by a unique shift among both-fail participants: A similar majority of control participants (n = 101; average estimate across domains: 78%, or roughly 79 of 101) and both-pass participants (n = 100; average estimate across domains: 75%, or roughly 75 of 100) simply chose the higher scoring entity as the superior one in the pair (Wald = 0.62, p = .430), yet significantly fewer both-fail participants (n = 100) shared this view (on average, across all eight domains: only 60% of these participants—roughly 60 of 100—made this choice), thus leaving the rest to dismiss the two entities as all the same (both fail vs. control: Wald = 16.30, p < .001; both fail vs. both pass: Wald = 9.77, p = .002).

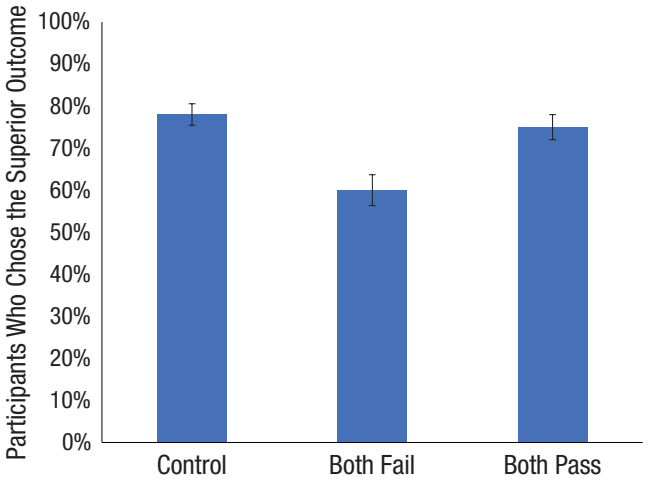


Fig. 2. Experiment 2: percentage of participants who chose the superior outcome in a pair as the superior one (vs. choosing "all the same"), separately for each framing condition. Values are estimated across eight domains (collapsed): sustainability, academics, health, technology, habits, athletics, happiness, and personality. Error bars show ±1 *SE*.

Results for each domain individually. As in Experiment 1, I explored domain-level fluctuations in the critical effect (i.e., the full U-shaped pattern in Fig. 2, showing the unique drop among both-fail participants). This pattern was observed in seven of the eight domains (for the full individual figures, see the Supplemental Material), listed from strongest to weakest effect size: personality (Wald = 13.47, $p = .001$), health (Wald = 12.67, $p = .002$), academics (Wald = 11.67, $p = .003$), athletics (Wald = 11.23, $p = .004$), habits (Wald = 10.29, $p = .006$), sustainability (Wald = 8.63, $p = .013$), and technology (Wald = 7.87, $p = .020$). No effects emerged for happiness (Wald = 0.57, $p = .753$).

Other variables. Finally, most participants passed the attention check (93%; 279 of 301). Study confusion was low (overall: $M = 1.92$, $SD = 1.63$); engagement was high (overall: $M = 4.91$, $SD = 1.72$), as was confidence (overall: $M = 5.98$, $SD = 1.23$). All patterns held when the analyses were rerun excluding attention-check failures and controlling for confusion, engagement, confidence, and demographics (for all of these results, see the Supplemental Material).

Discussion

Experiment 2 documented further evidence for a unique negative-lumping effect—here, even when all participants evaluated identical pairs of outcomes, merely framed differently.

Experiment 3: Major Versus Minor Reforms

Method

Next, I sought to replicate these effects in a new design. Participants evaluated specific reforms that problematic entities were called to make. Control participants lacked categorical cutoffs; experimental participants had them in the form of my flagging the key reform to make. I hypothesized that there would be a shift toward negative lumping if neither entity made this key reform.

Participants. I requested 600 participants from MTurk, yielding 613 individuals (age: $M = 35.90$ years, $SD = 10.99$; 37% women; 28% non-White) who participated for $1.00 each.

Procedure. Participants were randomly assigned to condition in a 3 (framing: control vs. neither make big change vs. both make big change; between-subjects) × 8 (domain: eight problems; within-subjects) design. Procedures were generally similar to those in Experiments 1 and 2: Participants evaluated eight problems, and for each, they chose which of two entities they viewed as the superior one in the pair (or categorized them as "all the same"). However, I made two key changes for the current purposes.

First, I used different domains and prompts, simply for further generalizability. Here, all actors were explicitly called to change. All domains involved organizational actors (e.g., as opposed to individuals working to improve themselves). Participants evaluated (a) the environment (two companies working to be more environmentally friendly), (b) schools (two schools working to address low achievement), (c) health (two cities working to address poor health care access), (d) technology (two platforms working to address free-speech issues), (e) culture (two teams working to address poor performance), (f) finances (two banks working to address slowdown), (g) transparency (two administrations working to address their lack of transparency), and (h) harassment (two organizations working to address their issues related to workplace harassment). For example, for harassment (for all domains, see Table A2 in the Appendix), participants read the following:

> Organization A and Organization B have both been involved with problems of workplace harassment. They have each been equally problematic on this front. . . . An external evaluator diagnosed their specific problems (assume this evaluator is completely unbiased; its assessments represent

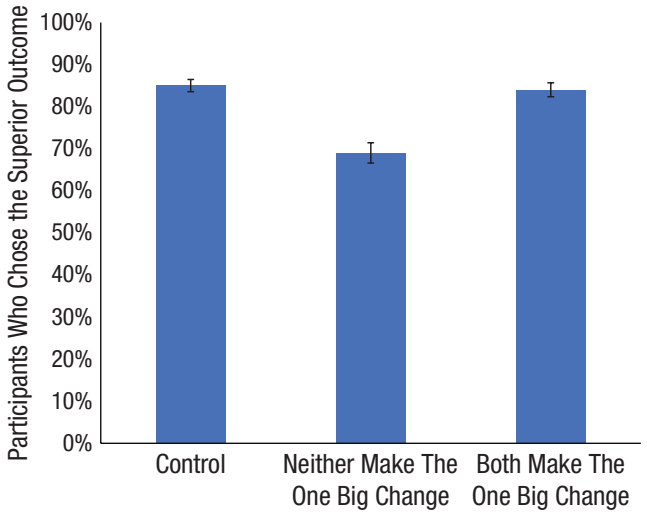


Fig. 3. Experiment 3: percentage of participants who chose the superior outcome in a pair as the superior one (vs. choosing "all the same"), separately for each framing condition. Values are estimated across eight domains (collapsed): environment, schools, health, technology, culture, finances, transparency, and harassment. Error bars show ±1 *SE*.

> true diagnoses). . . . Each organization was informed that they need to fix the following issues (listed in no particular order).

Second, rather than using the scoring system from Experiments 1 and 2, I instead showed participants three specific reforms (displayed in random order) that both entities were mandated to enact. I created unique lists for each domain. For example, for harassment, participants read that the two entities were instructed to "create more diverse teams for group tasks," "rotate team leaders more frequently," and "implement code of conduct training" (for the unique lists for all domains, see Table A2). Using specified reforms further expanded the experiment's ecological validity.

Then, all participants were informed that "time has now passed" and they would learn what each entity ended up doing; at this point, each participant was randomly assigned to one of three conditions.

Control participants learned that one entity successfully made one of the three changes (selected at random) but that the other entity made two of the three changes (which always contained this change, plus one more). I randomized which entity was which and their presentation order. They completed the same dependent measure from Experiments 1 and 2. For other participants, this procedure was identical—except that I additionally flagged one change on the initial list of three as "the single most important change" to make. In turn, for neither-make-big-change participants, I programmed the experiment such that whichever item was flagged as the big item ended up not being enacted by either entity; for both-make-big-change participants, whichever item was flagged did end up being enacted by both.

Last, after making their eight choices, all participants completed the end-of-study items from Experiments 1 and 2 (demographics; confusion, engagement, and confidence; and attention check).

Results

Main results. I conducted a repeated measures binary logistic regression analysis using the GEE procedure in SPSS, entering participant as a subject variable, domain as a within-subjects variable, framing as a between-subjects variable, and choice as the dependent variable.

The lumping effect again emerged, reflected in the main effect of framing (Wald = 42.65, $df = 2$, $p < .001$)—and again, with no interaction with domain (Wald = 18.07, $df = 14$, $p = .204$; see Fig. 3; for all other output, which is incidental to my hypothesis, see the Supplemental Material).

Pairwise comparisons confirmed that this main effect was driven by a unique shift among neither-make participants, yielding the same U-shaped pattern from Experiment 2: A similar majority of control participants ($n = 210$; average estimate across domains: 85%, or roughly 179 of 210) and both-make participants ($n = 200$; average estimate across domains: 84%, or roughly 168 of 200) simply chose the higher scoring entity as the superior one in the pair (Wald = 0.51, $p = .475$), yet significantly fewer neither-make participants ($n = 203$) shared this view (on average, across all eight domains: only 69% of these participants—roughly 140 of 203—made this choice), thus leaving the rest to dismiss the two entities as all the same (neither make vs. control: Wald = 36.02, $p < .001$; neither make vs. both make: Wald = 25.92, $p < .001$).

Results for each domain individually. Again, I was curious to assess fluctuations by domain (for the full individual figures, see the Supplemental Material), listed from strongest to weakest effect size: finances (Wald = 40.62, $p < .001$), harassment (Wald = 24.80, $p < .001$), transparency (Wald = 23.58, $p < .001$), health (Wald = 22.71, $p < .001$), culture (Wald = 16.92, $p < .001$), sustainability (Wald = 14.07, $p = .001$), technology (Wald = 12.52, $p = .002$), and academics (Wald = 11.34, $p = .003$).

Other variables. Finally, most participants passed the attention check (87%; 530 of 613). Study confusion was low (overall: $M = 2.05$, $SD = 1.67$); engagement was high (overall: $M = 4.96$, $SD = 1.69$), as was confidence (overall: $M = 5.98$, $SD = 1.17$). All patterns held when the analyses

were rerun excluding attention-check failures and controlling for confusion, engagement, confidence, and demographics (for all of these results, see the Supplemental Material).

Discussion

Experiment 3 further highlighted a unique negative-lumping effect. Presumably, it is more welcome news for a problematic entity to make two reforms than to make one—yet participants became more likely to dismiss these differences when neither entity achieved the major reform.

Experiments 4 to 7: Negative Lumping Is Robust

Method

To summarize the findings thus far, I found evidence for a unique negative-lumping effect across three different highly controlled study designs, each of which assessed a wide range of societal issues that held all degrees of relative progress precisely constant across conditions.

One additional advantage of these designs is that they are readily adaptable for testing boundaries. To this end, Experiments 4 to 7 followed the design of Experiment 2, which I took as a representative paradigm to assess the basic effect. However, I varied many other parameters and phrasings. Because these experiments are mostly identical to Experiment 2, I report them here in streamlined fashion (for full reporting, see the Supplemental Material). Together, my goal was to simply assess whether the effects observed so far indeed generalized beyond some idiosyncratic design feature.

Participants. In total, across Experiments 4 to 7 (conducted at separate times with unique participants), I requested 5,100 participants from MTurk, yielding 5,130 individuals (age: M = 38.79 years, SD = 12.30; 48% women; 25% non-White) who participated for $0.25 each. For each experiment, I requested a sample size that would yield approximately 100 participants per experimental cell.

Procedure.

Experiment 4. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 5 (threshold type: five sets of phrasings; between-subjects) design. My goal here was to vary the threshold. Thus far, I had used pass-versus-fail framing; entities that fall short count as a fail, and those that succeed count as a pass. Here, I randomly assigned each participant to see one of five threshold phrasings: pass versus fail (replicating previous experiments), no versus yes, low versus high, bad versus good, or poor versus excellent. I hypothesized that the lumping effect is not just a function of pass versus fail and so would not be moderated by threshold type. All other procedures were identical to those in Experiment 2—except that all participants evaluated the same single domain (personality; see Table A1) rather than eight.

Experiment 5. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 5 (choice type: five sets of phrasings; between-subjects) design. My goal here was to vary the dependent variable. Thus far, participants had chosen which entity they viewed as superior. Here, I randomly assigned each participant to make one of five choices: which entity was "superior" (replicating previous experiments), "better," "higher," "more noteworthy," or "more promising." I hypothesized that the lumping effect is not just a function of superior and so would not be moderated by choice type. All participants again evaluated the same personality domain as above, with all else being identical to Experiment 2.

Experiment 6. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 5 (choice type: five sets of phrasings; between-subjects) design. I again varied the dependent variable, but here, I randomly assigned each participant to see one of five sets of negatively phrased choices, affording a more conservative test: Participants chose which entity was "less negative," "less bad," "less harmful," "less problematic," or "less troublesome." I hypothesized that the lumping effect would still emerge even when participants were allowed to draw less favorable distinctions (i.e., no moderation by choice type). Again, they evaluated the same single domain, with all else being identical to Experiment 2—except that I assessed a new domain for further generalizability (corrupt governments; see Table A3 in the Appendix).

Experiment 7. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 2 (direction of change: higher is better vs. lower is better; between-subjects) design. Finally, my goal here was to vary the direction of change, providing another conservative test. In the scoring system from previous experiments, higher scores always meant better scores. Here, I randomly assigned each participant to one of two direction conditions: I informed

participants (across all framing conditions) that "higher numbers = more improvement" (replicating previous experiments) or that "lower numbers = more improvement." Perhaps Experiments 1 and 2 reflected some incidental number-based effect (e.g., whether "0" takes on special psychological meaning as a reference point; Shampanier et al., 2007). However, I hypothesized that the lumping effect would still emerge with this feature flipped (i.e., no moderation by direction). All participants again evaluated the same single domain, with all else being identical to Experiment 2—except that I again assessed another new domain (discriminatory hiring; see Tables A3 in the Appendix).

Results

For each experiment, I conducted the same analyses as in Experiment 2. For full reporting (including end-of-study variables and exclusion analyses, showing the same patterns), see the Supplemental Material. Most critically, negative lumping robustly emerged across these parameters as hypothesized.

Experiment 4 (beyond pass versus fail). The key effect of framing was significant (Wald = 25.65, $df = 1$, $p < .001$), showing the full U-shaped pattern of negative lumping (just as seen in Figs. 2 and 3). Moreover, negative lumping was robust to threshold type, showing no interaction (Wald = 2.21, $df = 1$, $p = .137$). The following are listed from strongest to weakest effect size: no versus yes (Wald = 13.69, $p < .001$), pass versus fail (Wald = 12.23, $p < .001$), poor versus excellent (Wald = 7.61, $p = .006$), bad versus good (Wald = 0.47, $p = .492$), and low versus high (Wald = 0.43, $p = .512$).

Experiment 5 (beyond "superior"—positive). The key effect of framing was significant in the same way (Wald = 32.07, $df = 1$, $p < .001$) and was robust to choice type, showing no interaction (Wald = 0.44, $df = 1$, $p = .508$). The following are listed from strongest to weakest effect size: more noteworthy (Wald = 21.00, $p < .001$), superior (Wald = 7.56, $p = .006$), higher (Wald = 6.80, $p = .009$), better (Wald = 6.12, $p = .013$), and more promising (Wald = 0.01, $p = .909$).

Experiment 6 (beyond "superior"—negative). These same patterns emerged for negative choices as well—as shown via the key effect of framing (Wald = 22.60, $df = 1$, $p < .001$), along with no interaction with choice type (Wald = 0.37, $df = 1$, $p = .545$). The following are listed from strongest to weakest effect size: less harmful (Wald = 9.56, $p = .002$), less problematic (Wald = 7.88, $p = .005$), less negative (Wald = 4.17, $p = .041$), less troublesome (Wald = 4.04, $p = .045$), and less bad (Wald = 0.61, $p = .437$).

Experiment 7 (different directions of change). Yet again, the key effect of framing was significant in the same way (Wald = 16.74, $df = 1$, $p < .001$) and likewise was robust to change direction, showing no interaction (Wald = 0.53, $df = 1$, $p = .466$). The negative-lumping effect emerged regardless of whether higher scores conveyed more improvement (Wald = 6.46, $p = .011$) or whether lower scores conveyed more improvement (Wald = 10.33, $p = .001$).

Discussion

All told, Experiments 4 to 7 suggested that negative lumping generalizes beyond various design features, at least in these study contexts.

Experiment 8: Falling Short Distorts Perceptions of Effort

Method

Next, I sought to unpack why this effect occurs. As theorized, falling short of full reform may be interpreted as a uniform lack of care to truly address the issue—and public opinion tends not to reward "unearned" achievements. Experiment 8 tested this possibility.

Participants. I requested 300 participants from MTurk, yielding 304 individuals (age: $M = 36.35$ years, $SD = 11.04$; 49% women; 29% non-White) who participated for $0.25 each.

Procedure. Participants were randomly assigned to condition in a single-factor, three-level (framing: control vs. both fail vs. both pass; between-subjects) design. Procedures resembled those used in previous experiments, except that they used a new domain and measures. All participants evaluated the same issue, involving reforms to workplace culture. They read the following:

> Industry X is due for some cultural reform. Many of its practices are woefully outdated by today's standards. Traditionally, only certain kinds of people hold all the power. Others regularly experience harassment and discrimination. People fear speaking out. Despite proclaiming to operate on democratic principles, it is far from a democracy.

As in previous experiments, participants then learned about two organizations in this industry ("Organization A" and "Organization B," with all else equal between them) that worked to reform their cultures over the

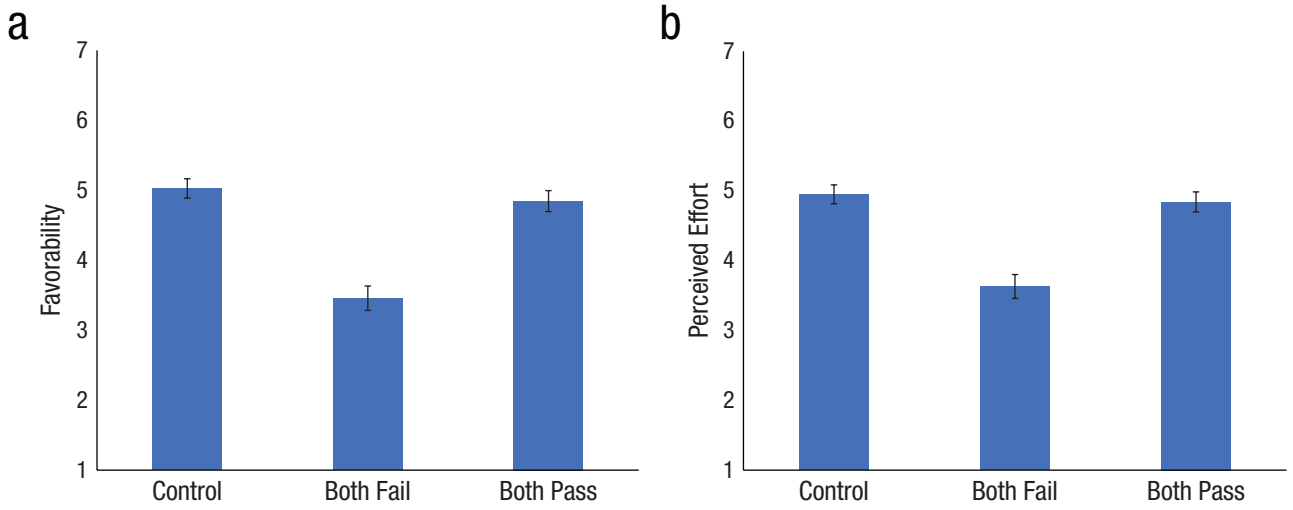


Fig. 4. Experiment 8: mean rating of (a) favorability and (b) perceived effort in each of the three framing conditions. Error bars show ±1 *SE*.

year, and participants then saw their improvement scores. I manipulated the scores just as before (e.g., as in Experiment 2); one organization always improved more than the other by 25 points on average, and participants were assigned to control, both-fail, or both-pass framings. Here, however, participants compared the two organizations with each other on new measures: They completed a dependent-variable block and a mediator block, and blocks were presented one at a time in random order (and the order of items in each block was also randomized).

For the dependent variable, all participants read, "So: Given this information, what are your reactions to Organization B vs. Organization A?" and rated five items each on a scale ranging from 1 (*same difference; at the end of the day, I feel they're basically the same*) to 7 (*totally different; I feel Organization B is clearly much more of this*). The items were "good," "positive," "impressive," "worthy of praise," and "deserving of reward." I collapsed these items into an overall favorability scale (α = .963). The study was programmed so that higher ratings on these scale items reflected a stronger preference for whatever organization had objectively improved more.

For the mediator block, participants rated five other items, which I intended to collapse into an overall perceived-effort scale: "put in serious effort," "made change their top priority," "truly wanted to change," "worked as hard as they could," and "tried everything possible to improve" (α = .958). Again, higher scores reflected a stronger preference for the objectively superior option.

As always, note that when the two organizations were compared with each other—which is what these measures directly asked participants to do—the objectively correct answer should have been obvious. However, I hypothesized that both-fail participants would uniquely express lower favorability than other participants and that this would be driven by correspondingly lower perceived effort.

All participants then completed the same end-of-study items from previous experiments (demographics; confusion, engagement, and confidence; attention check). Finally, at the end of the study, and simply to match the dependent measures from previous experiments, I also asked participants, in forced-choice fashion, to choose which organization they viewed "more positively" and which they believed put in "more serious effort" to change (a third lumping option, from previous experiments was "Ultimately, I view them as no different from each other").

Results

Main results. I conducted a one-way analysis of variance (ANOVA) with framing as the independent variable and the favorability and perceived-effort scales as dependent variables.

For favorability, I observed a significant omnibus effect of framing, $F(2, 303) = 30.82$, $p < .001$ (see Fig. 4a). Planned contrasts reveal the U-shaped pattern from previous experiments: Whereas control ($M = 5.03$, $SD = 1.42$) and both-pass ($M = 4.85$, $SD = 1.48$) participants similarly gave more credit to the more-improved organization, $t(301) = 0.81$, $p = .417$, $d = 0.12$, both-fail participants did not ($M = 3.46$, $SD = 1.77$)—compared with control, $t(301) = 7.21$, $p < .001$, $d = 0.98$; compared with both pass, $t(301) = 6.27$, $p < .001$, $d = 0.85$.

For perceived effort, I observed a parallel omnibus effect of framing, $F(2, 303) = 23.94$, $p < .001$ (see Fig. 4b): Again, control ($M = 4.95$, $SD = 1.37$) and both-pass ($M = 4.84$, $SD = 1.41$) participants similarly ascribed more effort to the more-improved organization, $t(301) = 0.51$, $p = .611$, $d = 0.08$, but again, both-fail participants did not ($M = 3.63$, $SD = 1.73$)—compared with control, $t(301) = 6.26$, $p < .001$, $d = 0.85$; compared with both pass, $t(301) = 5.65$, $p < .001$, $d = 0.77$—instead ascribing similarly low effort.

Note also that for both favorability and perceived effort, the mean ratings of control and both-pass participants each fell significantly above the scale midpoint (4.00), thus categorically favoring the more-improved organization (ps ≤ .001, ds ≥ 0.57), whereas the mean ratings for both-fail participants fell significantly below this midpoint, thus categorically favoring lumping (ps ≤ .032, ds ≥ 0.21).

Mediation. Next, I conducted mediation analyses (SPSS PROCESS Model 4; 5,000 iterations) examining the effect of framing (1 = control, both pass; 2 = both fail) on favorability via perceived effort. The indirect effect was significant ($b = –1.06$, $SE = 0.17$, bootstrapped 95% confidence interval = [–1.42, –0.73]).

Other variables. Finally, these continuous patterns were reflected in the forced-choice results, replicating the findings of prior experiments (see the Supplemental Material). Most participants passed the attention check (91%; 276 of 304). Study confusion was low (overall: $M = 2.04$, $SD = 1.53$); engagement was high (overall: $M = 4.35$, $SD = 1.81$), as was confidence (overall: $M = 5.77$, $SD = 1.18$). All patterns held when analyses were rerun excluding attention-check failures and controlling for confusion, engagement, confidence, and demographics (for all of these results, see the Supplemental Material).

Discussion

Experiment 8 shed light on why negative lumping occurs. Participants became more likely to uniquely lump entities that fell short because they inferred a shared lack of effort.

Experiments 9 to 12: Moderation by Effort Cues

Method

Next, I extended this mediation logic via moderation: If negative lumping is driven by observers' inferences about actors' genuine attempts to address the issue, then observers might not lump failed outcomes when falling short can be attributed to causes beyond lack of effort; the effect should be attenuated in contexts of higher effort failure. Experiments 9 to 12 tested this idea.

Again, as in Experiments 4 to 7, I mostly reran Experiment 2—a representative paradigm to assess the basic effect—except that I manipulated the presence of effort cues in varied ways. Thus, I report these experiments in streamlined fashion (for full reporting, see the Supplemental Material).

Participants. In total, across Experiments 9 to 12 (conducted at separate times with unique participants), I requested 2,600 participants from either MTurk or Cloud Research, yielding 2,622 individuals (age: $M = 37.90$ years, $SD = 12.12$; 50% women; 26% non-White) who each participated for either $0.25 or $0.50 (depending on the experiment). For each experiment, I requested a sample size that would yield approximately 100 participants per experimental cell.

Procedure.

Experiment 9. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 3 (goal difficulty: none vs. low difficulty vs. high difficulty; between-subjects) design. My goal here was to vary the threshold's difficulty. Using Experiment 2 as a base, I further assigned each participant to one of three difficulty conditions. Some participants were not informed about threshold difficulty (replicating the procedure of previous experiments). Others were informed that "this benchmark from this evaluator is exceptionally low; it's a very low bar for change." Still others were informed that "the benchmark from this evaluator is exceptionally high; it's a very high bar for change." I hypothesized that the lumping effect would be attenuated within high-difficulty conditions because observers need not infer that failed actors gave little effort. All else was identical to Experiment 2, except that a single (new) domain was used (outdated technology; see Table A3).

Experiment 10. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 3 (goal salience: none vs. low salience vs. high salience; between-subjects) design. My goal here was to vary whether actors knew of the threshold. Again using Experiment 2 as a base, I assigned each participant to one of three salience conditions. Some participants were not informed of actors' threshold knowledge. Others were informed that "they had full knowledge this cut-off score existed." Still others were informed that "they had zero knowledge this cut-off score existed." I hypothesized that the lumping effect would be attenuated within low-salience conditions because failed actors

did not fail because of low goal effort per se. Again, all else was identical to Experiment 2, except that a single (new) domain was used (green practices; see Table A3).

Experiment 11. Participants were randomly assigned to condition in a 3 (framing: control vs. both fail vs. both pass; between-subjects) × 2 (time left to achieve goal: little time left vs. long time left; between-subjects) design. My goal here was to vary the time left to improve. Again using Experiment 2 as a base, I further assigned each participant to one of two time-left conditions. Some participants were informed that actors' improvement was assessed "1 year into their 1 year window" (replicating previous experiments). Others were informed that it was assessed "1 month into their 1 year window." I hypothesized that the lumping effect would be attenuated with much time left to hit the mark because failed actors may still care but were given too little time to show it. Again, all else was identical to Experiment 2, except that a single (new) domain was used (public health; see Table A3).

Experiment 12. Participants were randomly assigned to condition in a single-factor, two-level (framing of identical outcomes: control vs. threshold; between-subjects) design. The goal here was to assess an inverse hypothesis that follows from my effort-inference proposal: Two identical improvement outcomes should be viewed as different (i.e., they should not be lumped together) when one surpasses a threshold and the other falls short. Such a design directly ruled out the possibility that negative lumping reflects the fact that progress within categorical failure often truly is less effortful than progress within categorical success, as opposed to distorted perceptions per se; if so, then people should indeed lump these same two outcomes (whereas my framework predicted that people would distinguish them when framed as success vs. failure).

All participants evaluated two companies that were described as being identical in every way, including needing to improve the diversity of their workforce (improving diversity; see Table A3). They learned that both companies were called to increase their diversity and that each indeed ended up increasing their diversity by 20% by year's end (thus, whatever it takes to increase diversity by 20% was held exactly constant for both companies). Each participant was then randomly assigned to one of two conditions. Control participants indicated, on the basis of this information alone, which company they viewed as superior (or dismissed them both as "all the same"), as in prior experiments. Threshold participants were also told that, via a random lottery, one company had been assigned to increase their diversity by 10%, whereas the other had been assigned to increase their diversity by 30%. I hypothesized that fewer threshold participants than control participants would lump the companies, despite all participants comparing entities that each increased their diversity by 20%.

Results

For each experiment, I conducted the same analyses as in Experiment 2 (except for Experiment 12 because its different design called for binary logistic regression). For full reporting (including end-of-study variables and exclusion analyses, showing the same patterns), see the Supplemental Material. Most critical to report, negative lumping was moderated by these factors, as hypothesized.

Experiment 9 (less lumping when falling short of high bars). The effect of framing (1 = control, both pass; 2 = both fail) was significant (Wald = 47.14, $df = 1$, $p < .001$) but was further qualified by an interaction with goal difficulty (Wald = 5.73, $df = 1$, $p = .017$). That is, the U-shaped pattern of negative lumping flattened out when the threshold was highly stringent: The key effect of framing (i.e., negative lumping) was large for no-information (Wald = 21.50, $p < .001$) and for low-difficulty (Wald = 29.28, $p < .001$) participants but significantly weaker for high-difficulty participants (Wald = 3.58, $p = .059$), who indeed raised their recognition of the more-improved (but still failed) entity.

Experiment 10 (less lumping when actors are unaware). The effect of framing was again significant (Wald = 8.65, $df = 1$, $p = .003$) but again qualified by an interaction with goal salience (Wald = 7.47, $df = 1$, $p = .006$). The U-shaped pattern flattened out in the same way when failed actors were unaware of the threshold: Negative lumping emerged among no-information (Wald = 4.37, $p = .037$) and among high-salience (Wald = 10.97, $p = .001$) participants but emerged less strongly among low-salience participants (Wald = 0.15, $p = .694$).

Experiment 11 (less lumping with a long time left). Here, too, I observed the same effect of framing (1 = control, both pass; 2 = both fail; Wald = 20.64, $df = 1$, $p < .001$), but it was driven by one condition, reflected in an interaction with time left (Wald = 8.15, $df = 1$, $p = .004$): Participants lumped undertakings together that fell short by year's end of their reform window (Wald = 27.26, $p < .001$) to a greater degree than they lumped undertakings together that fell short as assessed just 1 month into this window (Wald = 1.43, $p = .232$).

Experiment 12 (not lumping when lumping "should" occur). Serving as an inverse test of my hypothesis, nearly all control participants (correctly) lumped together the same improvement (i.e., two companies that each increased their diversity by 20%) as all the same (as chosen by 99%, 101 of 102, of control participants)—yet significantly fewer threshold participants did so when one company was framed as a categorical success and the other was framed as a categorical failure (55%, or 54 of 99, threshold participants, chose "all the same"; Wald = 18.70, $df = 1$, $p < .001$).

Discussion

All told, Experiments 9 to 12 further supported my theorizing about why negative lumping occurs; conveying serious intent by other means may help combat being lumped together (despite failing).

Experiment 13: Underinvesting in Relative Progress

Method

All of these results so far suggested problematic behavioral consequences, in that observers may become more likely to "check out" of supporting relative progress (vs. how much they would invest in that same degree of change but without categorical markers). Experiments 13 and 14 tested this possibility.

First, in Experiment 13, I assessed whether people underinvest in relatively more promising futures—in a context of tangible costs.

Participants. I recruited 285 participants from my university subject pool (age: $M = 31.88$ years, $SD = 13.64$; 51% women; 68% non-White) who participated for $3.00 each.[3] My pool drew from across the university and the surrounding community (37% of the sample were students).

Procedure. Participants were randomly assigned to condition in a single-factor, three-level (framing: control vs. both fail vs. both pass; between-subjects) design. I manipulated framing as in previous experiments, but all other procedures were new, including assessing real-time behavior with tangible stakes.

To begin, all participants were informed that they would compare two groups of lab subjects, just like them, who had allegedly completed a "motor-skills improvement" study over the prior month. The domain of motor skills is less related to social issues per se than were the previous domains I assessed, but it allowed me to precisely manipulate relative progress and maintain realism while holding all else equal between the entities. Also for these reasons, this other study was not real, but all participants were led to think it was until being debriefed.

Throughout the procedures, a research assistant guided participants through exhaustive details about these alleged subjects and their tasks (for the full materials, see https://osf.io/q7vj9/). The key points are highlighted here. All participants learned that these other subjects had worked to improve their motor-skills performance over the last month of practice and that they had now been organized them into two groups: the 100 best improvers and the 100 worst improvers. I then used the same randomized scoring system from previous experiments (e.g., Experiment 2): Participants were shown the average scores of each group; one group earned an improvement score from 26 to 50 and the other a score from 51 to 75 (randomly drawn). Participants then learned that "early next week," these two groups would complete another motor-skills test (just like the tests they had trained on)—and at this point, each participant was randomly assigned to one of three conditions.

Control participants proceeded directly to the dependent variable. They were invited to take a bet on which group would "score higher on average" on this test; they could wager $1.00 of their $2.00 advertised study payment on their choice or could simply opt to keep their $2.00 (forced choice: "yes, take the bet" vs. "no, do not take the bet").[4] I compared the percentage of control participants who took the bet with both-fail and both-pass participants—who followed identical procedures but with additional information about the scoring system. As in previous experiments (e.g., Experiment 2), both-fail participants learned that the external threshold for improving these abilities was a randomly drawn score from 76 to 100 (i.e., both groups fell short), whereas this threshold was drawn from 1 to 25 for both-pass participants (i.e., both groups passed).

After making their betting decision, all participants were debriefed and learned that everyone would receive a full $3.00 study payment. Finally, I gave participants an open-ended prompt to write any thoughts about their decision, and then they reported demographic information, rated how confusing they found the study (1 = *not confusing*, 7 = *very confusing*), and completed an attention check for whether they were shown threshold information (forced choice: yes vs. no).

Thus, I invited participants to take an informed (and obvious) bet on which group would outperform the other on an upcoming test. However, extending the negative-lumping effect from all previous experiments, I hypothesized that both-fail participants might uniquely opt out.

Results

Main results. I conducted a binary logistic regression with framing (1 = control, both pass; 2 = both fail) as a between-subjects variable and betting as the dependent variable. I observed the hypothesized effect of framing (Wald = 9.17, df = 1, p = .002): Pairwise comparisons revealed that a similarly large majority of participants indeed took the bet among the control condition (76%; 72 of 95) and among the both-pass condition (73%; 68 of 93; Wald = 0.18, p = .675)—yet significantly fewer both-fail participants (57%; 55 of 97) took this same bet (both fail vs. control: Wald = 7.64, p = .006; both fail vs. both pass: Wald = 5.53, p = .019).

Other variables. Finally, most participants passed the attention check (93%; 266 of 285). Study confusion was low (overall: M = 2.22, SD = 1.41). All patterns held when the analyses were rerun excluding attention-check failures and controlling for confusion and demographics (for all of these results, see the Supplemental Material).

Discussion

Experiment 13 advanced the evidence for negative lumping by highlighting downstream consequences. All participants were offered the same bet—a relative bet involving one clearly superior future in which to invest—yet both-fail participants were uniquely less likely to take it.

Experiment 14: Underrewarding Relative Progress

Method

In the final experiment, I explored another downstream consequence of negative lumping: People may underreward actual progress currently being made out in the world.

Participants. I requested 700 participants from MTurk, yielding 699 individuals (age: M = 40.65 years, SD = 13.10; 45% women; 25% non-White) who participated for $0.75 each.

Procedure. Participants were randomly assigned to condition in a 3 (category label: present vs. absent; between-subjects) × 2 (category pair: high-performing countries vs. low-performing countries; within-subjects) design. Participants evaluated the Climate Change Performance Index (CCPI), a real monitoring tool that tracks countries annually regarding their adherence to Paris Agreement improvement mandates on the basis of various objective metrics (Burck et al., 2019). Participants were shown CCPI's 2020 report (see https://osf.io/q7vj9/). Their report, as actually published, well captured the previous designs: It ranks 57 countries from 1 (best) to 57 (worst)—and each is assigned an improvement score (based on CCPI's algorithm), which is further assigned to one of five color-coded categories: "very high" (best), "high," "medium," "low," and "very low" (worst). These categories are essentially external thresholds (e.g., in the 2020 report, no country achieved "very high" status).

The dependent variable was a reward-allocation task. I informed participants about an annual fund intended to reward countries for their climate-change improvements efforts, with the entirety of each year's fund needing to be distributed across these countries. Using CCPI's actual list of rankings (except that I masked country names—using "Country 1" and so on—to help avoid other group dynamics), I introduced the within-subjects manipulation: For each participant, I randomly drew two countries from the best-improved category and two countries from the worst-improved category, and participants' task was to allocate a percentage of the fund to each. Participants typed their allocations (each from 0% to 100%) into individual boxes, with an additional box for "all remaining 53 countries" (everything had to sum to 100%, as enforced by the survey software). Of key interest, I compared participants' relative difference in allocation between the two top performers with their relative difference in allocation between the two bottom performers.

Importantly, I programmed these random draws to hold constant, between the two top draws and between the two bottom draws, their difference in ranks (which, on the basis of the natural distribution of CCPI's data, ended up being an average difference of 3.4 ranks) as well as their difference in scores (which ended up being an average difference of 5.0 improvement points). Thus, put concretely, all participants evaluated one country that outperformed another by about 3 ranks and 5 points, and I assessed how much more of the fund they allocated to the better performing country within that pair—and they did this for a pair of top performers and for a pair of bottom performers. In principle, whatever participants think "3 ranks and 5 points" are worth, this incremental value should be worth the same regardless of where it lies on the list. However, consistent with negative lumping, my hypothesis was that the better performing country within the bottom pair would earn a smaller relative gain in funding—despite improving the same degree.

On top of these procedures, I also included a between-subjects manipulation: Some participants completed these procedures as described, whereas other

participants did so without seeing the thresholds; they saw the ranks and scores, but CCPI's color-coded category information ("very low" and so on) was hidden. I understood this manipulation as mapping onto the previous designs, in which participants always viewed the same numerical difference between entities—but some also viewed threshold information, thus eliciting negative lumping. Here, the hypothesized allocation effect should have been weaker among participants who lacked category labels.

Finally, after making their allocations, all participants reported demographic information and rated the confusion, engagement, and confidence items from previous experiments. They also completed an attention check for whether they saw category information (forced choice: yes vs. no), and labels-present participants completed another check for which labels they saw (forced choice: high and low vs. medium). All participants also reported their familiarity with "climate change news" (forced choice: "not at all" vs. "moderate" vs. "very"), whether they believed that climate change is "real and human-caused" (forced choice: "don't believe at all" vs. "believe a moderate amount" vs. "very much believe"), and whether they had ever heard of the CCPI (forced choice: yes vs. no).

Results

Preregistered exclusions. At the time of the study, I learned of a new "Bot Check" feature offered by Qualtrics designed to flag bots by implicitly tracking mouse movements and scoring users from 0.00 ("likely a bot") to 1.00 ("likely a human"). Thus, I preregistered the intention to exclude scores below .70 and, for good measure, any participant who failed an attention check.

Most participants passed this Bot Check (94%; 655 of 699) and both attention checks (96%; 670 of 699). Together, after exclusions, 628 participants (90% of the sample) were retained.

Main results. I computed difference scores for each participant, for each pair of countries (i.e., participants' difference in allocation between their two top-performing countries and their difference in allocation between their two bottom-performing countries), and I then conducted a repeated measures ANOVA with category label as a between-subjects factor, category pair as a within-subjects factor, and differences in allocation as dependent variables.

I observed the hypothesized main effect of category pair, $F(1, 626) = 56.11$, $p < .001$, $d = 0.29$ (see Fig. 5): Overall, within the best-improving pairs, participants allocated a mean of 6.32% ($SD = 13.83\%$) more reward to the better performing country within that pair, but within the worst-improving pairs, participants allocated a mean of just 2.93% ($SD = 9.99\%$) more reward to the better performing country within that pair. This effect emerged despite similar numerical degrees of progress.

Unexpectedly, and contrary to my preregistered hypothesis, this main effect of category pair was not qualified by an interaction with category label, $F(1, 626) = 0.46$, $p = .497$; main effect of category label: $F(1, 626) = 1.15$, $p = .283$. That is, as revealed via pairwise comparisons, this same effect emerged regardless of whether participants saw CCPI's color-coded category column (where I indeed hypothesized that the effect would be found)—difference between best: $M = 6.03\%$, $SD = 15.01\%$; difference between worst: $M = 2.33\%$, $SD = 9.42\%$, $F(1, 626) = 33.92$, $p < .001$, $d = 0.30$, or whether this color-coded category column was hidden (where I hypothesized that the effect would be reduced)—difference between best: $M = 6.63\%$, $SD = 12.52\%$; difference between worst: $M = 3.55\%$, $SD = 10.53\%$, $F(1, 626) = 22.83$, $p < .001$, $d = 0.27$.

Although these patterns were still directionally consistent with the hypothesized interaction, I suspect (in hindsight) that the top and bottom of a long list of rankings likely struck people as being categorically different "on their own," without needing labels to denote such distinctions—a possibility that I then also empirically tested and indeed confirmed in a subsequent posttest.[5] Thus, I interpret these results as remaining consistent with and in support of my framework.

Other variables. Finally, study confusion was low (overall: $M = 2.28$, $SD = 1.57$), and engagement (overall: $M = 3.90$, $SD = 2.05$) and confidence (overall: $M = 4.69$, $SD = 1.61$) were relatively high. There was a mix in familiarity with climate-change news (11% not at all, 70 of 628; 69% moderate, 430 of 628; 20% very, 128 of 628), beliefs that climate change is real (8% do not believe at all, 49 of 628; 29% moderately believe, 183 of 628; 63% very much believe, 396 of 628), and CCPI knowledge (19% yes, 116 of 628; 82% no, 512 of 628). All patterns held when the analyses were rerun controlling for these variables (for all of these results, see the Supplemental Material).

Discussion

Experiment 14 revealed another consequence of negative lumping. Participants rewarded the same numerical strides differently depending on where they were made: Gains made in poor-performing countries were valued at less than half the rate as similar numerical gains made in high-performing countries.

2 countries from HIGH
(~3 ranks and ~5 points apart)

Difference in allocation:

$M = 6.32\%$
$SD = 13.83\%$

$F(1, 626) = 56.11$,
$p < .001$ ($d = 0.29$)

2 countries from VERY LOW
(~3 ranks and ~5 points apart)

Difference in allocation:

$M = 2.93\%$
$SD = 9.99\%$

Country	Score	Category
1. Sweden	75.77	HIGH
2. Denmark	71.14	HIGH
3. Morocco	70.63	HIGH
4. United Kingdom	69.8	HIGH
5. Lithuania	66.22	HIGH
6. India	66.02	HIGH
7. Finland	63.25	HIGH
8. Chile	62.88	HIGH
9. Norway	61.14	HIGH
10. Luxembourg	60.91	HIGH
11. Malta	60.6	HIGH
12. Latvia	60.75	HIGH
13. Switzerland	60.61	HIGH
14. Ukraine	60.6	HIGH
15. France	57.9	MEDIUM
16. Egypt	57.53	MEDIUM
17. Croatia	56.97	MEDIUM
18. Brazil	55.82	MEDIUM
19. Germany	55.78	MEDIUM
20. Romania	54.85	MEDIUM
21. Portugal	54.1	MEDIUM
22. Italy	53.92	MEDIUM
23. Slovak Republic	52.69	MEDIUM
24. Greece	52.59	MEDIUM
25. Netherlands	50.89	MEDIUM
26. China	48.16	MEDIUM
27. Estonia	48.05	MEDIUM
28. Mexico	47.01	LOW
29. Thailand	46.76	LOW
30. Spain	46.03	LOW
31. Belgium	45.73	LOW
32. South Africa	45.67	LOW
33. New Zealand	45.67	LOW
34. Austria	44.74	LOW
35. Indonesia	44.65	LOW
36. Belarus	44.18	LOW
37. Ireland	44.04	LOW
38. Argentina	43.77	LOW
39. Czech Republic	42.93	LOW
40. Slovenia	41.91	LOW
41. Cyprus	41.66	LOW
42. Algeria	41.45	LOW
43. Hungary	41.17	LOW
44. Turkey	40.76	VERY LOW
45. Bulgaria	40.12	VERY LOW
46. Poland	39.98	VERY LOW
47. Japan	39.03	VERY LOW
48. Russia	37.85	VERY LOW
49. Malaysia	34.21	VERY LOW
50. Kazakhstan	33.39	VERY LOW
51. Canada	31.01	VERY LOW
52. Australia	30.75	VERY LOW
53. Iran	28.41	VERY LOW
54. Korea	26.75	VERY LOW
55. Chinese Taipei	23.33	VERY LOW
56. Saudi Arabia	23.03	VERY LOW
57. United States	18.6	VERY LOW

Fig. 5. Experiment 14: difference in reward-fund allocation for relative climate-change progress (based on the Climate Change Performance Index; Burck et al., 2019). Country names were masked in the experiment. I also manipulated whether the category column was present or absent (results shown are collapsed across this factor).

General Discussion

Using controlled paradigms assessing more than 20,000 observed choices across 25 issues, I found that participants became more likely to lump undertakings that failed to make categorical change as "all the same" (vs. how they responded to that same degree of change but without categorical markers). Falling short conveyed an eschewal of doing the bare minimum without serious intent to change—shifting participants toward dismissing relative progress and withholding support from making it. Moreover, lumping was specific to the negative: Taking the same progress participants initially celebrated, about 23% shifted to dismiss it when framed as absolute failure—but only 5% shifted to lump absolute success (average "drops" across experiments).

Theoretical contributions

Much literature has examined how categorization affects comparison, showing lumping-type effects. A popular summary depicts people as cognitive misers (Fiske & Taylor, 1991) who jump to categorize, with mere category labels being sufficient to minimize perceived differences; in one study, varied-length lines looked more similar if labeled as "Group A" (Tajfel & Wilkes, 1963), highlighting this basic clouding effect. That the participants in the present experiments shifted to lump failures, but not equivalent successes, qualifies this depiction, at least in improvement contexts. Just as the categorization of a target (and thus its potential for lumping) depends on features of surrounding objects (Tversky, 1977), it may also depend on features of the target itself (e.g., valence) in ways that may be masked by an emphasis on how categories affect member differences.

The direction of this effect highlights further nuances. As discussed, negative outcomes may especially cloud differences. Yet other studies seem to predict the opposite: After all, if people more finely notice (Hansen & Hansen, 1988), process (Hastie, 1984), and recall (Pratto & John, 1991) negative than positive information, perhaps participants should have been especially sensitive to gradations of failure ("Each unhappy family is unhappy in its own way"; Tolstoy, 1878/2014; see also Alves et al., 2017). To explain this gap, there must be certain kinds of badness that elicit other (social) dynamics. My theorizing suggests that one culprit may entail norm-based contexts evoking violations of "decent" behavior. This idea highlights a simple but crucial point: Sometimes, people may not want to acknowledge distinctions in those they judge.

Moreover, although ample research has examined how goal concreteness affects actors' motivation, little research has examined observers' roles. The present findings underscore two key points: Concrete goals may have motivational benefits, but concrete failure may have unique motivational costs. We know even less about social costs, whereby public opinion could present its own barriers to an undertaking's fate.

Practical implications

These insights reveal psychological challenges for appreciating progress. Despite large gains in quality of life over historic time (Pinker, 2018), people believe that things have gotten worse (Roser & Nagdy, 2019)—perhaps because such changes are inevitably incremental. The present findings suggest that people will quickly dismiss the idea that relative progress is progress.

Experiments 8 to 12 encourage calls for change to strategically incorporate ongoing reminders of an undertaking's efforts. For example, an Earth that warms by 2.0 °C is profoundly more habitable than one that warms by 2.5 °C (Plumer & Popovich, 2018)—but the present findings warn that people may check out when 1.5 °C hits. By lumping a 2.0 °C world and 2.5 °C world as equally futile, people risk eschewing vital goals. Improvement efforts need not be done after they fail.

The findings hint at other overlooked gradations of badness across everyday life. A student who earns 60% likely cares more than one who earns 20%—yet teachers may quit on both. A pool of rejected applicants may contain some gems—yet evaluators may barely look. Research practices may be scorned quickly for evolving slowly. A failure to invest in failures suggests routine problems for maximizing potential. It also paints a different portrait of society's alleged celebration of growth and self-improvement (Dweck, 2006): Participants were less likely to celebrate "unofficial" change, which suggests that people value what is already improved more than what is currently improving. People may be less welcoming to one's past struggles than assumed. One can imagine, for instance, that participants would view any improvements in the worst countries as especially impressive (Experiment 14); I found the opposite. The lack of lumping of passing entities suggests the same: Even after one fully reforms, one's success story may be readily overshadowed by other individuals who do even more.

Next steps

These ideas invite fruitful research on negative lumping itself. First, lumping surely sometimes reflects reality (e.g., competitions award gold, silver, and bronze—and nothing else). In everyday life, passing often is graded, whereas failing means failing; students who have already

flunked a course cannot unflunk it with additional effort, just as teachers have little incentive to give them finely grained feedback. Likewise, it might really take less effort to raise a D to a C (for example) than a B to an A (although see Experiment 12); perhaps the findings reflect learned associations as they exist. Alas, perhaps tendencies toward negative lumping created such incentive systems to begin with; even if people are better off negatively lumping in classroom contexts, it is unclear whether classrooms are better off running that way. The findings suggest that negative lumping reflects a generalized heuristic—often valid, but unwittingly overapplied (Baron, 1990). Academics might view significance of $p = .19$ and $p = .99$ as equally pointless to pursue further (McShane & Gal, 2017)—even if one has a stronger signal. It is also informative to unpack the difference between evaluative and perceptual mechanisms: Do people stubbornly dismiss categorical failures (despite privately realizing they differ) or genuinely perceive them as indistinguishable? If the effect mostly reflects the former (which I assume it does because it ebbs and flows with effort cues, holding categorical failure constant), then other problems may arise simply from people thinking one thing and doing another.

Second, the robustly observed shift toward negative lumping was typically relative itself, meaning the larger fraction of these participants did not lump failures together (e.g., note that Figs. 1–3 show that more than 50% of both-fail participants still distinguished the entities). There is ample room to further assess boundaries. Other compositions of failed outcomes may convey differential effort in ways the present designs did not capture; people may indeed discriminate failed outcomes that are inordinately far apart, or cases in which one failed outcome moved up and the other got even worse. A related question is whether threshold proximity matters. Just missing a preferred category often prompts upward comparison ("It could've been better!"; Markman et al., 1995; Medvec & Savitsky, 1997), suggesting that near-misses are less likely to be lumped with far-misses if they remind observers of their closeness to full success. However, if effort inferences matter, then near-misses may be especially dismissed ("Why didn't they finish it?").

Finally, future research should assess the generalizability of the present findings. For example, participants in most of the experiments were asked to read descriptions of the target event via an online survey, which is just one of countless ways in which such information is delivered in everyday life (including, e.g., via social discussion and more extensive news reporting). These experiments assessed (largely online) American adult participants, but participants from other cultures, contexts, and life stages might show different patterns (e.g., cultural influences on what is "close vs. far" and "success vs. failure" should moderate the effects). Future research should likewise extend to other settings altogether (e.g., lesser evilism; Kruger et al., 2009). If norm violations matter, then positive lumping may emerge for abnormally positive behaviors (e.g., someone who donates $10 billion may seem equally saintly as someone who donates $20 billion; see also Klein & Epley, 2014). Other research could unpack whether actors lump their own failures. Because actors have intimate access to their underlying intentions, my framework suggests that self-lumping is less likely. Conversely, the "what-the-hell" effect (Cochran & Tesser, 1996) suggests that goal failure often triggers self-sabotaging spirals (e.g., binging after failing to lose 10 pounds)—echoing negative lumping (e.g., if one still lost a few). Both actors and observers may encounter hidden hardships from concrete goal setting that are veiled by its benefits.

Until these possibilities are tested, the current research suggests that people may indeed see "small choice in rotten apples"—true enough in some contexts but in others leaving everyone hungry.

Appendix

Table A1. Experiments 1 and 2: Relevant Excerpts for Each Domain

Domain	Sample text
Academics	Classroom A and Classroom B are in the same school district. They have similar students, and are similar on all other basic dimensions (e.g., size of class, age of class, grades, resources). At the start of last year, both decided to try to improve their quality of learning and achievement.
Athletics	Player A and Player B have the same coaching and training staff. They play similar positions, and are similar on all other basic dimensions (e.g., gender, age, dedication, fitness). At the start of last year, both decided to try to improve their skills and become the best at their positions.

(continued)

Table A1. *(continued)*

Domain	Sample text
Habits	Student A and Student B take part in the same wellbeing classes in the community. They have similar everyday lives, and are similar on all other basic dimensions (e.g., gender, age, dedication, resources). At the start of last year, both decided to try to improve their routines and habits.
Happiness	Person A and Person B live in the same town. They have similar interests, and are similar on all other basic dimensions (e.g., gender, age, finances, free time). At the start of last year, both decided to try to improve their outlook and think happier thoughts.
Health	City A and City B are in the same state. They have similar cultures, and are similar on all other basic dimensions (e.g., population, demographics, wealth). At the start of last year, both decided to try to improve their access to health-promoting features, such as building greener spaces and making it easier for citizens to engage in exercise and healthier diets.
Personality	Patient A and Patient B are patients in the same psychology clinic. They have similar social lives, and are similar on all other basic dimensions (e.g., gender, age, finances, family background). At the start of last year, both decided to try to improve their personality and become kinder and more empathetic.
Sustainability	Organization A and Organization B are manufacturing companies. They develop similar products, and are similar on all other basic dimensions (e.g., size, revenue, networking, connections). At the start of last year, both decided to try to improve their position in terms of issues related to sustainability and environmental impact.
Technology	Platform A and Platform B are social media platforms. They have similar designs, and are similar on all other basic dimensions (e.g., content, users, traffic). At the start of last year, both decided to try to improve their position on social progress, like being more inclusive and encouraging free speech.

Note: For the full materials, see https://osf.io/q7vj9/.

Table A2. Experiment 3: Relevant Excerpts for Each Domain

Domain	Sample text
Culture	Team A and Team B have been involved with problems of culture.
Environment	Company A and Company B have been involved with problems of sustainability and being environmentally friendly.
Finances	Bank A and Bank B have been involved with problems of financial slowdown.
Harassment	Organization A and Organization B have been involved with problems of workplace harassment.
Health	City A and City B have been involved with problems of health access.
Schools	School A and School B have been involved with problems of learning and achievement.
Technology	Platform A and Platform B have been involved with problems of free speech.
Transparency	Administration A and Administration B have been involved with problems of transparency.
All domains (end of each)	They have each been equally problematic on this front. Both are now trying to improve and fix this problem moving forward. Each was informed they need to fix these issues[list of 3 issues]. . . . Ok: Time has now passed, and we can see what each organization did.
Three issues (unique to each domain)	*Culture* (change "practice structure," "workouts," "in-game structure") *Environment* (use different "lightbulbs," "recycling bins," "water stations") *Finances* (update "data system," "reporting system," "investment system") *Harassment* ("create diverse teams," "rotate leaders," "implement conduct code") *Health* (boost "text-based," "phone-based," "email-based" communication) *Schools* (improve "classroom," "lunchroom," "after-school activities") *Technology* (allow "subgroups," "symbol sharing," "cross-text") *Transparency* (increase access to "employee records," "clients," "network")

Note: For the full materials, see https://osf.io/q7vj9/.

Table A3. Stimuli Used in Experiments 6, 7, 9, 10, 11, and 12

Experiment	Stimulus (summary of relevant excerpts)
Experiment 6 (corrupt governments)	Local Government A and Local Government B preside over similar towns. They have similar populations of citizens, and they are similar on all other basic dimensions. | At the start of last year, the people called for change. Both were in need of reforming their policies to ensure better treatment of all citizens. Thus, they decided to try to improve their policies and become fairer and more just governing bodies. | They both started at the same point. Now, the year has come and gone. Here are their improvement scores.
Experiment 7 (discriminatory hiring)	Organization A and Organization B have the same problem: They need to improve their hiring practices to ensure fairer and more equitable treatment of their applicants, allowing them to hire the best possible people for the job. Thus, at the start of last year, both organizations decided to look into this issue and work on making structural changes to address it. | They both started at the same point. Now, the year has come and gone. An evaluator scores how much each organization improved. Here are their scores.
Experiment 9 (outdated technology)	Workplace A and Workplace B have the same problem: They need to update with the times. Their technologies are slow and outdated. Their scheduling systems are confusing and inefficient. They still use the same systems from years ago. It's time for some serious updating. | They both started at the same point. Now, the year has come and gone. An evaluator scores how much each workplace improved. Here are their scores.
Experiment 10 (green practices)	Company A and Company B have the same problem: They need to improve their environmental practices and do better on the sustainability front. A number of their manufacturing methods are outdated and are now believed to be contributing to climate change. A number of their everyday workplace practices are wasteful and inefficient. Companies in this industry, including these two, need to improve. | At the start of last year, both companies were tracked in terms of their improvement. They both started at the same point. Now, the year has come and gone. An evaluator scores how much each company improved. Here are their scores.
Experiment 11 (public health)	City A and City B face the same city-wide issue: They need to improve the quality of their public health resources. They need to increase access to more of their citizens. They need to update their equipment and the tools they use for communication. A number of their practices could use some fine-tuning. It's time for a change in these cities. | At the start of last year, both began being tracked by an evaluator in terms of their improvement. They both started at the same point. | Ok. Time has now passed. The evaluator scores how much each has improved to this point [manipulation: 1 month in vs. 12 months in, of a 12-month clock]. Here are their scores.
Experiment 12 (increasing diversity)	Imagine there are two companies that are identical in every way (e.g., they work on the same things; they have the same resources; they have the same kinds of workers; they're the same size; etc). | In addition, both companies are equally problematic on the front of diversity. They're both being called to increase the diversity of their workforce. They've both acknowledged the problem and have vowed to do better over the course of the year. [Threshold participants additionally saw the following text:] A town regulator is using this as a perfect opportunity to conduct an experiment on how companies work to increase diversity. Based on a random lottery, the regulator has called each company to make the following changes: Company X has been informed that they need to increase their diversity by 10% by year's end; Company Z has been informed that they need to increase their diversity by 30% by year's end. | Now, at year's end, here's what each company actually did.

Note: For the full materials, see https://osf.io/q7vj9/.

Transparency

Action Editor: Lasana Harris
Editor: Patricia J. Bauer
Author Contributions
E. O'Brien is the sole author of this article and is responsible for its content.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This research was supported by the Willard Graham and Charles E. Merrill faculty research funds at The University of Chicago Booth School of Business.
Open Practices
Data and materials for all the experiments have been made publicly available via OSF and can be accessed at https://osf.io/q7vj9/. The design and analysis plans for all experiments were preregistered on AsPredicted (copies are available at https://osf.io/q7vj9/). This article has received the

badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Ed O'Brien https://orcid.org/0000-0002-4481-8408

Acknowledgments

Kaushal Addanki, Bryan Baird, Morgan Britt, and Becky White assisted with data collection. Linda Hagen, George Wu, Nick Epley, Ann McGill, Carey Morewedge, Carol Dweck, Eugene Caruso, Stephen Spiller, Nicole Mead, Clayton Critcher, Craig Fox, Brent McFerran, Zak Tormala, Amit Kumar, and Anuj Shah provided helpful feedback on previous drafts and/or at various stages of presenting this project.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976221075302

Notes

1. Similar comparison conditions were included in all experiments.
2. Across experiments, just 6% of all observations (of more than 12,000 eligible observations) involved choosing the less-improved entity as the superior one in the pair—a number so small that no result was affected if I instead opted to exclude these responses altogether rather than do what I did, which was to group them with those who chose the more-improved entity (see the Supplemental Material available online). I grouped them because, by choosing any "victor," participants were conveying that they were willing or able to discriminate the entities rather than lump them—which was the psychology of interest.
3. The experiment was advertised for $2.00. Unbeknownst to participants (until revealed during debriefing), they would all be sent home with $3.00 as part of the study procedures.
4. That is, I measured whether participants accepted the offer to enter the bet (not which entity they would bet on). Presumably, one would enter the bet only if one were not lumping the entities (e.g., "I feel confident in a victor").
5. In this posttest (for the full details, see the Supplemental Material; for the full materials, see https://osf.io/q7vj9/), I showed 201 participants from the same population (age: M = 38.74 years, SD = 10.81; 46% women; 11% non-White; $0.50 pay) the label-present list (n = 99) or the label-absent list (n = 102) exactly as they were shown in the main experiment. I randomly drew one country from the top category and one country from the bottom category and asked participants to rate the extent to which they viewed these two countries as categorically different in terms of success versus failure (1 = *not at all*, 7 = *very*). Label-present raters indeed rated the two as categorically different (M = 5.69, SD = 1.01)—versus the scale midpoint: $t(98) = 16.67$, $p < .001$, $d = 1.68$—which is unsurprising because these participants saw explicitly labeled different categories for each country. Yet label-absent participants also viewed the two as categorically different (M = 5.32, SD = 1.20)—versus the scale midpoint: $t(101) = 11.18$, $p < .001$, $d = 1.11$—even though these participants saw no explicit category cutoffs.

References

Aengenheyster, M., Feng, Q. Y., van der Ploeg, F., & Dijkstra, H. A. (2018). The point of no return for climate action: Effects of climate uncertainty and risk tolerance. *Earth System Dynamics*, *9*, 1085–1095.

Alves, H., Koch, A., & Unklebach, C. (2017). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences*, *21*, 69–79.

Bandura, A., & Simon, K. M. (1977). The role of proximal intentions in self-regulation of refractory behavior. *Cognitive Therapy and Research*, *1*, 177–193.

Baron, J. (1990). Harmful heuristics and the improvement of thinking. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills* (pp. 28–47). Karger.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323–370.

Burck, J., Hagen, U., Höhne, N., Nascimento, L., & Bals, C. (2019). *Climate Change Performance Index 2020: Results*. Germanwatch. https://ccpi.org/download/the-climate-change-performance-index-2020/

Cochran, W., & Tesser, A. (1996). The "what the hell" effect: Some effects of goal proximity and goal framing on performance. In L. L. Martin & A. Tesser (Eds.), *Striving and feeling: Interactions among goals, affect, and self-regulation* (pp. 99–120). Erlbaum.

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., Carlton, J. S., Lewandowsky, S., Skuce, A. G., Green, S. A., Nuccitelli, D., Jacobs, P., Richardson, M., Winkler, B., Painting, R., & Rice, K. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4), Article 048002. https://doi.org/10.1088/1748-9326/11/4/048002

Dawes, R. A., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Ballantine Books.

Eskreis-Winkler, L., & Fishbach, A. (2019). Not learning from failure—the greatest failure of all. *Psychological Science*, *30*, 1733–1744. https://doi.org/10.1177/0956797619881133

Eskreis-Winkler, L., & Fishbach, A. (2020). Hidden failures. *Organizational Behavior and Human Decision Processes*, *157*, 57–67.

Ferguson, S. J. (2020). *Race, gender, sexuality, and social class: Dimensions of inequality and identity*. SAGE.

Fischoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge University Press.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). McGraw-Hill.

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences, USA*, *114*, 6068–6073.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*, 493–503.

Gollwitzer, P. M., & Oettingen, G. (2011). Planning promotes goal striving. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 223–246). Guilford Press.

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology*, *54*, 917–924.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology*, *46*, 44–56.

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, *38*, 79–109.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General*, *143*, 2366–2379.

Klein, N., & O'Brien, E. (2016). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social Cognition*, *34*, 149–166.

Klein, N., & O'Brien, E. (2017). The power and limits of personal change: When a bad past does (and does not) inspire in the present. *Journal of Personality and Social Psychology*, *113*, 210–229.

Klein, N., & O'Brien, E. (2018). People use less information than they think to make up their minds. *Proceedings of the National Academy of Sciences, USA*, *115*, 13222–13227.

Kruger, J., Burrus, J., & Kressel, L. M. (2009). Between a rock and a hard place: Damned if you do, damned if you don't. *Journal of Experimental Social Psychology*, *45*, 1286–1290.

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*, 91–98.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal systems. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 331–378). Academic Press.

Ledgerwood, A., & Boydstun, A. E. (2014). Sticky prospects: Loss frames are cognitively stickier than gain frames. *Journal of Experimental Psychology: General*, *143*, 376–385.

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (Ed.), *Personality and the behavior disorders* (pp. 333–378). Ronald Press.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Prentice Hall.

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1995). The impact of perceived control on the imagination of better and worse possible worlds. *Personality and Social Psychology Bulletin*, *21*, 588–595.

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, *112*, 885–895.

Medvec, V. H., & Savitsky, K. (1997). When doing better means feeling worse: The effects of categorical cutoff points on counterfactual thinking and satisfaction. *Journal of Personality and Social Psychology*, *72*, 1284–1296.

Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966-1984. *Organizational Behavior and Human Decision Processes*, *39*, 52–83.

Morales, A. C. (2005). Giving firms an "e" for effort: Consumer responses to high-effort firms. *Journal of Consumer Research*, *31*, 806–812.

O'Brien, E. (2020). When small signs of change add up: The psychology of tipping points. *Current Directions in Psychological Science*, *29*(1), 55–62. https://doi.org/10.1177/0963721419884313

O'Brien, E., & Klein, N. (2017). The tipping point of perceived change: Asymmetric thresholds in diagnosing improvement versus decline. *Journal of Personality and Social Psychology*, *112*, 161–185.

Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin Books.

Plumer, B., & Popovich, N. (2018, October 7). Why half a degree of global warming is a big deal. *The New York Times*. https://www.nytimes.com/interactive/2018/10/07/climate/ipcc-report-half-degree.html

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*, 380–391.

Roser, M., & Nagdy, M. (2019). *Optimism and pessimism*. Our World in Data. http://www.ourworldindata.org/optimism-pessimism

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296–320.

Shakespeare, W. (2003). *The taming of the shrew* (E. Schafer, Ed.). Cambridge University Press. (Original work published 1593).

Shampanier, K., Mažar, N., & Ariely, D. (2007). Zero as a special price: The true value of free products. *Marketing Science*, *26*, 742–757.

Simon, H. A. (1979). Rational decision making in business organizations. *American Economic Review*, *69*, 493–513.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin*, *129*, 475–494.

Soman, D., & Cheema, A. (2004). When goals are counterproductive: The effects of violation of a behavioral goal on subsequent performance. *Journal of Consumer Research*, *31*, 52–62.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgment. *British Journal of Psychology*, *54*, 101–114.

Tolstoy, L. (2014). *Anna Karenina* (R. Bartlett, Trans.). Oxford University Press. (Original work published 1878).

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*, 548–573.

Weiner, B. (1995). *Judgments of responsibility*. Guilford Press.

Supplemental Online Materials for:

Losing Sight of Piecemeal Progress: People Lump and Dismiss

Improvement Efforts That Fall Short of Categorical Change—Despite Improving

---

Experiment 1

1. Remaining output of main results: Main effect of Score Pair, Wald = 6.23, df = 1, p = .013; Main effect of Domain: Wald = 77.38, df = 7, $p < .001$; Threshold by Domain: Wald = 24.36, df = 7, p = .001; Score Pair by Domain: Wald = 9.34, df = 7, p = .229).

2. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, and confidence, and demographics: Main effect of Threshold, Wald = 34.94, df = 1, $p < .001$, and its critical 2-way interaction with Score Pair, Wald = 15.02, df = 1, $p < .001$; null 3-way interaction with Domain, Wald = 7.73, df = 7, p = .357.

3. All observations (N = 3,208 choices): 73% superior, 26% both the same, 1% inferior.

4. For full individual figures of each Domain, see Figure S1.

5. Confusion:
    a. Main effect of Threshold, $F(1, 397) = 2.03$, p = .155.
    b. Main effect of Score Pair: Both-Low participants (M = 1.90, SD = 1.44) vs. Both-High participants (M = 1.46, SD = 0.99), $F(1, 397) = 12.56$, $p < .001$.
    c. Threshold*Score Pair, $F(1, 397) = 0.05$, p = .817.

6. Confidence:
    a. Main effect of Threshold, $F(1, 397) = 0.73$, p = .393.
    b. Main effect of Score Pair: Both-Low participants (M = 5.56, SD = 1.36) vs. Both-High participants (M = 5.93, SD = 1.11), $F(1, 397) = 9.17$, p = .003.
    c. Threshold*Score Pair, $F(1, 397) = 0.68$, p = .409.

7. Engagement:
    a. Main effect of Threshold: $F(1, 397) = 0.51$, p = .476.
    b. Main effect of Score Pair: $F(1, 397) = 0.24$, p = .623.
    c. Threshold*Score Pair: $F(1, 397) = 0.95$, p = .330.

Experiment 2

1. Remaining output of main results: Main effect of Domain: Wald = 29.15, df = 7, $p <$ .001).

2. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 19.02, df = 2, $p < .001$; interaction with Domain, Wald = 22.02, df = 14, $p = .078$).

3. All observations ($N = 2{,}408$ choices): 66% superior, 30% both the same, 4% inferior.

4. For full individual figures of each Domain, see Figure S2.

5. Study confusion was low and did not differ by condition (overall: $M = 1.92$, $SD = 1.63$; Framing effect, $F(2, 298) = 0.73$, $p = .481$); engagement was high and did not differ by condition (overall: $M = 4.91$, $SD = 1.72$; Framing effect, $F(2, 298) = 0.01$, $p = .990$); and confidence was high and did not differ by condition (overall: $M = 5.98$, $SD = 1.23$; Framing effect, $F(2, 298) = 0.40$, $p = .672$).

Experiment 3

1. Remaining output of main results: Main effect of Domain: Wald = 31.96, df = 7, $p < .001$).

2. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Main effect of Framing, Wald = 43.77, df = 2, $p < .001$; interaction with Domain, Wald = 22.17, df = 14, $p = .075$).

3. All observations ($N = 4{,}904$ choices): 70% superior, 21% both the same, 9% inferior.

4. For full individual figures of each Domain, see Figure S3.

5. Study confusion was low and did not differ by condition (overall: $M = 2.05$, $SD = 1.67$; Framing effect, $F(2, 612) = 0.23$, $p = .794$); engagement was high and did not differ by condition (overall: $M = 4.96$, $SD = 1.69$; Framing effect, $F(2, 612) = 1.69$, $p = .185$); and confidence was high and did not differ by condition (overall: $M = 5.98$, $SD = 1.17$; Framing effect, $F(2, 612) = 1.27$, $p = .282$).

Experiment 4

1. We requested 1,500 participants from Amazon's Mechanical Turk, yielding 1,504 ($M_{age} = 38.94$, $SD_{age} = 12.27$; 47% women; 26% non-White) who participated for $0.25.

2. Pairwise results collapsing across Threshold Type: Both Fail (44%) uniquely-strongly dipped compared to Control (70%) and Both Pass (60%)
    a. Both Fail vs. Control: Wald = 69.10, $p < .001$

   b. Both Fail vs. Both Pass: Wald = 67.59, $p < .001$
   c. Control vs. Both Pass: Wald = 11.38, $p = .001$

3. Most participants passed the attention check regarding whether they had seen threshold information (93%).

4. Remaining output: Main effect of Threshold Type: Wald = 0.83, df = 1, $p = .362$.

5. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 24.95, df = 1, $p < .001$; interaction with Threshold Type, Wald = 3.14, df = 1, $p = .076$).

6. All observations (N = 1,504 choices): 55% superior, 42% both the same, 3% inferior.

7. For full individual figures of each Threshold Type, see Figure S4.

8. Study confusion did not differ by condition (overall: $M = 1.86$, $SD = 1.36$; all ps $\geq .154$); nor did engagement (overall: $M = 4.23$, $SD = 1.91$; all ps $\geq .184$); nor did confidence (overall: $M = 5.66$, $SD = 1.29$; all ps $\geq .052$).

Experiment 5

1. We requested 1,500 participants from Amazon's Mechanical Turk, yielding 1,507 (M_{age} = 38.58, SD_{age} = 12.37; 49% women; 24% non-White) who participated for $0.25.

2. Pairwise results collapsing across Choice Type: Both Fail (51%) uniquely-strongly dipped compared to Control (77%) and Both Pass (69%)
   a. Both Fail vs. Control: Wald = 71.07, $p < .001$
   b. Both Fail vs. Both Pass: Wald = 30.59, $p < .001$
   c. Control vs. Both Pass: Wald = 10.19, $p = .001$

3. Most participants passed the attention check regarding whether they had seen threshold information (93%). This study also included a second attention check regarding which Choice Type participants saw; most passed (81%).

4. Remaining output: Main effect of Choice Type: Wald = 11.71, df = 1, $p = .001$.

5. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 26.65, df = 1, $p < .001$; interaction with Choice Type, Wald = 0.64, df = 1, $p = .424$).

6. All observations (N = 1,507 choices): 63% superior, 35% both the same, 2% inferior.

7. For full individual figures of each Choice Type, see Figure S5.

8. Study confusion (overall: $M = 1.86$, $SD = 1.36$) happened to differ by condition such that Control was less confusing ($M = 1.93$, $SD = 1.39$) than either Both Fail ($M = 2.17$, $SD = 1.52$) or Both Pass ($M = 2.24$, $SD = 1.60$), $p = .002$ (all other ps $\geq .415$). Confidence (overall: $M = 5.12$, $SD = 1.46$) showed a parallel effect, such that Control was more confident ($M = 5.27$, $SD = 1.43$) than either Both Fail ($M = 5.06$, $SD = 1.43$) or Both Pass ($M = 5.02$, $SD = 1.50$), $p = .011$ (all other ps $\geq .059$). There were no differences in engagement (overall: $M = 4.18$, $SD = 1.85$; all ps $\geq .123$).

9. Unique to this experiment, simply because we were curious, participants also completed open-ended exploratory measures regarding their estimates of the lowest possible score of the scale (overall: $M = 4.93$ points, $SD = 16.71$ points), the highest possible score of the scale (overall: $M = 99.65$ points, $SD = 130.87$ points), the total sample size of all entities that the evaluator had assessed (overall: $M = 1{,}898.46$ entities, $SD = 36{,}797.64$ entities), and the average improvement score of this total number of entities (overall: $M = 50.81$ points, $SD = 25.88$ points). All of these responses are retained in our data file.

Experiment 6

1. We requested 1,500 participants from Amazon's Mechanical Turk, yielding 1,514 ($M_{age} = 38.74$, $SD_{age} = 12.36$; 48% women; 24% non-White) who participated for $0.25.

2. Pairwise results collapsing across Choice Type: Both Fail (61%) uniquely-strongly dipped compared to Control (85%) and Both Pass (74%)
    a. Both Fail vs. Control: Wald = 70.71, $p < .001$
    b. Both Fail vs. Both Pass: Wald = 20.20, $p < .001$
    c. Control vs. Both Pass: Wald = 18.28, $p < .001$

3. Most participants passed the attention check regarding whether they had seen threshold information (92%). This study also included a second attention check regarding which Choice Type participants saw; most passed (79%).

4. Remaining output: Main effect of Choice Type: Wald = 0.24, df = 1, $p = .626$.

5. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 26.84, df = 1, $p < .001$; interaction with Choice Type, Wald = 0.01, df = 1, $p = .908$).

6. All observations ($N = 1{,}507$ choices): 64% superior, 27% both the same, 9% inferior.

7. For full individual figures of each Choice Type, see Figure S6.

8. Study confusion (overall: $M = 2.06$, $SD = 1.58$) happened to differ by condition such that Both Fail was less confusing ($M = 1.78$, $SD = 1.45$) than either Control ($M = 2.17$, $SD = 1.61$) or Both Pass ($M = 2.24$, $SD = 1.65$), $p < .001$ (all other ps $\geq .072$). Confidence (overall: $M = 5.79$, $SD = 1.23$) showed a parallel effect, such that Both Fail was more confident ($M = 5.96$, $SD = 1.13$) than either Control ($M = 5.65$, $SD = 1.33$) or Both Pass ($M = 5.76$, $SD = 1.21$), $p < .001$ (all other ps $\geq .087$). There were no differences in engagement (overall: $M = 4.41$, $SD = 1.83$; all ps $\geq .194$).

Experiment 7

1. We requested 600 participants from Amazon's Mechanical Turk, yielding 605 ($M_{age} = 39.07$, $SD_{age} = 12.08$; 47% women; 25% non-White) who participated for $0.25.

2. Pairwise results collapsing across Direction: Both Fail (68%) uniquely-strongly dipped compared to Control (90%) and Both Pass (84%)
    a. Both Fail vs. Control: Wald = 27.19, $p < .001$
    b. Both Fail vs. Both Pass: Wald = 14.32, $p < .001$
    c. Control vs. Both Pass: Wald = 3.03, $p = .082$

3. Most participants passed the attention check regarding whether they had seen threshold information (83%). This study also included a second attention check regarding which Choice Type participants saw; most passed (87%).

4. Remaining output: Main effect of Direction: Wald = 1.29, df = 1, $p = .255$.

5. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 22.19, df = 1, $p < .001$; interaction with Direction, Wald = 0.37, df = 1, $p = .541$).

6. All observations ($N = 605$ choices): 69% superior, 20% both the same, 11% inferior.

7. For full individual figures of each Direction, see Figure S7.

8. Study confusion did not differ by condition (overall: $M = 2.09$, $SD = 1.57$; all ps $\geq .060$); nor did engagement (overall: $M = 3.97$, $SD = 1.88$; all ps $\geq .334$); nor did confidence (overall: $M = 5.84$, $SD = 1.17$; all ps $\geq .242$).

Experiment 8

1. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect on Favorability, $F(2, 267) = 27.71$, $p < .001$, and on Effort, $F(2, 267) = 21.34$, $p < .001$; mediation, $b = -1.14$, SE = .18, 95% CI_{Boot} [-1.52, -0.80]).

2. Study confusion was low and did not differ by condition (overall: $M = 2.04$, $SD = 1.53$; Framing effect, $F(2, 303) = 0.35$, $p = .704$); engagement was high and did not differ by condition (overall: $M = 4.35$, $SD = 1.81$; Framing effect, $F(2, 303) = 2.21$, $p = .111$); and confidence was high and did not differ by condition (overall: $M = 5.77$, $SD = 1.18$; Framing effect, $F(2, 303) = 0.60$, $p = .552$).

3. Forced-choice (Favorability), overall effect of Framing, Wald = 11.04, df = 1, $p = .001$: Most Control (74%) and Both Pass (70%) participants simply chose the more-improved organization as the superior one (Wald = 0.39, $p = .534$)—yet Both Fail participants were less likely to share this view (48%), instead lumping them as "all the same" (Both Fail vs. Control: Wald = 14.74, $p < .001$; Both Fail vs. Both Pass: Wald = 10.23, $p = .001$).
    a. All observations ($N = 304$ choices): 60% superior, 36% both the same, 4% inferior.

4. Forced-choice (Perceived Effort), overall effect of Framing, Wald = 15.96, df = 1, $p = .001$: Most Control (76%) and Both Pass (81%) participants simply chose the more-improved organization as the superior one (Wald = 0.89, $p = .345$)—yet Both Fail participants were less likely to share this view (55%), instead lumping them as "all the same" (Both Fail vs. Control: Wald = 9.53, $p = .002$; Both Fail vs. Both Pass: Wald = 14.88, $p < .001$).
    a. All observations ($N = 304$ choices): 68% superior, 29% both the same, 3% inferior.

Experiment 9

1. We requested 900 participants from Amazon's Mechanical Turk, yielding 908 ($M_{age} = 37.93$, $SD_{age} = 12.21$; 56% women; 24% non-White) who participated for $0.25.

2. Pairwise results collapsing across Difficulty: Both Fail (57%) uniquely-strongly dipped compared to Control (82%) and Both Pass (78%)
    a. Both Fail vs. Control: Wald = 40.63, $p < .001$
    b. Both Fail vs. Both Pass: Wald = 26.97, $p < .001$
    c. Control vs. Both Pass: Wald = 1.87, $p = .172$

3. Critically, however, the significant interaction indicates these patterns vary depending on level of Difficulty. For full individual figures of each level of Difficulty, see Figure S8. The relevant statistical output is:

    a. No Information (negative lumping replicates):
        i. Overall effect of Framing: Wald = 21.51, $p < .001$
            1. Both Fail vs. Control: Wald = 16.49, $p < .001$
            2. Both Fail vs. Both Pass: Wald = 13.55, $p < .001$
            3. Control vs. Both Pass: Wald = 0.18, $p = .670$

b. Low Difficulty (negative lumping replicates):
    i. Overall effect of Framing: Wald = 29.28, $p < .001$
        1. Both Fail vs. Control: Wald = 28.40, $p < .001$
        2. Both Fail vs. Both Pass: Wald = 13.42, $p < .001$
        3. Control vs. Both Pass: Wald = 4.43, $p = .035$

c. High Difficulty (**negative lumping is uniquely attenuated*):
    i. Overall effect of Framing: Wald = 3.58, $p = .059$
        1. Both Fail vs. Control: Wald = 2.18, $p = .140$
        2. Both Fail vs. Both Pass: Wald = 2.99, $p = .084$
        3. Control vs. Both Pass: Wald = 0.71, $p = .790$

4. Most participants passed the attention check regarding whether they had seen threshold information (92%). This study also included a second attention check regarding which level of Difficulty participants saw; most passed (66%).

5. Remaining output: Main effect of Difficulty: Wald = 339, df = 1, $p = .065$.

6. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 40.49, df = 1, $p < .001$; interaction with Difficulty, Wald = 5.90, df = 1, $p = .015$).

7. All observations ($N = 908$ choices): 70% superior, 28% both the same, 2% inferior.

8. Study confusion (overall: $M = 1.62$, $SD = 1.11$) happened to differ by condition such that Both Fail was less confusing ($M = 1.49$, $SD = 0.98$) than either Control ($M = 1.63$, $SD = 1.12$) or Both Pass ($M = 1.74$, $SD = 1.22$), $p = .017$ (all other ps $\geq .091$). There were no differences in confidence (overall: $M = 5.84$, $SD = 1.19$; all ps $\geq .140$) or engagement (overall: $M = 4.23$, $SD = 1.89$; all ps $\geq .419$).

Experiment 10

1. We requested 900 participants from Amazon's Mechanical Turk, yielding 912 ($M_{age} = 37.14$, $SD_{age} = 11.81$; 46% women; 26% non-White) who participated for $0.25.

2. Pairwise results collapsing across Knowledge: Both Fail (63%) uniquely-strongly dipped compared to Control (83%) and Both Pass (73%)
    a. Both Fail vs. Control: Wald = 31.86, $p < .001$
    b. Both Fail vs. Both Pass: Wald = 8.21, $p = .004$
    c. Control vs. Both Pass: Wald = 8.81, $p = .003$

3. Critically, however, the significant interaction indicates these patterns vary depending on level of Knowledge. For full individual figures of each level of Knowledge, see Figure S9. The relevant statistical output is:

a. No Information (negative lumping replicates):
    i. Overall effect of Framing: Wald = 4.37, $p = .037$
        1. Both Fail vs. Control: Wald = 12.41, $p < .001$
        2. Both Fail vs. Both Pass: Wald = 3.94, $p = .047$
        3. Control vs. Both Pass: Wald = 3.12, $p = .077$

b. High Salience (negative lumping replicates):
    i. Overall effect of Framing: Wald = 10.97, $p = .001$
        1. Both Fail vs. Control: Wald = 9.46, $p = .002$
        2. Both Fail vs. Both Pass: Wald = 10.29, $p = .001$
        3. Control vs. Both Pass: Wald = 0.03, $p = .854$

c. Low Salience (**negative lumping is uniquely attenuated*):
    i. Overall effect of Framing: Wald = 0.15, $p = .694$
        1. Both Fail vs. Control: Wald = 10.35, $p = .001$
        2. Both Fail vs. Both Pass: Wald = 0.15, $p = .696$
        3. Control vs. Both Pass: Wald = 39.10, $p < .001$

4. Most participants passed the attention check regarding whether they had seen threshold information (89%). This study also included a second attention check regarding which level of Knowledge participants saw; most passed (62%).

5. Remaining output: Main effect of Knowledge: Wald < 0.00, df = 1, $p = .998$.

6. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 9.91, df = 1, $p = .002$; interaction with Knowledge, Wald = 6.79, df = 1, $p = .009$).

7. All observations ($N = 912$ choices): 68% superior, 27% both the same, 5% inferior.

8. Study confusion did not differ by condition (overall: $M = 1.88$, $SD = 1.48$; all ps $\geq .269$); nor did confidence (overall: $M = 5.79$, $SD = 1.20$; all ps $\geq .478$). Engagement (overall: $M = 4.24$, $SD = 1.81$) happened to differ by condition such that High Salience was less engaging ($M = 4.04$, $SD = 1.89$) than either Control ($M = 4.27$, $SD = 1.81$) or Low Salience ($M = 4.40$, $SD = 1.72$), $p = .042$ (all other ps $\geq .381$).

Experiment 11

1. We requested 600 participants from Amazon's Mechanical Turk, yielding 601 ($M_{age} = 38.75$, $SD_{age} = 12.52$; 52% women; 27% non-White) who participated for $0.25.

2. Pairwise results collapsing across Window: Both Fail (67%) uniquely-strongly dipped compared to Control (88%) and Both Pass (80%)

a. Both Fail vs. Control: Wald = 23.24, $p < .001$
b. Both Fail vs. Both Pass: Wald = 7.60, $p = .006$
c. Control vs. Both Pass: Wald = 5.29, $p = .021$

3. Critically, however, the significant interaction indicates these patterns vary depending on level of Window. For full individual figures of each level of Window, see Figure S10. The relevant statistical output is:

    a. One Year Into One-Year Window (negative lumping replicates):
        i. Overall effect of Framing: Wald = 27.26, $p < .001$
            1. Both Fail vs. Control: Wald = 21.55, $p < .001$
            2. Both Fail vs. Both Pass: Wald = 14.53, $p < .001$
            3. Control vs. Both Pass: Wald = 1.44, $p = .231$

    b. One Month Into One-Year Window (**negative lumping is uniquely attenuated*):
        i. Overall effect of Framing: Wald = 1.43, $p = .232$
            1. Both Fail vs. Control: Wald = 4.52, $p = .032$
            2. Both Fail vs. Both Pass: Wald < 0.00, $p = 1.00$
            3. Control vs. Both Pass: Wald = 4.52, $p = .033$.

4. Most participants passed the attention check regarding whether they had seen threshold information (87%). This study also included a second attention check regarding which level of Window participants saw; most passed (92%).

5. Remaining output: Main effect of Window: Wald = 0.82, df = 1, $p = .365$.

6. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (Framing effect, Wald = 15.22, df = 1, $p < .001$; interaction with Window, Wald = 8.29, df = 1, $p = .004$).

7. All observations ($N = 601$ choices): 74% superior, 22% both the same, 4% inferior.

8. Study confusion did not differ by condition (overall: $M = 2.15$, $SD = 1.52$; all ps $\geq .071$); nor did engagement (overall: $M = 4.24$, $SD = 1.82$; all ps $\geq .608$); nor did confidence (overall: $M = 5.59$, $SD = 1.28$; all ps $\geq .153$).

Experiment 12

1. We requested 200 "Cloud Approved" participants from Cloud Research, yielding 201 ($M_{age} = 38.66$, $SD_{age} = 11.80$; 39% women; 25% non-White) who participated for $0.50.

2. Most participants passed the attention check regarding whether they had seen threshold information (95%). This study also included a second attention check regarding how much progress each company actually made by year's end; most passed (92%).

3. All main results hold when re-running our analyses excluding attention-check failures, and controlling for confusion, engagement, confidence, and demographics (effect of condition on choice, $F(1, 168) = 64.26$, $p < .001$). Note that this re-conducted analysis reflects a Univariate GLM rather than Binary Logistic Regression (after our attention-check exclusions, there were zero Control participants who chose a "winner," thereby leaving an empty cell that renders Binary Logistic Regression unable to be conducted).

4. All observations ($N = 201$ choices): 20% surpass-is-superior, 77% both the same, 3% fall-short-is-superior.

5. Study confusion was low and did not differ by condition (overall: $M = 1.26$, $SD = 0.70$; $t(199) = 1.33$, $p = .185$); engagement was relatively high and did not differ by condition (overall: $M = 4.12$, $SD = 1.95$; $t(199) = -1.21$, $p = .228$); and confidence was high and did not differ by condition (overall: $M = 6.34$, $SD = 1.02$; $t(199) = 0.62$, $p = .537$).

Experiment 13

1. All main results hold when re-running our analyses excluding attention-check failures, and when controlling for confusion and demographics (Framing effect: Wald = 9.69, $p = .002$).

2. Study confusion was low across conditions, although we did happen to observe an omnibus effect of Framing such that Both Fail participants reported higher confusion ($M = 2.49$, $SD = 1.60$) than Control participants ($M = 2.02$, $SD = 1.25$) and Both Pass participants ($M = 2.13$, $SD = 1.30$), $F(2, 284) = 3.05$, $p = .049$.

Experiment 14

1. All main results hold when re-running our analyses when controlling for additional variables (critical main effect of Category Pair, $F(1, 620) = 5.16$, $p = .023$; null interaction with Category Label, $F(1, 620) = 0.14$, $p = .710$).

2. Study confusion was low and did not differ by condition (overall: $M = 2.28$, $SD = 1.57$; $t(626) = 0.15$, $p = .884$); engagement was relatively high and did not differ by condition (overall: $M = 3.90$, $SD = 2.05$; $t(626) = 0.04$, $p = .969$); and confidence was high and did not differ by condition (overall: $M = 4.69$, $SD = 1.61$; $t(626) = 1.54$, $p = .123$).

Experiment 14 (Post-test)

1. We requested 200 participants from Amazon's Mechanical Turk, yielding 201 ($M_{age} = 38.74$, $SD_{age} = 10.81$; 46% women; 11% non-White) who participated for $0.50.

2. Phrasing of item: *To what extent do you view these 2 countries as categorically "different" from each other, based on where they rank on this list? (Meaning: Does one country strike you as fundamentally better/worse than the other country, as if these 2 countries belong on different "good" vs. "bad" lists altogether?* (1 = *these country ranks are not at all different in this way*; 7 = *these country ranks are very different in this way*).

3. One country was randomly drawn from ranks 1-14 (top category) and the other country was randomly drawn from ranks 44-57 (bottom category).

4. The only other item that participants completed was an attention check regarding whether they had seen threshold information; most passed (78%), and the effect holds when re-running the analysis while excluding attention-check failures (Label-Present participants vs. scale midpoint: $t(78) = 17.47$, $p < .001$, $d = 1.97$; Label-Absent participants vs. scale midpoint, $t(76) = 9.54$, $p < .001$, $d = 1.09$).

Figure S1. Experiment 1: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 8 = weakest.

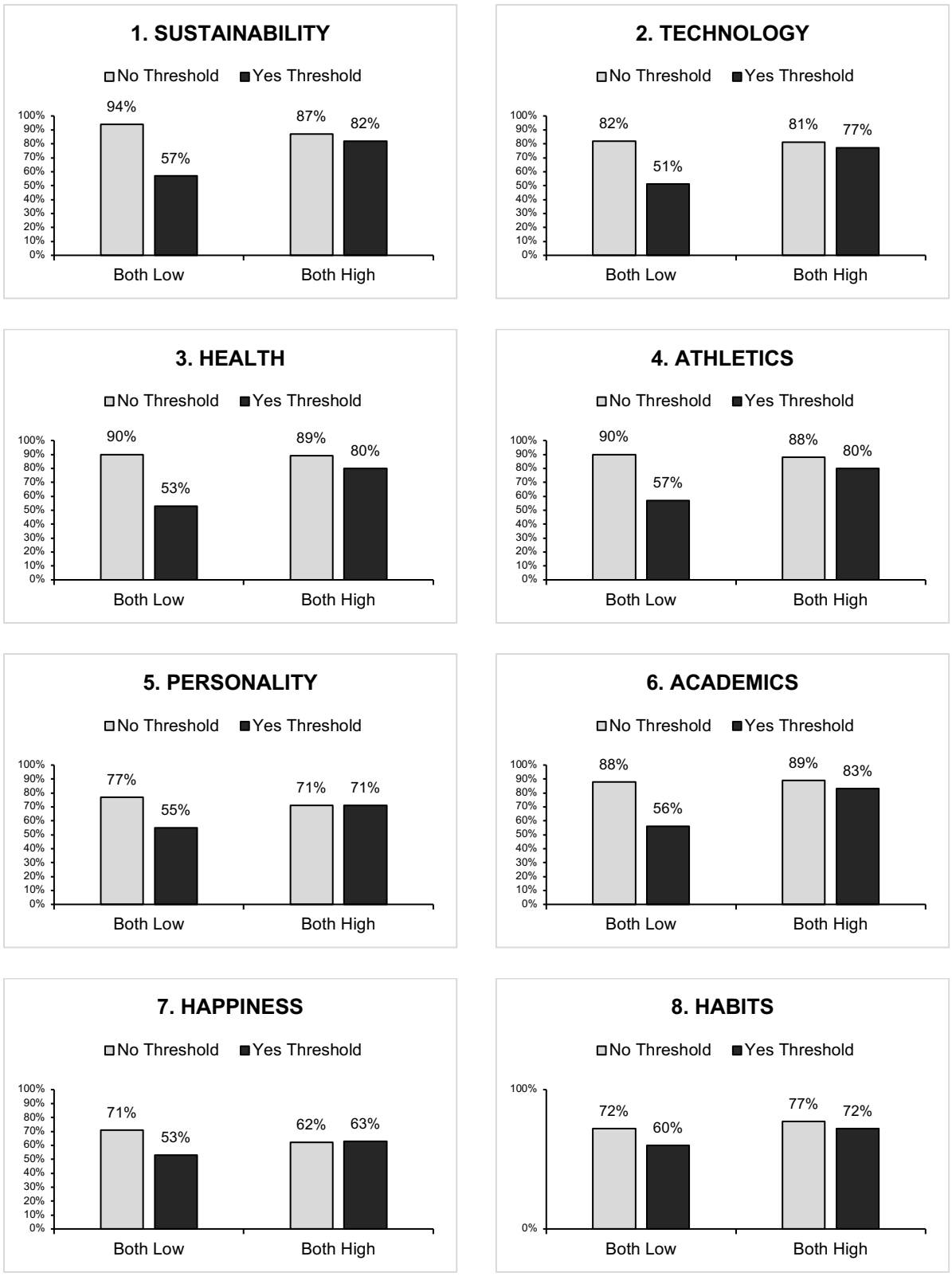


Figure S2. Experiment 2: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 8 = weakest.

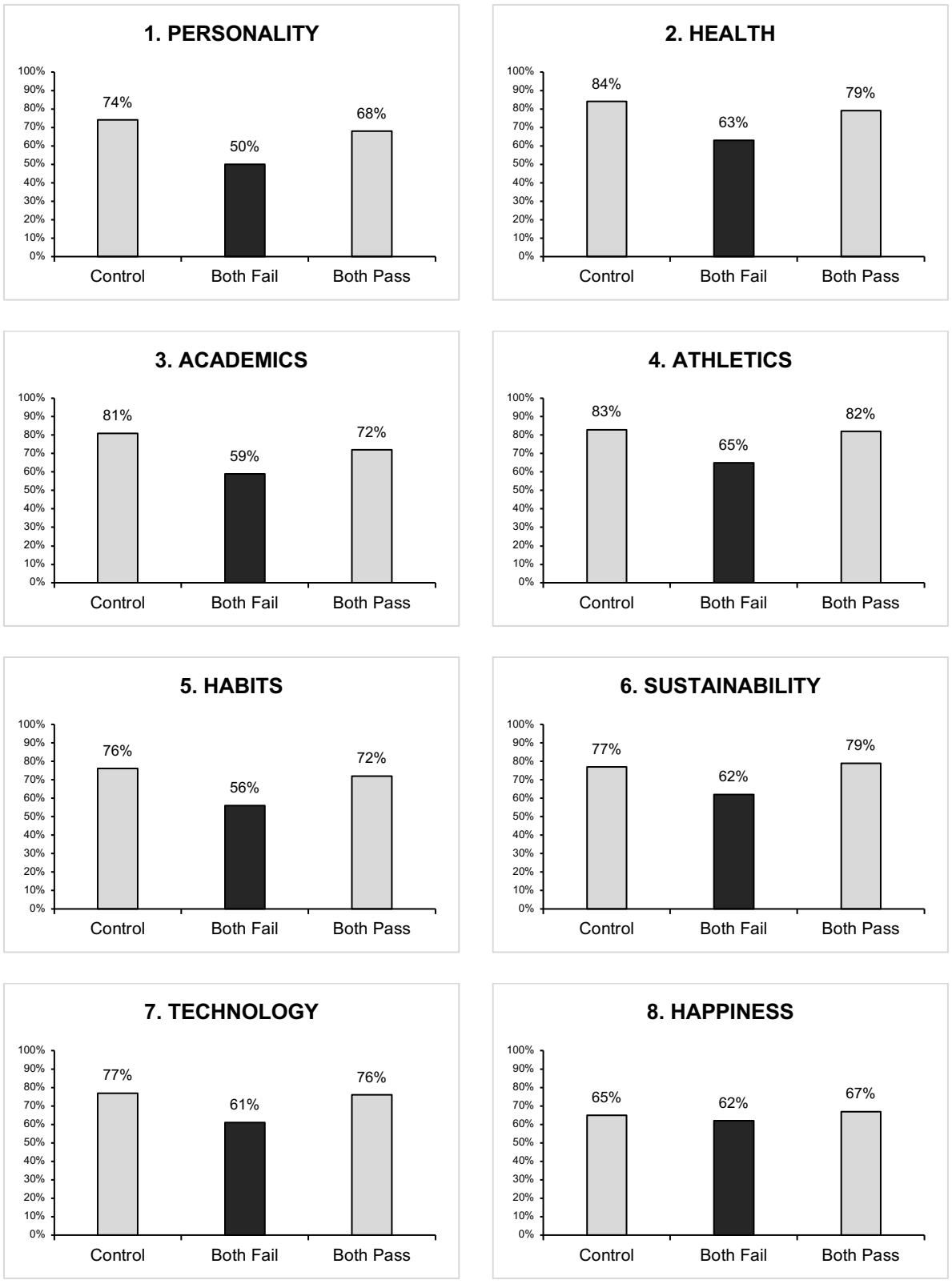


Figure S3. Experiment 3: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 8 = weakest.

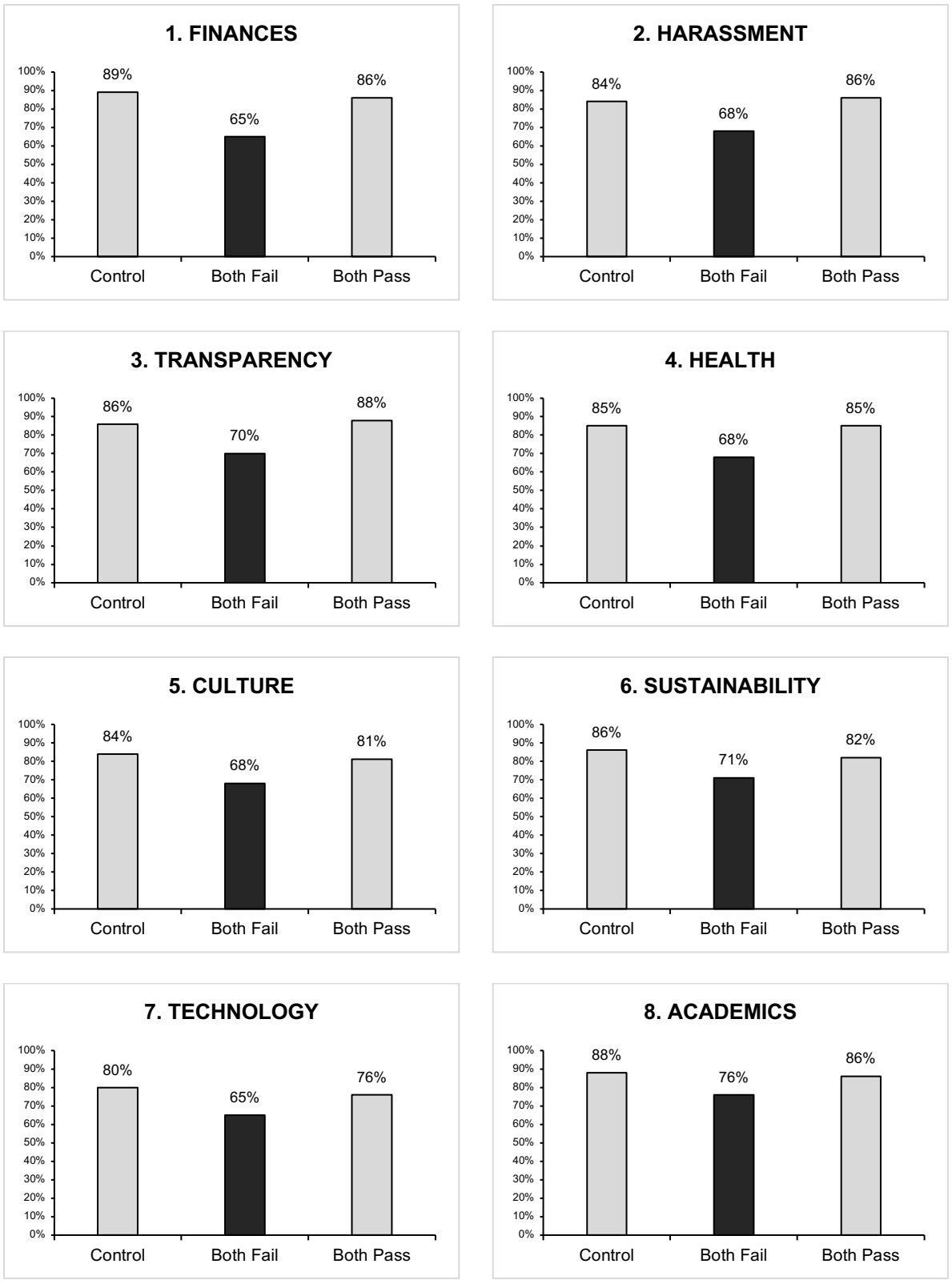


Figure S4. Experiment 4: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 5 = weakest.

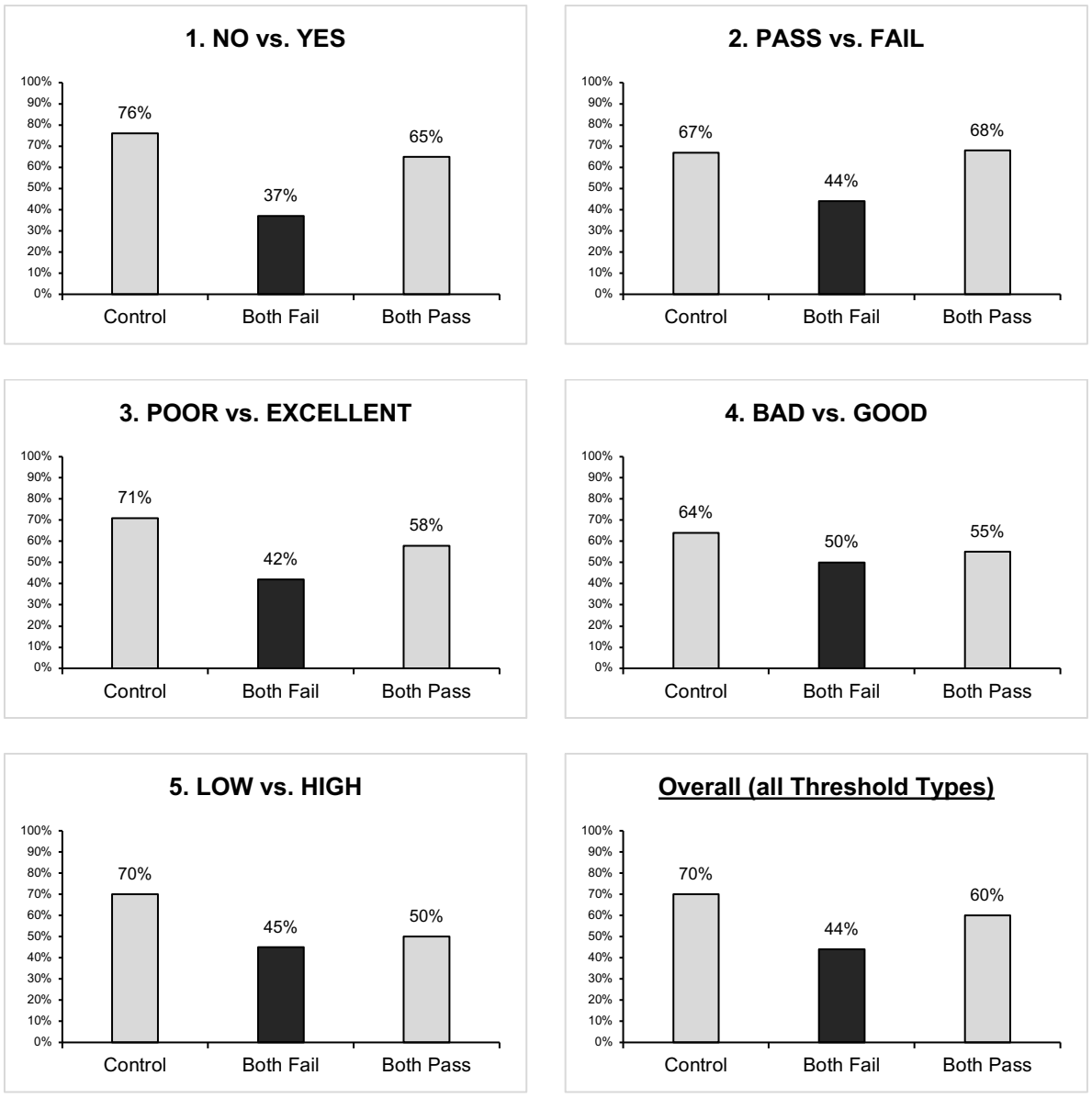


Figure S5. Experiment 5: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 5 = weakest.

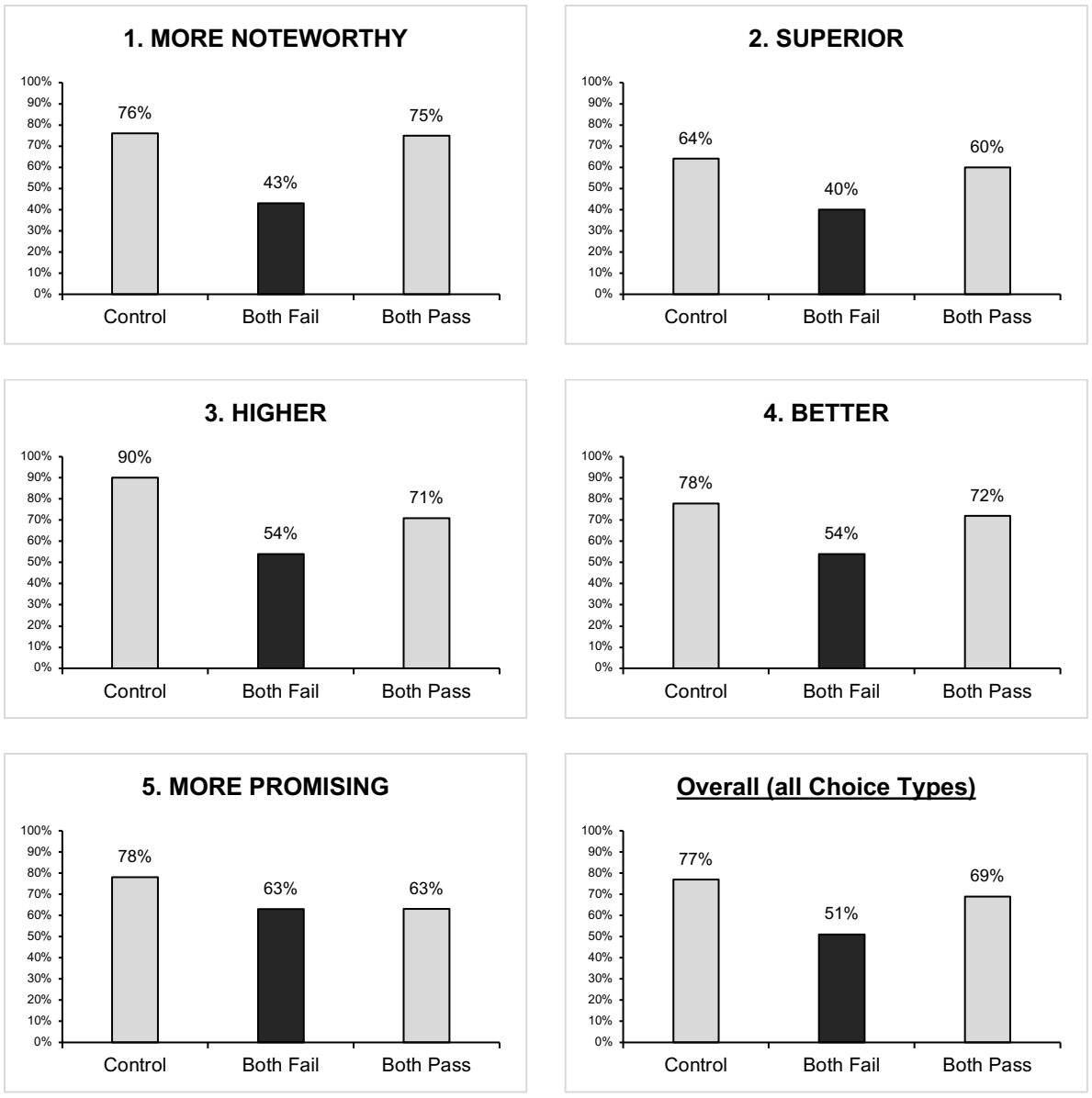


Figure S6. Experiment 6: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 5 = weakest.

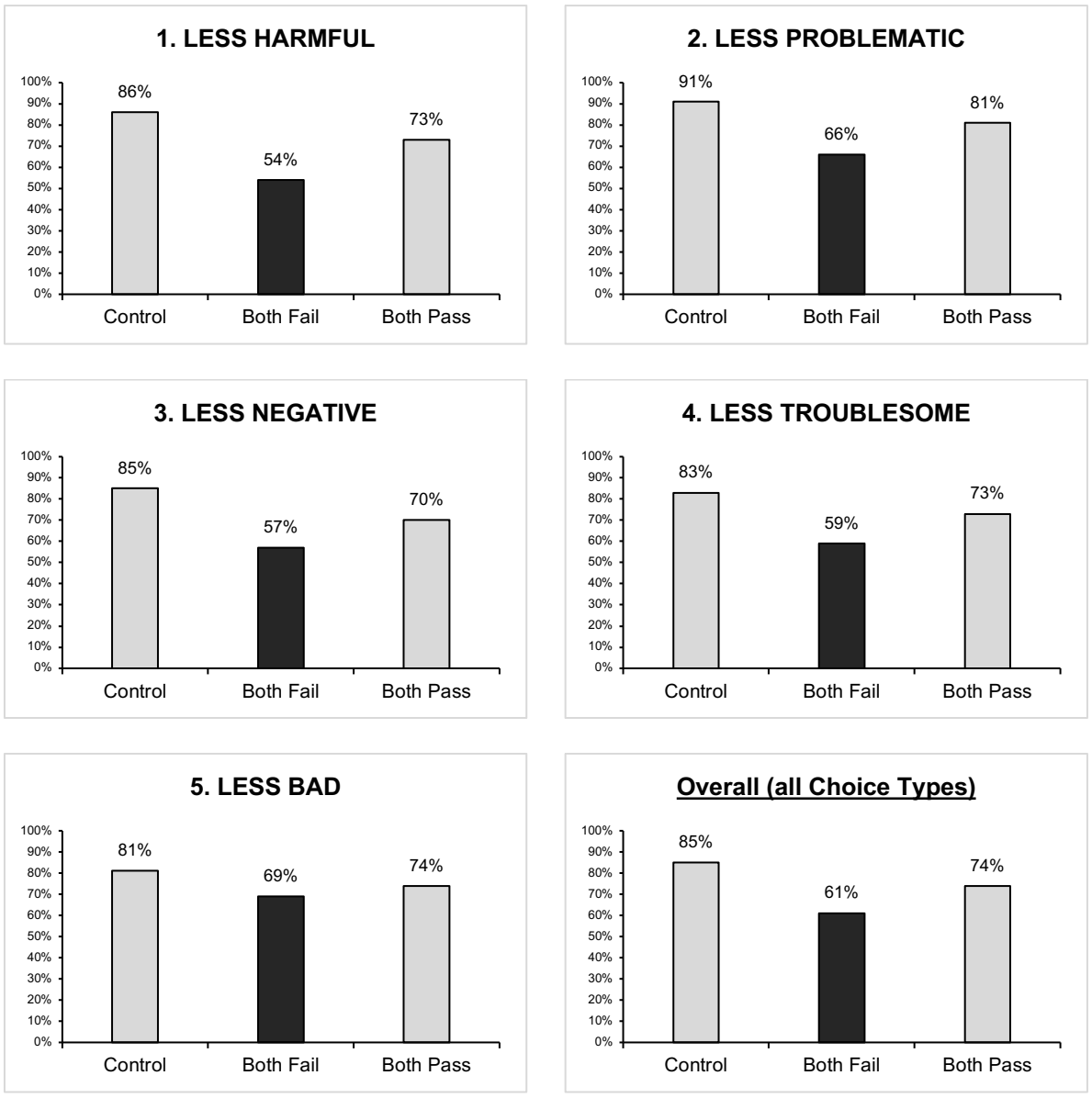
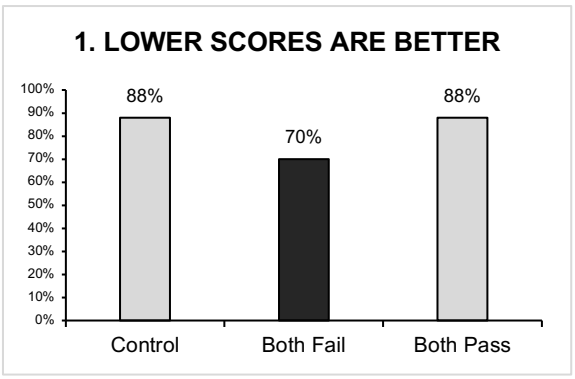
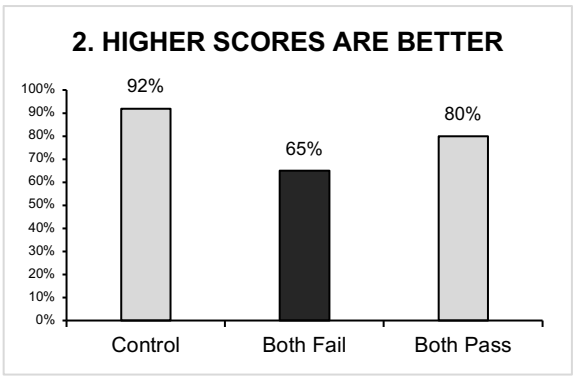


Figure S7. Experiment 7: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 2 = weakest.





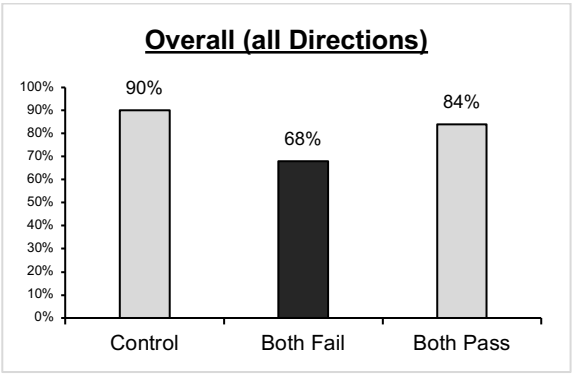


Figure S8. Experiment 9: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 3 = weakest.

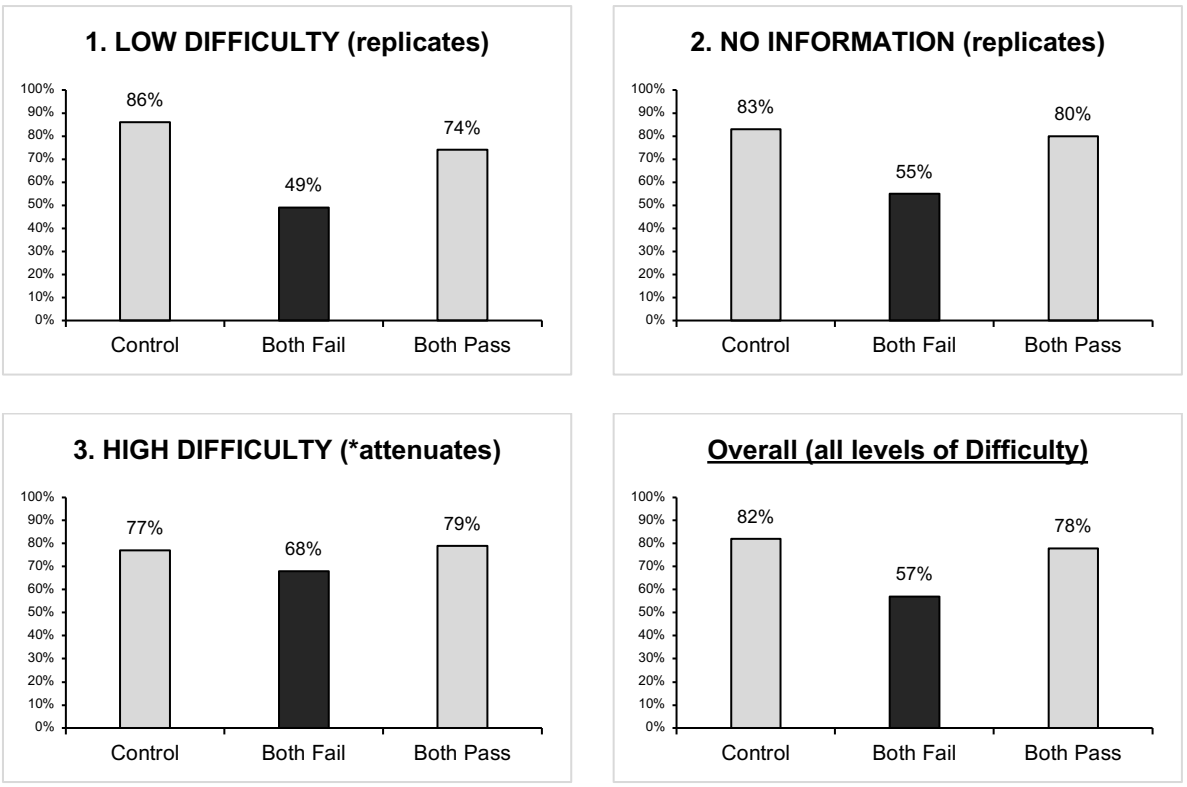


Figure S9. Experiment 10: Percentage of participants choosing the superior outcome in a pair as the superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 3 = weakest.

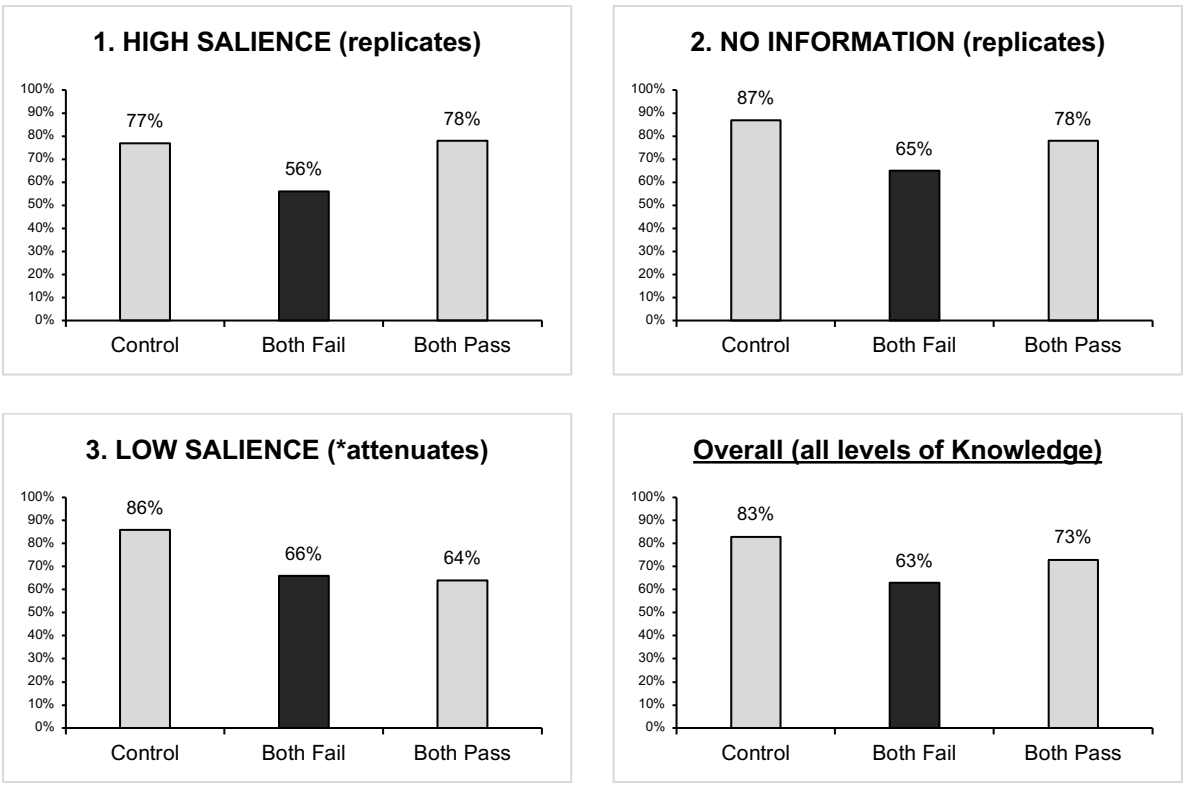


Figure S10. Experiment 11: Percentage of participants choosing the superior outcome in a pair as superior one (vs. choosing "all the same"). Ranked by effect size: 1 = strongest, 2 = weakest.