AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Personality and Social Psychology:
Attitudes and Social Cognition

© 2023 American Psychological Association
ISSN: 0022-3514

https://doi.org/10.1037/pspa0000339

Threshold Violations in Social Judgment

Nadav Klein[1] and Ed O'Brien[2]
[1] INSEAD, Organizational Behavior, Fontainebleau, France
[2] University of Chicago Booth School of Business

People commonly establish in advance the thresholds they use to pass social judgment (e.g., promising reward/punishment after a fixed number of good/bad behaviors). Ten preregistered experiments ($N = 5{,}542$) reveal when, why, and how people *violate* their social judgment thresholds, even after formally establishing them based on having full information about what might unfold. People can be swayed to be both "quicker to judge" (e.g., promising reward/punishment after 3 good/bad behaviors, yet then acting after 2 such behaviors) and "slower to judge" (e.g., promising reward/punishment after 3 good/bad behaviors, yet then withholding until 4 such behaviors)—despite all behaviors obeying their threshold. We document these discrepancies across many parameters. We also propose and test an integrative theoretical framework to explain them, rooted in *psychological support*: Being both "quicker" and "slower" to judge reflect a shared function of the distinct modes of evaluation involved in the act of setting social judgment thresholds (involving a packed summary judgment extending across myriad possible realities) versus following them in real time (involving an unpacked focus on whatever specific reality unfolds, which could provide higher or lower support than threshold setters had accounted for). Manipulating the degree of psychological support thus determines the direction of threshold violations: Higher support produces "quicker to judge" effects while lower support produces "slower to judge" effects. Finally, although violating one's preset threshold may sometimes be to one's benefit, we document initial evidence that it also risks damaging people's reputations and relationships. When it comes to treating others, making exceptions to the rule may often be the rule—for better or worse.

Keywords: change perception, thresholds, social judgment, reputation, reward/punishment

Supplemental materials: https://doi.org/10.1037/pspa0000339.supp

Being deemed a sinner or a saint requires some history of bad or good deeds. To take a rather literal example, the Catholic Church demands evidence of two miracles after a person's death to be canonized for sainthood. When Mother Teresa was posthumously considered for sainthood, however, her supporters grew impatient with the rigidity of this threshold (Perry & Hume, 2016). In prospect, accomplishing two miracles seemed uncontroversial; in practice, it seemed increasingly draconian as Mother Teresa's legacy grew.

This example highlights the fact that—far beyond matters spiritual—people often predetermine thresholds for when to pass social judgment. Indeed, people constantly hold others to certain expectations for good and bad behavior (Kahneman et al., 1986), and such expectations often are quantified in advance. Teachers might set a fixed number of reprimands that may go on a student's record before expelling them; managers might set a sales target that an employee must hit before bonusing them; parents might set ground rules that a child must abide by before rewarding or punishing them; and so on.

How do people's preset thresholds for passing social judgment compare with those they adhere to in practice? Rules such as thresholds necessitate that they apply. Our opening example, however, hints at a discrepancy: People might act more patiently in prospect (e.g., setting "2 miracles for sainthood") than in practice (e.g., calling for sainthood after "just 1 miracle"). Initial findings from Klein and O'Brien (2018) support this possibility. In these studies, participants overestimated their own sampling behavior in preference-formation contexts (e.g., participants overestimated how many sample artworks they would view before deciding whether they liked or disliked the style: Study 2)—leading Klein and O'Brien (2018) to conclude "people use less information than they think to make up their minds" (p. 13222). These findings suggest people may set higher social judgment thresholds in prospect than in practice; people may be quicker to commend, and quicker to condemn, than they think.

Critically, however, these findings remain unexplained—and, we suspect, may be overstated. In the current article, we propose a broader framework to explain them, and to understand the

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

This research was supported by the Willard Graham Faculty Research Fund and the Charles E. Merrill Faculty Research Fund, both from the University of Chicago Booth School of Business, and by the Hoffmann Global Insitute for Business and Society at INSEAD. The authors thank Bryan Baird, Morgan Britt, Jasmine Kwong, Erin Shirtz, and Becky White for collecting data, and Linda Hagen, Nick Epley, George Wu, Alex Imas, Amit Kumar, Anuj Shah, Emma Levine, Josh Klayman, Oleg Urminsky, and Jane Risen for helpful comments on presentations of this research.

Both Nadav Klein and Ed O'Brien contributed to conceptualization, formal analysis, funding acquisition, methodology, resources, writing of original draft, and writing of review and editing.

Correspondence concerning this article should be addressed to Ed O'Brien, University of Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL, 60637, United States. Email: eob@chicagboooth.edu

psychology of social judgment thresholds more generally. We also identify circumstances that reverse the direction of the discrepancy between thresholds set in prospect and those followed in practice. Being "quicker" to judge may represent just one side of the equation; indeed, note that for every proponent clamoring to lower Mother Teresa's threshold for canonization, there seemed to be an opponent demanding to raise it (Taylor, 2016). As we will put forth, there is a fundamental psychological difference between how people set thresholds in prospect versus how people follow thresholds in practice, reflecting a distinction between "packed" judgment (i.e., having to consider myriad possible outcomes in the aggregate) versus "unpacked" judgment (i.e., reacting to the specific outcome that ends up unfolding). Preset thresholds may therefore be rendered too high ("quicker" to judge, replicating Klein & O'Brien's, 2018 effect) *or* too *low* ("slower" to judge), both as a function of how people come to view the particular reality that subsequently manifests.

Setting Versus Following Social Judgment Thresholds

People pass judgment when they believe sufficient evidence is available (Klein & O'Brien, 2016; O'Brien, 2020, 2022; O'Brien & Klein, 2017). We propose, however, that these two modes of judgment—presetting thresholds versus following them in real time—differentially shift when people hit this point. We draw on support theory for inspiration.

Support theory (Tversky & Koehler, 1994) explains judgments under uncertainty. It applies mainly to probability judgments. However, its main propositions also apply to setting versus following social judgment thresholds.

First, support theory posits that people judge the probability of events based on how they are described. In turn, event descriptions vary in the degree to which they make it easy or difficult to imagine the events happening, which support theory defines as the psychological "support" people can summon for those events. This can be represented as:

$$P(A, B) = \frac{s(A)}{s(A) + s(B)}. \tag{1}$$

Here in Equation 1, (A) and (B) are mutually exclusive events, $P(A, B)$ is the judged probability of those events, and $s(A)$ and $s(B)$ are how the events are described. For example, suppose people are asked to estimate the chance that they will "die from any natural cause" (A) rather than from something else (B); according to Equation 1, these estimates are a function of the relative strength of support (s) people can summon for (A) over (B).

Second, support theory posits that psychological support depends on whether information is presented as *packed* or *unpacked*. This can be represented as:

$$s(A) \leq s(A_1 \vee \ldots \vee A_n). \tag{2}$$

Here in Equation 2, "packed" categories elicit less (or no more) support than their "unpacked" components when disjunctively presented. For example, consider the comparison between "dying from any natural cause" versus "dying from heart disease, cancer, or any other natural cause." This information is statistically equivalent—yet numerous tests of support theory find that people judge its latter "unpacked" description as more probable (e.g., Fischhoff et al., 1978; Johnson et al., 1993; Mulford & Dawes, 1999; Redelmeier et al., 1995; Rottenstreich & Tversky, 1997; Russo & Kolzow, 1994; Van Boven & Epley, 2003).

Support theory suggests this effect of piecemeal unpacking "represents a basic principle of human judgment," as it "enhances [the component's] salience and hence its support" (Tversky & Koehler, 1994, p. 549). Many categories contain too many layers for people to consider (e.g., it is impossible to bring to mind every cause of death), leading people to change their estimates based on what is unpacked ("Oh, right; heart disease and cancer are big ones. Natural causes are a big problem"). By this same rationale, however, the effect is *reversed* when unpacking draws attention to components with *weak* support (Macchi et al., 1999; Sloman et al., 2004). For example, people may believe they are more likely to die from "heart disease, cancer, or any other natural cause" than from "any natural cause"—but they may *also* believe they are *less* likely to die from "septicemia, malaria, or any other natural cause" than from "any natural cause" ("Oh, right; there are lots of things that won't affect me. Natural causes aren't always a big problem").

Put in terms of our research, the act of setting thresholds in advance resembles support theory's "packed" judgments. For example, a manager might set a reward threshold of "5 stars get a bonus" or a reprimand threshold of "3 strikes get a penalty." Yet in presetting such thresholds, note how difficult it is for managers to have summoned each particular version or combination of how "5 stars" or "3 strikes" can play out (across employees, contexts, time, and so on); such a task is effectively impossible, akin to having to bring to mind every possible cause of death to inform one's global estimate. People are found to make prediction errors even when tasked with imagining a single reality, faithfully described (Wilson & Gilbert, 2005)—let alone when tasked with imagining the myriad of realities that can unfold to begin with. All else equal, people's preset thresholds are therefore likely summoned based on some salient representative average or summary ("What is a star or strike here typically like? … How many of those may make a fair threshold for all?"), just as how support theory assumes people generate "packed" probabilities (e.g., "What are some typical causes of death? … How likely are they?")—which presetters may take as the next best proxy for determining a fair threshold (Anderson, 1965; Asch, 1946; Birnbaum, 1972).

Conversely, the act of following thresholds as they actually unfold resembles support theory's "unpacked" judgments; indeed, reality *always* plays out as an unpacked state. Yet note that the unpacked pieces that reality unveils need not correspond to some representative summary experience; *myriad* specific versions or combinations of events could end up unfolding, ranging from completed tasks that wield higher support (akin to "heart disease and cancer") to those that wield lower support (akin to "septicemia and malaria"). Consistent with the logic of support theory, these differences in support should change how people respond in real time, in ways presetters might not have accounted for.

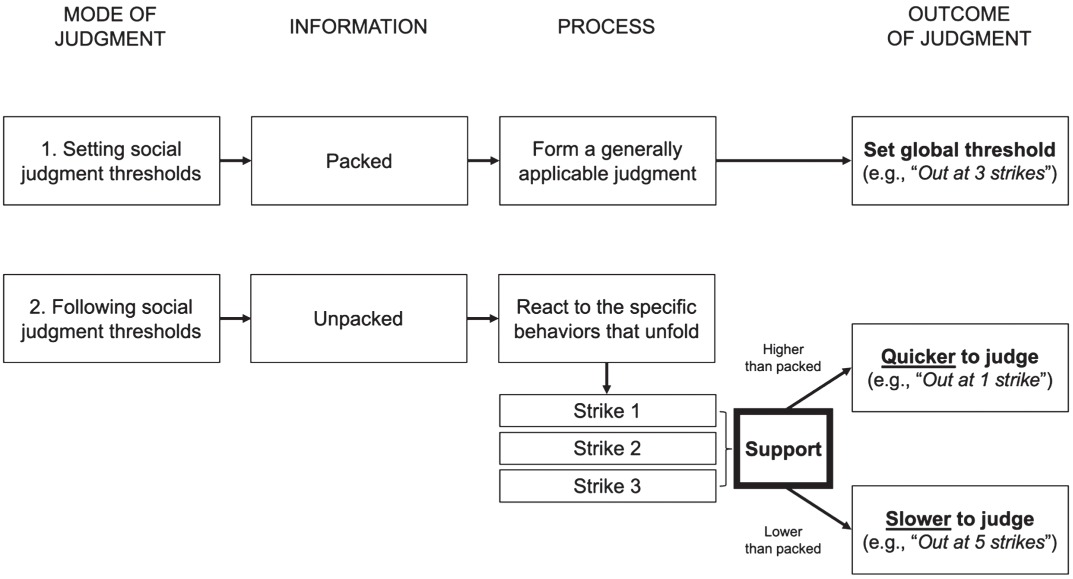
Our Proposed Model

Putting these ideas together, we propose that the points at which people pass social judgment will depend on the support they can summon for what unfolds—which will vary by their mode of judgment (see Figure 1).

Setting thresholds in advance is a form of prediction whereby judges must make a decision about multiple individual behaviors

Figure 1

Framework for Understanding the Psychology of Social Judgment Thresholds


MODE OF JUDGMENT
INFORMATION
PROCESS
OUTCOME OF JUDGMENT
1. Setting social judgment thresholds
Packed
Form a generally applicable judgment
Set global threshold (e.g., "Out at 3 strikes")
2. Following social judgment thresholds
Unpacked
React to the specific behaviors that unfold
Strike 1
Strike 2
Strike 3
Support
Higher than packed
Lower than packed
Quicker to judge (e.g., "Out at 1 strike")
Slower to judge (e.g., "Out at 5 strikes")

packed together, leading them to generate some global estimate that may fairly apply as a threshold. In contrast, following thresholds is a form of experience whereby judges encounter each piecemeal behavior as it unfolds, with each piece wielding its own support—in specific individual ways that presetters might not have fully accounted for (even when knowing that all such pieces were possible). Higher support should therefore lower thresholds and hasten judgment (vs. what was envisioned); lower support should raise thresholds and slow judgment. This model allows one to predict not only when people are "quicker" to judge relative to preset thresholds (as put forth by Klein & O'Brien, 2018) but also when people are "slower" to judge—both depending on the support provided by what unfolds.

In addition, given this model, one can ask: What are the psychological inputs into support (bold box in Figure 1)? The current research will test some (nonexhaustive) possibilities (e.g., as we will discuss, unpacked behaviors that elicit higher vs. lower emotions tend to elicit higher vs. lower support); we will expand on each as they arise in our experiments. Our broader point here is not about any one such factor per se; rather, we propose that they are examples of the same broader construct—*support*—that is the key driver explaining when and why people may violate social judgment thresholds, in *either* direction.

Overview of Experiments

Ten preregistered experiments (total $N = 5{,}542$) tested this hypothesized framework. Table 1 provides an overview.

First, Experiment 1 sought to confirm Klein and O'Brien's (2018) basic "quicker to judge" effect in social judgment contexts. We held all information exactly constant but manipulated whether it was presented in a packed versus unpacked format. Thus, because all information is explicitly identical, with no extra layers to unpack, the unpacking manipulation here simply serves to increase its salience and thus its support—and so we hypothesized that unpacked participants will be "quicker to judge" relative to packed participants.

Next, Experiments 2a–5b sought to test our key research question: Does this discrepancy reflect a universal effect of being "quicker" to judge? Or, as we hypothesize, does it depend on the support provided by what exactly is unpacked? According to our theorizing, unpacked experiences that elicit higher support should indeed produce a "quicker to judge" effect while those that elicit lower support should instead produce a "slower to judge" effect. We tested this hypothesis across various reward and punishment contexts, spanning three sources of support—the emotionality of the unpacked behaviors (i.e., how much positive or negative emotion they elicit), the clearness of the unpacked behaviors (i.e., how clearly good or bad they seem), and the consistency of the unpacked behaviors (i.e., how consistently they play out)—each of which should feed into support upon being unpacked and thus promote *either* effect, with higher levels hastening judgment and lower levels slowing judgment.

Finally, Experiments 6a–6b shifted to a different question: While our prior set of experiments tested conditions for threshold violations, this set tested downstream consequences for specific examples from the real world. We tested, for example, whether violating one's social judgment threshold can risk damaging people's reputations and relationships (e.g., by disrupting learning, making one seem hypocritical, and inviting retaliation).

For all experiments, we report all measures, manipulations, and exclusions (if any). We predetermined sample sizes by conducting a power analysis (G*Power; Faul et al., 2007) of all seven studies reported in Klein and O'Brien (2018); the average size of their key effect was $d = 0.81$, yielding a recommended cell size of 41 participants for comparing two means (two-tailed, $\alpha = 0.05$, 95% power). Using this number as a reference point, we rounded up (to 50)

Table 1
Overview of Experiments

Experiment	Goal of study	Study task	Manipulation of unpacking	Dependent variable	Hypothesis	Why?
Experiment 1	Test basic unpacking effect	P's read about a person's repeated good/bad behaviors	Between subjects: The repeated behaviors are presented globally versus act-by-act	Yes/no: Whether P's draw dispositional attributions about person's character	Unpacking increases rates of dispositional attributions	Unpacking increases piecemeal salience and thus support
Experiment 2a	Test Moderator 1: Manipulate unpacked *emotion* (positive)	P's can choose a person as partner in econonic games involving trust	Between subjects: P's imagine versus experience a trusting interaction with person, before choice	Yes/no: Whether P's choose person as partner	Unpacking increases rates of choosing person as partner	Higher emotion = higher support
Experiment 2b	Test Moderator 1: Manipulate unpacked *emotion* (negative)	P's can accuse a partner of cheating in economic games involving die rolls	Between subjects: The partner rolls repeated wins, which P's imagine globally versus experience roll-by-roll	Yes/no: Whether P's accuse person of cheating	Unpacking increases rates of accusing person of cheating	Higher emotion = higher support
Experiment 3a	Test Moderator 1: Manipulate unpacked *emotion* (positive)	P's set threshold for promoting a helpful user of forum	Between subjects: P's imagine versus experience a helpful post of user, before setting threshold	1–10 ratings: P's self-set threshold; user should be promoted after 1–10 kind posts	(i) Unpacking leads P's to soften (i.e., lower) promotion threshold; (ii) This effect is attenuated by weakening P's emotional reactions	Higher emotion = higher support; lower emotion = lower support
Experiment 3b	Test Moderator 1: Manipulate unpacked *emotion* (negative)	P's set threshold for banning a "troll" user of forum	Between subjects: P's imagine versus experience a troll post of user, before setting threshold	1–10 ratings: P's self-set threshold; user should be banned after 1–10 troll posts	(i) Unpacking leads P's to harshen (i.e., lower) banning threshold; (ii) This effect is attenuated by weakening P's emotional reactions	Higher emotion = higher support; Lower emotion = lower support
Experiment 4	Test Moderator 2: Manipulate unpacked *clearness of goodness/badness*	P's can agree to bonus/penalize a worker after fixed number of early/late arrivals	Within subjects: P's indicate their agreement before arrivals occur (packed), then again after arrivals occur (unpacked)	Yes/no: Whether P's agree to issue bonus/penalty	(i) Unpacking increases action when arrivals are clearly good/bad; (ii) Unpacking decreases action when arrival are ambiguously good/bad	Higher clearness = higher support; Lower clearness = lower support
Experiment 5a	Test Moderator 3: Manipulate unpacked *consistency*	P's set threshold for bonusing/penalizing a player for playing helpfully/harmfully across series of economic games	Within subjects: P's set their threshold before games occur (packed), then judge a player who falls short of it (unpacked)	Yes/no: Whether P's act anyway, despite the worker falling short of threshold	Unpacking increases rates of acting-anyway when worker is consistently helpful/harmful otherwise	Higher consistency = higher support
Experiment 5b	Test Moderator 3: Manipulate unpacked *consistency*	P's set threshold for bonusing/penalizing a player for playing helpfully/harmfully across series of economic games	Within subjects: P's set their threshold before games occur (packed), then judge a player who meets it (unpacked)	Yes/no: Whether P's withhold action, despite the worker meeting threshold	Unpacking increases rates of witholding when worker is inconsistently helpful/harmful otherwise	Lower consistency = Lower support

(*table continues*)

Table 1 *(continued)*

Experiment	Goal of study	Study task	Manipulation of unpacking	Dependent variable	Hypothesis	Why?
Experiment 6a	Test downstream consequences (real-world stimuli)	P's recall others' threshold violations from their own lives	N/A	P's self-reported judgments of themselves and others	P's judge others negatively when others violate their preset thresholds	Threshold violations may seem unfair/ hypocritical if P's cannot contextualize them
Experiment 6b	Test downstream consequences (controlled stimuli)	P's read about various kinds of threshold violators	N/A	P's self-reported judgments of the violators	P's judge others negatively when others violate their preset thresholds	Threshold violations may seem unfair/ hypocritical if P's cannot contextualize them

and doubled it (to 100) for good measure to account for our varied designs; that is, for all experiments, we predetermined sample sizes of at least 100 participants per cell (or more, as resources allowed). All data, materials, and preregistrations, as well as a copy of our Supplemental Materials document, can be found on the Open Science Framework (OSF) at https://osf.io/b4h8d/.

We refined our thinking and terminology as this research developed. In Experiment 1, for example, we originally conceptualized participants as being "predictors versus experiencers" who make "tipping point" judgments (as seen in the preregistration and in the study measures)—but we now conceptualize these features as "packed versus unpacked" and "threshold" judgments. Also, our preregistrations for Experiments 2a–2b–3a–3b included supplementary mediation analyses, which worked as predicted (see OSF for these results).

Experiment 1: Unpacking Social Judgment

In Experiment 1, we assessed thresholds for drawing dispositional attributions. Participants evaluated a target who repeatedly committed the same good or bad act, and indicated their threshold for judging the target as a good or bad actor—as a function of evaluating those behaviors all together up front (i.e., packed) or as they unfolded one by one (i.e., unpacked). We hypothesized that predicted thresholds would be higher than actual thresholds—because in this basic unpacking task, all information is explicitly identical and thus the unpacked (vs. packed) condition simply increases the salience of each relevant piece, providing more support. Our goal was to establish a foundation for our main experiments by first confirming Klein and O'Brien's (2018) "quicker to judge" effect.

We also further advanced Klein and O'Brien (2018) by introducing a unique design feature in this experiment (and in all our experiments)—one that rules out random error in inadvertently manufacturing their effect. In Klein and O'Brien (2018), the typical study asked predictor participants to report predictions on a larger scale than experiencer participants could report on (e.g., asking predictors, "How many consecutive behaviors, of 10, would lead you to tip?" vs. asking experiencers, after each behavior, "Have you tipped yet?" [and counting until they clicked "yes" over "no"])—meaning that random error had worked in favor of their "quicker to judge" hypothesis. Experiment 1 (and all our experiments) will avoid this concern by matching the scales of measurement across conditions.

Method

Participants

We requested 800 "Cloud Approved" participants from Cloud Research, yielding 804 participants (45.15% women; 27.49% non-White; $M_{\text{age}} = 39.73$, $SD_{\text{age}} = 12.40$) who completed the experiment for $0.30.

Procedure

The experiment followed a 2 (judgment type, between subjects: packed vs. unpacked) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior) design.

First, all participants learned they would evaluate "Person E." Their task was to figure out whether they view Person E as a "good person (officially has positive character; their good behaviors aren't just a fluke)" or as a "bad person (officially has negative character; their bad behaviors aren't just a fluke)," and that to do so, they would have the opportunity to see how Person E behaves over the next six consecutive observations of them encountering someone in need. We operationalized good behavior as "Person E tries to help the person in need" and bad behavior as "Person E tries to avoid the person in need." We assessed simple descriptions of identically repeated behaviors—equally for all participants—for a pure test of basic unpacking (such that each individual event is made more salient by separating them piece by piece). All participants observed Person E's behavior three times (out of the 6 purported available observations), but we conveyed these three observations differently across conditions. We informed all participants that they could not end participation early, regardless of their responses (thus, they could not strategically end early by indicating that they had already "seen enough" to judge).

We then randomly assigned participants to 1 of 4 conditions. Packed participants read (nonbrackets show Good-Packed condition; brackets show Bad-Packed condition):

> When do you think you'd hit your tipping point—the very first point at which you'd feel like Person E has exhibited good behavior [bad behavior] enough times such that you'd officially view them as a good person (i.e., that their good behavior isn't just a fluke and they must indeed be a good person) [bad person (i.e., that their bad behavior isn't just a fluke, and they must indeed be a bad person)] …?

Serving as our key dependent variable, these Packed participants then chose from 1 of 2 forced-choice options, presented in randomized order, each prefaced with "If I learned that Person E exhibits good behavior [bad behavior] at Observations 1–3 …" One option corresponded to hitting their judgment threshold at this point ("I would tip in judging them at this point; that's enough for now; no need to see what they do for Observations 4–6"), and the other corresponded to not yet hitting this threshold ("I wouldn't tip in judging them at this point; that's not enough for now; still need to see what they do for Observations 4–6").

We compared these responses to those of Unpacked participants, who followed identical prompts up to the point that we operationally defined good and bad behavior, then read: "Click >> for Observation 1." These participants then proceeded to observe, screen by screen, Person E's behavior—for each of Observations 1–3, separately—learning Person E exhibited "good behavior [bad behavior]" each time. After clicking through each screen, Unpacked participants completed the same dependent variable as Packed participants, phrased in the present tense ("I do tip …"; "I don't tip …").

Thus, if we find that more Unpacked versus Packed participants pass judgment at this point as hypothesized, note that this difference cannot be explained by random error—because participants in both conditions report their responses on the same binary scale (unlike the typical study in Klein & O'Brien, 2018, which compared participants who made binary ratings to participants who made ratings on a larger continuous scale).

Finally, all participants reported demographic information and completed two forced-choice attention checks: whether they evaluated a person who committed "good behavior" versus "bad behavior," and whether they had been instructed to indicate their threshold at the "first point" versus "last point" they had formed judgment. They also completed a no-penalty check about whether their judgment truly reflected having hit (or not yet hit) their threshold, or whether it reflected something else (forced choice: "reported my genuine tipping-point behavior at this point"; "wanted to end the study ASAP; "wasn't paying attention and clicked at random"; "other [please describe]"), plus a no-penalty honesty check regarding whether we should trust their data as genuine (forced choice: "yes" or "no").

Results and Discussion

Main Results: Threshold Behavior

We conducted a Binary Logistic Regression with Judgment Type (Packed vs. Unpacked) and Valence of Judged Behavior (Good Behavior vs. Bad Behavior) as between-subjects factors, and threshold behavior (yes-judge or no-judge) as the dependent variable.

First, there was a main effect of Valence, such that participants were quicker to judge bad actors versus good actors, $\beta = 0.16$, $SE = .07$, $p = .025$. This result is incidental to the current research, but it is consistent with the valence asymmetry we have found in our previous research on threshold judgments (Klein & O'Brien, 2016; O'Brien, 2020, 2022; O'Brien & Klein, 2017)—a point to which we will return in the General Discussion.

Second, and more critical for the current research, there was also the hypothesized main effect Judgment Type, $\beta = -0.31$, $SE = .07$, $p < .001$—which emerged across Valence (null interaction: $\beta = 0.02$, $SE = .07$, $p = .830$). Pairwise comparisons reveal that Packed participants were more likely to predict they would withhold judgment after the target exhibited the same behavior for Observations 1–3, as compared to what Unpacked participants reported after experiencing Observations 1–3 one by one. When observing three consecutive good acts, more Unpacked participants (48.08%, 100 of 208) judged the target as a good person, as compared to the number of Packed participants who thought they would pass judgment at this point (33.84%, 67 of 198), $\beta = -0.30$, $SE = .10$, $p = .004$; when observing three consecutive bad acts, more Unpacked participants (56.87%, 120 of 211) judged the target as a bad person, as compared to the number of Packed participants who thought they would pass judgment at this point (40.64%, 76 of 187), $\beta = -0.33$, $SE = .10$, $p = .001$.

Other Variables

Most participants passed the attention checks (valence: 91.92%, 739 of 804; "first/last": 70.90%, 570 of 804); reported their genuine threshold (92.91%, 747 of 804); and passed the honesty check (98.51%, 792 of 804). When rerunning our analyses while excluding participants who failed any of these checks (leaving $N = 489$), results are unchanged (main effect of Judgment Type: $\beta = -0.28$, $SE = .09$, $p = .002$).

Experiment 1 confirms Klein and O'Brien's (2018) "quicker to judge" effect. When unpacked, others' behaviors provided more support for passing social judgment.

Next, we turn to the main goal of the current research. We test whether this effect varies with the support provided by what is unpacked (as opposed to reflecting a universal effect of being "quicker" per se). We assess three sources of support: Unpacked behaviors can vary in how much emotion they elicit (Experiments

2a−2b−3a−3b), how clearly good or bad they seem (Experiment 4), and how consistently they play out (Experiments 5a−5b), each of which should produce "quicker" *or* "slower" judgment upon being unpacked.

Experiments 2a–2b–3a–3b: Emotionality As an Input Into Support

Unpacked events can vary in the extent to which they stir strong versus weak emotions. In turn, higher (vs. lower) emotions should provide higher (vs. lower) support, because people tend to draw on perceived emotionality as information for forming judgment (e.g., Hutcherson & Gross, 2011; Jenni & Loewenstein, 1997; Keltner et al., 1993; Lerner & Keltner, 2000; Schwarz & Clore, 2007; Van Boven et al., 2013).

Experiments 2a–2b–3a–3b tested whether the higher (vs. lower) support provided by higher (vs. lower) unpacked emotionality affects judgment thresholds—with higher support producing "quicker" judgment but lower support producing "slower" judgment.

Experiment 2a: Choosing Partners in Economic Games

In Experiment 2a, participants chose a partner for "The Trust Game." To aid their choice, we allowed them to chat with another alleged person to assess their trustworthiness—and we rigged the procedure such that this person was very kind to the participant. We randomly assigned participants to either experience this kindness play out or to merely imagine it all up front. We hypothesized that participants would be more likely to choose the person as their partner after experiencing this kind interaction unfold versus after simply imagining it all up front, with the former providing more support (via positive emotion).

Method

Participants. We requested 500 "Cloud Approved" participants from Cloud Research, yielding 499 participants (41.08% women; 27.66% non-White; $M_{\text{age}} = 39.93$, $SD_{\text{age}} = 11.45$) who completed the experiment for $3.00.

Procedure. The experiment followed a 2-cell (judgment type, between subjects: packed vs. unpacked) design.

All participants learned they would be playing "The Trust Game" with another worker—specifically, based on random assignment, that they would either simply imagine playing it (Packed condition) or that they would actually play it (Unpacked condition).

All participants read the same game instructions (see OSF for all materials), which stated that there are two players (two current workers, i.e., participants now taking the study), playing at the same time, on the same team; that the game "involves quick tasks that require trust between you and your partner; for example, one such task will give your partner a chance to cheat you out of money and keep it all for themselves"; and that each person can "win up to $5.00 in bonuses, depending on how your team plays." All instructions were identical; for Packed participants, each statement began with "Imagine …" (e.g., "Imagine the game involves quick tasks" vs. "The game involves quick tasks").

Furthermore, and maintaining this "Imagine" language throughout for Packed participants, all participants learned they would choose their own partner—and, to help them do so, we would pair them for a quick conversation with another current worker, after which they would be able to "freely choose whether you want to play with this person as your teammate (and thus start the game), or if you want to chat more with them, or move to someone new (and thus continue your teammate search, before starting the game)." We emphasized to all participants, identically, to "choose your partner wisely," meaning it was in their "best interest to choose a partner who you find trustworthy, before playing The Trust Game"; and that, "if they happen to be kind to you during your interaction," participants should choose them only if they "feel sure their kindness is genuine."

This process of choosing a partner for The Trust Game to come later was, unbeknownst to participants, our *real* measure of trust. At this point, we led Unpacked participants to believe that they were actually being paired with such a worker (e.g., via timed loading screens, programed by us). They then learned they were successfully paired, and that we were randomly assigning them to a particular conversation context whereby they would first respond to "3 personal questions"; we would then send their responses to the other (alleged) worker; the other worker would then write a reply; at which point they (the real participant) would control what to do next. This procedure then unfolded as described, whereby these Unpacked participants typed their "Top 3 favorite TV shows at the moment" (Question 1); they selected which "guilty pleasure" they "admit to enjoying the most from time to time" (forced choice, 1 of 10 options, presented in randomized order: Sleeping in; Enjoying a drink; Zoning out; Treating myself; Gaming; Late-night snacking; Wasting time online; Indulging a sweet tooth; Feigning busyness for alone time; Blowing off exercise"; Question 2); and, as they typed via an open-ended text box (500-character minimum), "What is the one, single lasting goal that you most hope to achieve with your life?" (Question 3). Afterward, we led them to believe that we had sent these responses to the other worker, and that the other worker then sent a reply. In the reply (always the same written paragraph; see OSF), the other worker thanked the participant, complimented them, and stated that they also happened to love one of the same shows (always piped in as whatever participants had typed in as their second top show); that they also happened to share the same guilty pleasure (using customized text to match whatever participants chose); and that the participant can trust them to be their partner.

These Unpacked participants then completed the key dependent variable: "So: What do you want to do? (Entirely your choice! They are currently awaiting your response, after which we'll assign things based on whatever you choose below [they won't know that you chose anything!])." They chose from 1 of 3 options, presented in random order: "I'm ready: I choose this partner to play The Trust Game with (let's start The Trust Game)"; "I'm not yet ready: I choose to keep chatting with this partner before knowing for sure if I want to play The Trust Game with them (don't yet start The Trust Game)"; or "I'm not yet ready: I choose to drop this partner and chat with a new randomly drawn person to figure them out as my possible partner to play The Trust Game with (don't yet start The Trust Game)." After making their choice, they rated a manipulation check for support, here in terms of emotion: "How'd you feel about how your partner responded?," rated from 1 (*I felt as "warmed" [moved/taken/etc.] as I assumed they'd make me feel]*) to 10 (*I felt more "warmed" [more moved/taken/etc.] than I assumed they'd make me feel]*).

For comparison, Packed participants read these same instructions for The Trust Game, and imagined all of the procedures Unpacked participants completed. Then, they did not see an actual reply, but instead read a summary of it:

> Imagine they then thank you, compliment you, and write that they also happen to love one of the same shows you listed; that they also happen to enjoy your same guilty pleasure; and they say you can trust them to be your partner.

They then completed the same key dependent variable (choice of partner at that point), followed by the same manipulation check, with phrasings adapted for "imagining" how they "thought" they would respond.

Finally, all participants reported demographic information and completed an attention check regarding the name of the game in the study (forced choice from 1 of 3 options: Trust Game, California Game, Halloween Game), plus the same honesty check from Experiment 1. Also, we showed Packed participants the text of the alleged partner's reply (pasted from the Unpacked condition), and asked whether it fairly matched our opening summary description (forced choice: "yes" or "no")[1]; likewise, we debriefed Unpacked participants that the other worker was not real, and asked them whether they believed the study and worker were real before this end-of-study reveal (forced choice: "yes" or "no").

Results and Discussion

Manipulation Check: Support (Via Emotion). First, there was a significant effect of our manipulation on support: An independent-samples t test confirmed that Unpacked participants reported higher support (i.e., emotion: $M = 6.85$, $SD = 2.81$) than Packed participants ($M = 5.09$, $SD = 2.35$), $t(497) = 7.60$, $p < .001$, $d = 0.68$.

Main Results: Threshold Behavior. Next, we conducted a Binary Logistic Regression with Judgment Type (Packed vs. Unpacked) as a between-subjects factor and choice of partner (recoded as binary, as preregistered: yes-choose or no-choose) as the dependent variable.

There was a significant effect of Judgment Type: As hypothesized, more Unpacked participants chose the other person as their partner for The Trust Game (88.94%, 209 of 235; other responses: 4.68% keep chatting, 11 of 235; 6.38% drop and chat with someone new, 15 of 235) as compared with Packed participants who imagined being at this same point (40.91%, 108 of 264; other responses: 50.76% keep chatting, 134 of 264; 8.33% drop and chat with someone new, 22 of 264), $\beta = 2.45$, $SE = .24$, $p < .001$.

Other Variables. Most participants passed the attention check (99.80%, 498 of 499) and the honesty check (99.60%, 497 of 499). At the end of the study, recall that we showed Packed participants the actual reply of the other worker; most (90.53%, 239 of 264) reported this reply fairly matched our opening summary description. Likewise, most Unpacked participants (78.30%, 184 of 235) believed everything was real (before being debriefed). When rerunning our analyses while excluding participants who failed the attention or honesty check, as well as any Packed participant who reported unfair expectations and any Unpacked participant who did not believe the study (leaving $N = 422$), results are unchanged: Effect of Judgment Type on support, $t(420) = 8.37$, $p < .001$, $d = 0.82$; on partner choice, $\beta = 2.60$, $SE = .28$, $p < .001$.

Posttest. To further ensure this discrepancy was driven by differential support—as opposed Packed versus Unpacked participants judging categorically different targets (see Footnote 1 from earlier)—we recruited a separate sample of participants from the same population ($N = 250$, for \$3.00; 43.20% women; 23.20% non-White; $M_{\text{age}} = 39.65$, $SD_{\text{age}} = 11.56$). They completed the same procedures as Unpacked participants—but before completing the dependent measures, we alerted them that the person was not real. This manipulation should reduce support (i.e., emotion) while otherwise matching the Unpacked condition. Indeed, these participants reported moderate support on our manipulation check ($M = 5.83$, $SD = 2.74$), with 73.60% (184 of 250) indicating they would choose the partner—both of which fell in between our Packed participants ($M = 5.09$, $SD = 2.35$; 40.91%, 108 of 264) and Unpacked participants ($M = 6.85$, $SD = 2.81$; 88.94%, 209 of 235), all $ps < .002$ (see OSF for details and files). We will be following up on these posttest results by directly testing the notion of differential unpacked support in Experiments 3a–3b–4–5a–5b.

Experiment 2a provides initial evidence for emotion as an input into unpacked support. Participants who actually experienced a kind interaction unfold (vs. merely imagining it all up front) lowered their threshold for choosing the person as a trustworthy partner.

Experiment 2b: Accusing Partners in Economic Games

Experiment 2b resembled Experiment 2a but focused on *dis*trust. Participants completed a task involving economic games where their opponents' behavior could be interpreted as cheating. We hypothesized that participants would be more likely to report them for cheating after actually playing these games one by one versus merely imagining them all up front, with the former providing more support (via negative emotion).

Method

Participants. We requested 500 "Cloud Approved" participants from Cloud Research, yielding 491 participants (43.58% women; 28.11% non-White; $M_{\text{age}} = 39.56$, $SD_{\text{age}} = 11.42$) who completed the experiment for \$0.60.

Procedure. The experiment followed a 2-cell (judgment type, between subjects: packed vs. unpacked) design.

All participants learned they would be playing "The Numbers Game" with another worker—specifically, based on random assignment, that they would either imagine playing it (Packed condition) or that they would actually play it (Unpacked condition).

All participants read the same game instructions (see OSF for all materials), which stated that there are two players (two workers currently taking the study) playing against each other; that the game involves each player rolling "a randomly assigned number from virtual dice ranging from 1–6"; that each player will privately roll and report the number; that whoever reports the higher number "wins that round" and will receive a \$1.00 bonus; and that they would

[1] Such a measure, as assessed in this experiment and in all other relevant experiments, allows us to rule out a rather uninteresting alternative explanation: If we ask Packed participants to imagine Category X, but give Unpacked participants examples that are not part of Category X, then this would mean participants across conditions judge categorically different targets (and this might obviously explain why they respond differently).

play each other for five rounds in total ("thus, one can win up to $5.00 in bonuses"). Critical for our procedure, all participants read that "Based on a random drawing, this survey has determined that if any round ends in a tie, the other worker automatically wins." This feature is critical because it ensures that all participants know that their opponent knows that they (the opponent) are guaranteed to win each round, no matter what, so long as they (the opponent) report rolling six (i.e., Beating 1–5, and breaking a tie with another 6).

We led Unpacked participants to believe they were actually being paired with such a worker (e.g., via timed loading screens, programed by us). They then learned they were successfully paired, and the game began as described. They rolled their number for Round 1 via a click, at which point they learned their outcome, which we programed to be a random number from 2–5. They typed their number into a box and then clicked to proceed, after which they learned their opponent reported a "6 (highest possible number; automatic win for them)." They then repeated this procedure for Rounds 2–5, programed this same way. Thus, all Unpacked participants lost five rounds in a row, with their opponent reporting a six each time.

At this point, these Unpacked participants then completed what was our *real* dependent variable (unbeknownst to participants): "Help us sort through our participant pool! Do you want to report this worker for possible cheating? If you say yes, we'll look into it (you'll remain anonymous)." They made their choice from 1 of 2 options, presented in randomized order: "yes" or "no." After making their choice, they rated a manipulation check for support (i.e., emotion): "What is your emotional state at this point in the game?," rated from 1 (*as emotional [angry/upset/taken aback, etc.] as I imagined I'd be*) to 10 (*surprisingly emotional [angry/upset/taken back, etc.], more than I imagined I'd be*).

For comparison, after Packed participants read these same instructions for The Numbers Game, they further imagined the following: "Imagine that the other worker reports a 6 (highest possible number; automatic win for them) for Rounds 1–5." They saw the prompt that Unpacked participants responded to regarding whether they wanted to report this worker for possible cheating, and read: "How would you respond?" They predicted their response via the same scale—followed by the same manipulation check, with phrasings adapted for "imagining" how they "thought" they would respond.

Finally, all participants reported demographic information and completed an attention check regarding the name of the game in the study (forced choice from 1 of 3 options: Numbers Game, Food Game, Winter Game), an attention check regarding the bonus winnings for each round (forced choice from 1 of 3 options: $0.01, $1.00, $100.00), and the same honesty check from prior experiments. Also, we debriefed Unpacked participants that the other worker was not real, and asked them whether they believed the study and worker were real before this end-of-study reveal (forced choice: "yes" or "no").

Results and Discussion

Manipulation Check: Support (Via Emotion). First, and unexpectedly, we did not find our hypothesized effect on support. Although Unpacked participants indeed reported higher support (i.e., emotion: M = 4.48, SD = 2.66) than Packed participants (M = 4.36, SD = 2.40), this difference was not statistically significant, independent-samples t test: $t(489) = 0.54$, $p = .589$, $d = 0.05$. After we report our preregistered analyses, we will return to these null manipulation-check results with a potential post hoc explanation.

Main Results: Threshold Behavior. Next, we conducted a Binary Logistic Regression with Judgment Type (Packed vs. Unpacked) as a between-subjects factor and reporting of the partner for cheating (yes-report or no-report) as the dependent variable.

There was a significant effect of Judgment Type: As hypothesized, more Unpacked participants reported their partner for cheating (80.67%, 192 of 238) as compared with Packed participants who imagined being at this same point (66.80%, 169 of 253), $\beta = 0.73$, $SE = .21$, $p < .001$.

Other Variables. Most participants passed the attention checks ("name": 99.59%, 489 of 491; "earnings": 99.39%, 488 of 491) and the honesty check (100%, 491 of 491). At the end of the study, recall that we asked Unpacked participants if they believed they were playing with a real partner; most did (59.24%, 141 of 238). When rerunning our analyses while excluding participants who failed any of these checks, as well as any Unpacked participant who did not believe they were playing with a real partner (leaving N = 390), our main result is unchanged (Effect of Judgment Type on reported cheating, $\beta = 0.66$, $SE = .25$, $p = .008$), but the manipulation check results *are* changed—such that they now indeed become *significant* (as we initially expected) as opposed to remaining unexpectedly null, Effect of Judgment Type on support, $t(388) = 2.40$, $p = .017$, $d = 0.25$.

Possible Explanation for Null Effect on Support: An Exploratory Reanalysis

Because we had preregistered to find a significant effect on the manipulation check, we did not preregister plans for what we would do if this effect was nonsignificant. In any case, we conducted exploratory analyses to better understand this null effect after observing it, specifically regarding the role of believability; indeed, if Unpacked participants did not believe they were actually playing with a real partner—and note that a full 40.76% (97 of 238) of them were *non*believers—then this lack of belief should itself reduce support (thereby explaining a weaker effect on our support measure).

Our exclusion results from earlier hint at this possibility. We followed up on these results in two ways.

First, we compared the results of the manipulation check among the 97 nonbeliever/Unpacked participants to the full sample of 253 Packed participants. Indeed, nonbeliever/Unpacked participants reported marginally *lower* support (i.e., less emotion: M = 3.82, SD = 2.66) as compared to Packed participants (M = 4.36, SD = 2.40), $t(348) = 1.80$, $p = .073$, $d = 0.22$; results of a Welch's t test that accounts for these unbalanced groups, $t(159.24) = 1.72$, $p = .088$, $d = 0.21$.

Second, we reconducted all our original analyses while excluding the 97 nonbeliever/Unpacked participants (leaving N = 394; the remaining 141 Unpacked participants [believers only] vs. the full sample of 253 Packed participants). Here, using this nonpreregistered exclusion, our preregistered hypothesis was supported on all measures.

Manipulation Check: Support (Via Emotion). There was a significant effect of our manipulation on support: An independent-samples t test confirmed that Unpacked participants reported higher support (i.e., emotion: M = 4.93, SD = 2.57) than Packed participants (M = 4.36, SD = 2.40), t(392) = 2.22, p = .027, d = 0.23, Welch's t test: t(272.53) = 2.17, p = .031, d = 0.23.

Main Results: Threshold Behavior. The effect of Judgment Type was significant, such that more Unpacked participants (79.43%, 112 of 141) than Packed participants (66.80%, 169 of 253) reported their partner for cheating, $\beta = 0.65$, $SE = .25$, $p = .008$.

Other Variables. Most participants passed the attention checks ("name": 99.49%, 392 of 394; "earnings": 99.24%, 391 of 394) and the honesty check (100%, 394 of 394).

Finally, another question on this front is why the effect on thresholds emerged even when including nonbelievers; in fact, nonbelievers also showed the effect themselves (Nonbeliever/ Unpacked vs. Packed: Effect of Judgment Type on threshold behavior, β = 0.85, SE = .30, p = .004). Presumably, both here and across Experiments 2a–2b–3a–3b, participants who have the full unpacked experience—which includes these nonbelievers—can also summon higher support for passing judgment from sources beyond emotion. As emphasized throughout, we aim to highlight the more general input of perceived support, which may be drawn from many sources beyond the ones we test in the current research.

Taken together, these results resemble those of Experiment 2a, here in terms of *dis*trust. Participants who experienced signs of cheating unfold one by one (vs. merely imagining them all up front) lowered their threshold for accusing the person of cheating.

Next, Experiments 3a–3b again focused on support via emotion—both positive (Experiment 3a) and negative (Experiment 3b) emotion—but we sought to replicate these patterns across further contexts, including tests of *slowing* judgment with *lower* support.

Experiment 3a: Promoting Message-Board Users

In Experiment 3a, participants set a threshold for how many positive message-board interactions a user must post before being invited to be a moderator. Conceptually replicating Experiments 2a–2b, we hypothesized that participants would set softer (i.e., lower) promotion thresholds after seeing one such positive interaction unfold (Unpacked condition) versus merely imagining it all up front (Packed condition), with the former providing more support (via positive emotion).

We also assessed another Unpacked condition whereby participants *also* saw this same interaction as it unfolded (as opposed to imagining it up front)—but we blacked out the emotional content of the user's post. We refer to this as the "Unpacked-Blunted" condition, as blacking out the post in this way should blunt participants' reactions to it—and therefore provide lower unpacked support as compared to our full Unpacking condition. As such, we hypothesized that this manipulation should attenuate the effect (i.e., that Unpacked-Blunted participants might set thresholds that fall somewhere in between our full Unpacked participants and our Packed participants).

Method

Participants. We requested 600 "Cloud Approved" participants from Cloud Research, yielding 599 participants (45.15% women; 27.49% non-White; M_{age} = 40.38, SD_{age} = 12.91) who completed the experiment for $0.75.

Procedure. The experiment followed a 3-cell (judgment type, between subjects: packed vs. unpacked vs. unpacked-blunted) design.

To begin, all participants learned we had allegedly been piloting a "Helpers Forum" over the past few months. In the forum, "Askers" are encouraged to "safely and freely post about problems in their life" while "Responders" can "search these posts and leave replies." Participants learned we were recruiting them to help us identify an optimal threshold for identifying and promoting forum moderators—"truly good" Responders who "leave a kind/encouraging comment and, in turn, seem truly invested in helping and supporting the Asker"—as opposed to falsely identifying such users who may "leave a kind/encouraging comment, but it turns out they were just sarcastically trolling." We informed participants it was very important for us to correctly identify who to promote to the moderating team—which can be best figured out by "track[ing] the same user over time, to see if their comments remain consistently supportive"—and that we were currently crowdsourcing opinions on the optimal threshold for identifying them.

First, to ensure that participants read all the prompts, we required them to summarize their task instructions via an open-ended text box. Next, they learned they had been randomly paired with one such Responder from our database of users, who we identified using the anonymized username "h52315." We then randomly assigned participants to 1 of 3 conditions.

We asked Packed participants: "From the following options, what do you think our policy should be, in terms of promoting this particular Responder to be a Forum Moderator in our Helpers Forum?," and they indicated their threshold from 1 (*1 kind/ encouraging interaction, and h52315 should be officially promoted*) to 10 (*10 kind/encouraging interactions, and h52315 should be officially promoted*), serving as our key dependent variable.

Unpacked participants completed this same dependent variable, but first read that we would show them a randomly selected forum interaction of this user. We showed them an image of an interaction from our alleged forum with "Asker e11386," a student who posts about wanting to give up on school due to various hardships, with Responder h52315 responding with an encouraging comment to not give up (see OSF for all materials). We added a parenthesis to their choice options in the dependent variable to ensure a fair numerical comparison: Option 1 was phrased, "1 kind/encouraging interaction, and h52315 should be officially promoted (i.e., they should be promoted now)," and so on through Option 10, which was phrased, "10 kind/encouraging interactions, and h52315 should be officially promoted (i.e., they should be promoted after 9 more like this").

Participants in a third condition—Unpacked-Blunted participants—followed these same procedures as Unpacked participants, but we blacked out the kind words of Responder h52315.

If differences in threshold-setting are indeed driven by differences in support, then we should find that Unpacked participants set softer (i.e., lower) thresholds than Packed participants, with Unpacked-Blunted participants falling somewhere in between.

After reporting their promotion threshold, all participants rated a manipulation check for support (i.e., emotion): "Based on what we showed you from Responder h52315, how emotionally moving did this particular responder make you feel?," rated from 1 (*felt a little bit moved, given what I saw*) to 10 (*felt extra moved, given what I saw*).

Finally, all participants reported demographic information and completed an attention check regarding their condition (forced choice from 1 of 3 options, describing each condition), and the same honesty check from prior experiments. Also, we asked Unpacked and Unpacked-Blunted participants whether the unpacked interaction they viewed fairly matched our opening summary description of the forum (forced choice: "yes" or "no").

Results and Discussion

Manipulation Check: Support (Via Emotion). First, there was a significant effect of Judgment Type on support, One-Way Analysis of Variance (ANOVA) $F(2, 596) = 111.01$, $p < .001$, $\eta_p^2 = .27$, such that Unpacked participants reported higher support (i.e., emotion: $M = 7.37$, $SD = 2.35$) than Packed participants ($M = 3.95$, $SD = 2.32$), while Unpacked-Blunted participants fell in between ($M = 4.61$, $SD = 2.63$). To tease apart this omnibus effect, we conducted a set of two pairwise contrasts.[2] First, Unpacked participants reported more support than Packed participants and Unpacked-Blunted participants (contrast weights: Packed = +2; Unpacked = −1; Unpacked-Blunted = −1), $t(596) = 9.59$, $p < .001$, $d = 0.79$. Second, Unpacked-Blunted participants reported less support than Unpacked participants (contrast weights: Packed = 0; Unpacked = −1; Unpacked-Blunted = +1), $t(596) = 11.39$, $p < .001$, $d = 0.93$.

Main Results: Threshold Behavior. Next, we conducted an ANOVA with Judgment Type (Packed vs. Unpacked vs. Unpacked-Blunted) as the independent variable and promotion threshold (promoted after 1 kind/encouraging interaction through 10 kind/encouraging interactions) as the dependent variable.

As hypothesized, there was a significant omnibus effect of Judgment Type, $F(2, 596) = 21.25$, $p < .001$, $\eta_p^2 = .07$. Packed participants believed our promotion threshold should be about seven kind interactions ($M = 7.06$, $SD = 2.84$)—yet Unpacked participants set *softer* thresholds, lowering them to about five kind interactions ($M = 5.18$, $SD = 2.98$). Unpacked-Blunted participants fell in between with a threshold of about six kind interactions ($M = 6.36$, $SD = 2.92$)—which was softer than Packed participants but not as soft as Unpacked participants.

A set of two pairwise contrasts confirmed this pattern. First, Unpacked and Unpacked-Blunted participants provided softer (i.e., lower) thresholds than Packed participants (contrast weights: Packed = +2; Unpacked = −1; Unpacked-Blunted = −1), $t(596) = 5.09$, $p < .001$, $d = 0.42$. Second, Unpacked-Blunted participants provided a threshold that was not as soft (i.e., not as low) as Unpacked participants (contrast weights: Packed = 0; Unpacked = −1; Unpacked-Blunted = +1), $t(596) = 4.06$, $p < .001$, $d = 0.33$.

Other Variables. Most participants passed the attention check (96.16%, 576 of 599) and the honesty check (99.33%, 595 of 599). Most Unpacked participants (99.50%, 201 of 202) and Unpacked-Blunted participants (86.57%, 174 of 201) reported that we fairly set their expectations. When rerunning our analyses while excluding participants who failed either check, or reported that unfair expectations (leaving $N = 547$), results are unchanged, omnibus effect of Judgment Type on support: $F(2, 544) = 114.52$, $p < .001$, $\eta_p^2 = .30$; on promotion threshold: $F(2, 544) = 25.03$, $p < .001$, $\eta_p^2 = .08$.

Experiment 3a provides further evidence for our support-based framework. Unpacking a positive interaction led participants to soften their packed promotion thresholds—but participants did so to a lesser degree when that unpacked support was blunted.

Experiment 3b: Banning Message-Board Users

Experiment 3b resembled Experiment 3a but focused on banning thresholds. Participants set a threshold for how many negative message-board interactions (e.g., trolling) a user must commit before being banned. We hypothesized that participants would set harsher (i.e., lower) banning thresholds after seeing one such negative interaction unfold (Unpacked condition) versus merely imagining it all up front (Packed condition), with the former providing more support (via negative emotion).

We also assessed another Unpacked condition, conceptually resembling the Unpacked-Blunted condition in Experiment 3a. Here, participants *also* saw one such negative interaction unfold (as opposed to imagining it up front), but they saw a different negative interaction from the one that our full Unpacked participants saw—one that was weakly negatively charged. We refer to this as the "Unpacked-Weak" condition, as this particular post should elicit weaker reactions among participants—and therefore provide lower unpacked support as compared to our full Unpacking condition. As in Experiment 3a, we hypothesized that this manipulation should attenuate the effect (i.e., that Unpacked-Weak participants might set thresholds that fall somewhere in between our full Unpacked participants and our Packed participants).

Moreover, we further speculated in our preregistration that this Unpacked-Weak condition might even *flip* the effect, given that it uses an entirely different stimulus from the Unpacked condition (as opposed to using the same stimulus but blacking parts out to blunt reactions, as in Experiment 3a). That is, Unpacked-Weak participants might go beyond setting less-harsh thresholds than Unpacked participants (as in Experiment 3a)—here they might even set less-harsh thresholds than *Packed* participants, if the specific stimulus we used happens to provide objectively low (not just relatively lower) support. (Experiments 4–5a–5b will then directly manipulate and test this idea in more controlled settings.)

Method

Participants. We requested 300 "Cloud Approved" participants from Cloud Research, yielding 285 participants (61.05% women; 29.82% non-White; $M_{\text{age}} = 41.87$, $SD_{\text{age}} = 13.03$) who completed the experiment for $0.55.[3]

Procedure. The experiment followed a 3-cell (judgment type, between subjects: packed vs. unpacked vs. unpacked-weak) design.

Participants learned we had allegedly been piloting a message board, and that we wanted to "make sure that we are promoting a healthy conversational culture"; thus, we were recruiting them to "assume the job of monitoring for rude comments." We explained that, sometimes, a user might share something they deem worthy of attention, but others respond harshly and put the user down. Their

[2] Both here in Experiment 3a and also in Experiment 3b—regarding our tests to tease apart omnibus effects on the manipulation check and on the dependent variable—we post hoc decided to tease them apart via these reported sets of pairwise contrasts. In Experiment 3a, we had preregistered to conduct pairwise contrasts, but we did not specify the weights. In Experiment 3b, we had preregistered to conduct independent samples *t*-tests (which yields similar results either way: see OSF).

[3] This requested sample size of Experiment 3b ($N = 300$) was half that of Experiment 3a ($N = 600$) because we conducted Experiment 3b first, with a smaller research budget.

job was to "help us determine how many (i.e., rude comments) we should tolerate before banning someone."

First, all participants read, "Consider the following post from our message board," and saw an image of our alleged forum showing a post from "User eob444." The image depicted a person smiling, with text from the user stating they were proud of themselves for working through a stressful academic program (see OSF for all materials). This post was real (Reddit, 2021), except we photoshopped it into our own alleged forum (without any reference to Reddit). We then randomly assigned participants to 1 of 3 conditions.

Packed participants read: "Let's take a randomly selected user—let's call them New User X—who decided to respond to this post." Without seeing New User X's response, these participants immediately completed the key dependent variable: "How many 'strikes' (i.e., rude comments in response to posts) do you think you would give New User X before banning them from the message board?" and indicated their threshold from 1 (*I'd say, 1 strike and they're out*) to 10 (*I'd say, 10 strikes and they're out*).

Unpacked participants and Unpacked-Weak participants completed this same dependent variable, but before doing so, they read: "We're going to show you how a randomly selected user—let's call them New User X—responded in the following way to this message …" We then showed these participants New User X's response to the post. This too was real, taken from pilot research where we asked other participants to write a rude reply to an online braggart (as written broadly speaking, without reference to this experiment).[4] We then took these replies for use here in Experiment 3b.

For Unpacked participants, we took the Top 3 reply posts from our pilot research (again, see Footnote 4) that generated the strongest emotional reactions among viewers. These participants saw one of these three posts, selected at random, for example, one was:

> You idiot. You absolute buffoon. NOBODY GIVES A SHIT! Please stop wasting everyone's time with your obnoxious gloating

We sampled multiple posts simply for further generalizability (see OSF for all of the posts).

Conversely, for Unpacked-Weak participants, we took the Bottom 3 reply posts that generated the weakest emotional reactions among viewers. These participants saw one of these three posts, selected at random, for example, one was:

> I wish you could see how you come across to other people who see your posts. I think you should really reflect on why you feel the need to brag about every single thing in your life and figure out why your esteem is so low that you need praise for everything you do

Both Unpacked participants and Unpacked-Weak participants then responded to the same dependent variable as Packed participants. We added a parenthesis to their choice options to ensure a fair numerical comparison: Option 1 was phrased, "*I'd say, 1 strike and they're out (they're out now)*," and so on for each option all through option 10, which was phrased, "*I'd say, 10 strikes and they're out (they're out after 9 more strikes"*).

Thus, if differences in threshold-setting are indeed driven by differences in support, then we should find that Unpacked participants set harsher (i.e., lower) thresholds than Packed participants, with Unpacked-Weak participants falling somewhere in between—or, as we reviewed in Experiment 3b's introduction, these Unpacked-Weak participants might even set *less-harsh* (i.e., higher) thresholds than *Packed* participants.

After reporting their banning threshold, all participants rated a manipulation check for support (i.e., emotion): "In seeing [for Packed conditions, "imagining"] the reply post by New User X, how strong were your visceral/emotional reactions?," rated from 1 (*not very strong; less impassioned than I'd usually be about these things*) to 10 (*very strong; more impassioned than I'd usually be about these things*).

Finally, all participants reported demographic information and completed an attention check regarding their assigned condition (forced choice from 1 of 2 options, describing whether or not they saw a specific reply from New User X), and the same honesty check from prior experiments. All participants completed an additional check for whether the experiment fairly previewed the forum and the post (forced choice: "yes" vs. "no").

Results and Discussion

Manipulation Check: Support (Via Emotion). First, there was a significant effect of Judgment Type on support, ANOVA $F(2, 282) = 18.89, p < .001, \eta_p^2 = .12$, such that Packed participants reported less support (i.e., emotion: M = 4.53, SD = 2.09) than Unpacked participants (M = 6.35, SD = 2.30), while Unpacked-Weak participants fell in between (M = 4.76, SD = 2.28). A set of two pairwise contrasts confirmed this pattern. First, Unpacked participants reported more support than Packed and Unpacked-Weak participants (contrast weights: Packed = +2; Unpacked = −1; Unpacked-Weak = −1), $t(282) = 3.63$, $p < .001$, $d = 0.43$. Second, Unpacked-Weak participants reported less support than Unpacked participants (contrast weights: Packed = 0; Unpacked = −1; Unpacked-Weak = +1), $t(282) = 4.96$, $p < .001$, $d = 0.59$.

Main Results: Threshold Behavior. Next, we conducted an ANOVA with Judgment Type (Packed vs. Unpacked vs. Unpacked-Weak) as the independent variable and banning threshold (banned after 1 strike through 10 strikes) as the dependent variable.

As hypothesized, there was a significant omnibus effect of Judgment Type, $F(2, 282) = 11.56, p < .001, \eta_p^2 = .08$. Packed participants believed our banning threshold should be about 3–4 strikes (M = 3.49, SD = 2.02)—yet Unpacked participants set *harsher* thresholds, lowering them by about one strike (M = 2.76, SD = 2.12). Moreover, among the Unpacked-Weak condition, this effect was not just attenuated—it flipped. Unpacked-Weak participants set the *softest* thresholds of all, *raising* them by about one strike compared with Packed participants and by about two strikes compared with Unpacked participants (M = 4.30, SD = 2.50).

[4] We hesitated to craft a rude reply ourselves to use as a stimulus, or to ask others to craft a rude reply for us, as this would likely foster a toned-down, nongenuine response. To circumvent this issue, we conducted two pilot studies (see OSF for full details). First, we asked a separate sample of participants (from the same population) to bring to mind someone they know who brags on the Internet, and to anonymously "let it rip; put the person in their place the way you've always wished you could." They freely typed their reply via an open-ended text box. Second, we then took a random selection from that list of written rude comments, paired them so as to look like they were responding to our actual "User eob444" stimulus, and asked another separate sample (also from this same population) to rate their emotional reactions to seeing this interaction (1 = *not very strong; less impassioned than I'd usually be about these things*, to 10 = *very strong; more impassioned than I'd usually be about these things*). We then ranked these average emotion ratings, and we chose the Top 3 most emotion-evoking replies (M = 6.95, SD = 2.48), and the Bottom 3 least emotion-evoking replies (M = 4.63, SD = 2.35), which we then used in Experiment 3b as reported here in the main text.

A set of two pairwise contrasts confirmed this pattern. First, Unpacked participants provided a lower (i.e., harsher) threshold than Packed participants (contrast weights: Packed = +1; Unpacked = −1; Unpacked-Weak = 0), $t(282) = 2.25$, $p = .025$, $d = 0.27$. Second, Unpacked-Weak participants provided a higher (i.e., softer) threshold than Packed and Unpacked participants (contrast weights: Packed = −1; Unpacked = −1; Unpacked-Weak = +2), $t(282) = 14.14$, $p < .001$, $d = 1.68$.

Readers may wonder why Unpacked-Weak participants softened (i.e., raised) their threshold relative to Packed participants, despite no significant difference on our manipulation check for support (via emotion). As we had noted in our Experiment 2b discussion, unpacking here may invite participants to summon lower (or higher) support from other sources beyond lower (or higher) emotion per se; again, the broader point is not about support-via-emotion but about support-via-many potential inputs (which may be drawn from many sources beyond the ones we assess here). Experiments 4–5a–5b will test some others.

Other Variables. Most participants passed the attention check (90.18%, 257 of 285) and the honesty check (98.60%, 281 of 285), and reported we fairly set their expectations (Packed: 92.31%, 84 of 91; Unpacked: 98.97%, 96 of 97; Unpacked-Weak: 100.00%, 97 of 97). When rerunning our analyses while excluding participants who failed either check, or reported unfair expectations (leaving N = 247), results are unchanged, Omnibus effect of Judgment Threshold on support: $F(2, 244) = 16.75$, $p < .001$, $\eta_p^2 = .12$; on banning threshold: $F(2, 244) = 14.56$, $p < .001$, $\eta_p^2 = .11$.

Experiment 3b builds on Experiment 3a by extending to banning contexts, further generalizing our findings. Unpacking led participants to set different banning thresholds than Packed participants—but in different ways depending on what was unpacked.

Next, we move beyond emotion to other potential inputs into support. We also shift to within-subject designs, with the same participants setting a packed threshold and then responding to an unpacked piece. Throughout, we further test differential effects of unpacking, such that it should hasten *or* slow judgment depending on the support provided by what is unpacked (despite participants knowing about those unpacked possibilities beforehand). Other such inputs into support are how clearly good or bad unpacked behaviors seem (Experiment 4), and how consistently they play out (Experiments 5a−5b)—both of which we assess next.

Experiment 4: Clearness of Goodness/Badness As an Input Into Support

Back in Experiment 3b's Unpacked-Weak condition, recall that we had showed participants a less-nasty "trolling" stimulus (relative to the blatantly nasty stimulus that we had showed Unpacked participants)—which we did in order to test the idea that viewing a less-nasty stimulus should elicit weaker emotional reactions among viewers and thereby provide less support. Taking a step back, however, note that—as we have emphasized throughout—our support-based theorizing is not restricted to the emotion piece itself. Lower (vs. higher) support in Experiment 3b could have also come from other sources—such as from having had the knowledge that, by design, one of these stimuli was clearly less (vs. more) nasty than the other (independent of one's emotional reactions per se).

Experiment 4 tested this idea: How *clearly good or bad* an unpacked event seems might serve as its own input into support that alters threshold judgments accordingly. By "clearly good or bad," we mean the extent to which an unpacked event seems unqualified by alternative explanations for its occurrence, allowing people to feel confident in judging its goodness or badness (Jones & Davis, 1965); its goodness or badness seems obvious.

Unpacked events can indeed vary on clearness in this way; in turn, higher (vs. lower) clearness should provide higher (vs. lower) support, because people tend to draw on perceived clearness of goodness or badness as information in forming judgment. For example, parents are found to give more praise to children who complete chores willingly than to children who complete chores resentfully (Krull et al., 2008), bystanders are found to give more praise to helpers who intervene immediately than to helpers who intervene after deliberation (Critcher et al., 2012), and so on—despite all targets successfully committing the same good behavior. Put in terms of our research, *how* actors end up meeting (or failing to meet) a threshold might affect observers' decisions to pass or withhold judgment anyway, despite those different possibilities being known beforehand—such that acting in clearly good or clearly bad ways (i.e., higher support ways) may elicit "quicker" judgment whereas acting in ambiguously good or ambiguously bad ways (i.e., lower support ways) may elicit "slower" judgment.

Specifically, participants indicated their thresholds for rewarding or punishing a target based on their arrival times (early vs. late) across a series of studies that required on-time arrival. We hypothesized that participants would set higher or lower thresholds depending on the differential support provided by the clearness of what is unpacked.

Method

Participants

We requested 600 "Cloud Approved" participants from Cloud Research, yielding 601 participants (48.09% women; 27.62% non-White; $M_{\text{age}} = 40.96$, $SD_{\text{age}} = 12.18$) who completed the experiment for $0.30.

Procedure

The experiment followed a 2 (judgment type, within subjects: packed vs. unpacked) × 2 (what's unpacked, between subjects: high support vs. low support) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior) design.

Participants learned we would be running another study involving putting workers (from their same population) into pairs and having them schedule 10 video-chat conversations with each other (via Zoom: Howlett, 2022), spread out over a month—and so it is "essential that workers show up to their scheduled times," or else we cannot conduct that session. We therefore explained that we would try to motivate on-time behavior via bonuses and penalties from the study's set pay of $20: that we would "issue a $10 bonus if they end up showing up early across the 10 sessions (making their total pay $30)" and "issue a −$10 penalty if they end up showing up late across the 10 sessions (making their total pay $10)." We further explained that we were "trying to come up with a fair number that, if hit, should trigger the bonus/penalty (i.e., for how many of these 10 sessions do they need to show up early/late to earn the bonus/penalty?)"—so we were recruiting the current participants to share their feedback on what this number should be, which would "inform our decision."

We then randomly assigned participants to their conditions. For ease, we report the Valence conditions separately: We first report the Judge-Good conditions in full, followed by their "Judge-Bad" counterparts.

Judge-Good Conditions

Judge-Good participants evaluated the number of early arrivals needed for the bonus. They read:

> Just like with anything in life, there's a huge variety of ways in which someone can "show up early." For example, one could show up 1 min early versus 15 min early; one could show up early with an eager readiness versus more lazily show up early; and so on.

We then instructed participants to think of all these ways as falling into one of five buckets: "Showing up early with exceptional passion/hardworkingness"; "Showing up early with some passion/hardworkingness"; "Showing up early normally/neutrally"; "Showing up early with some disrespect/fakeness"; and "Showing up early with exceptional disrespect/fakeness"—and

> to keep things fair, assume any of these buckets is always equally likely; if a worker indeed shows up early, and thus gets closer to hitting our promised number for the bonus, assume they could do it in any of these ways, each time.

First, these Judge-Good participants made their *Packed* judgment: "With this in mind, let's say our workers end up showing up early for five of the 10 sessions. Do you think this number (5) should earn them the bonus?" (forced choice: "yes" or "no"). In this packed task, note that participants must consider all the many possible combinations of (those prestated buckets of) "arriving early" in order for them to figure out for themselves whether "5 early arrivals out of 10" would make for a fair threshold for us to implement.

Second, we then put this threshold to the test. These Judge-Good participants then made their *Unpacked* judgment, as randomly assigned to one of two "What's Unpacked" conditions. We told participants that we would show them the actual behavior of a randomly selected pilot worker, "Worker N"—and for all participants, Worker *N* "ended up showing up early, for five of the 10 sessions." Critically, however, we also described *how* each of these early arrivals had played out. For High Support participants, each of these five early arrivals was described as "showed up early with exceptional passion/hardworkingness" (the most positive of our 5 buckets); for Low Support participants, each of these five early arrivals was described as "showed up early with exceptional disrespect/fakeness" (the most negative of our 5 buckets). We then asked all participants: "Do you think this number (5) should earn Worker *N* the bonus?" (forced choice: "yes" or "no").

Thus, all told, note our key test: If participants first indicate that "5 of 10" makes a fair bonus threshold (packed), then we are testing whether they indeed bonus Worker N, as Worker N is simply an unpacked example of someone who meets it. However, we hypothesized that unpacking clearly good versus ambiguously good cases differentially affects support, despite those possibilities being known beforehand—swaying a *larger* number of High Support (i.e., clearly good) participants to bonus Worker *N* relative to their own packed judgment ("quicker to judge"), but a *smaller* number of Low Support (i.e., ambiguously good) participants to do so ("slower to judge").

Judge-Good participants then completed a manipulation check for support, regarding the clearness of the examples of early arrivals that now came to mind: We asked them to "bring to mind a few examples of a randomly selected worker showing up early a few times" and to indicate "how obviously good" this worker's early arrivals are, rated from 1 (*not obviously good*) to 7 (*very obviously good*; with 4 = *normal/average*).

Judge-Bad Conditions

Judge-Bad participants followed this same procedure, except evaluated the number of late arrivals needed for the penalty. They read:

> Just like with anything in life, there's a huge variety of ways in which someone can "show up late." For example, one could show up 1 min late versus 15 min late; one could show up late for good reasons, apologetically versus with no excuses, unapologetically; and so on.

We then instructed them to think of all these ways as falling into one of five buckets: "Showing up late with exceptional rudeness/recklessness"; "Showing up late with some rudeness/recklessness"; "Showing up late normally/neutrally"; "Showing up late with some respect/reasonableness"; and "Showing up late with exceptional respect/reasonableness."

In turn, following the same (converse) phrasings, they first evaluated whether being late to five of the 10 sessions should earn our workers the penalty (forced choice: "yes" or "no"); and then evaluated whether unpacked Worker N, who indeed met this prestated threshold by showing up late for five of the 10 sessions—either with "exceptional rudeness/recklessness" (High Support; clearly bad) or with "exceptional respect/reasonableness (Low Support; ambiguously bad)—should get the penalty (forced choice: "yes" or "no").

Judge-Bad participants then completed the same (converse) manipulation check: We asked them to "bring to mind a few examples of a randomly selected worker showing up late a few times" and to indicate "how obviously bad" this worker's late arrivals are, rated from 1 (*not obviously bad*) to 7 (*very obviously bad*; with 4 = *normal/average*).

Finally, all participants reported demographic information and completed an attention check regarding their judged thresholds (forced choice from 1 of 2 options corresponding to their Valence condition); an attention check regarding what information was unpacked (forced choice from 1 of 2 options, describing the clear vs. ambiguous texts); and the same honesty check and "fairly previewed" check from prior experiments.

Results and Discussion

Manipulation Check: Support (Via Clearness)

First, there was a significant effect of What's Unpacked on support (i.e., clearness), Univariate GLM $F(1, 597) = 430.99$, $p < .001$, $\eta_p^2 = .42$. High Support participants were more likely to have more supportive (i.e., clearly good or bad) examples top of mind ($M_{\text{all}} = 6.17$, $SD_{\text{all}} = 1.06$) as compared with Low Support participants ($M_{\text{all}} = 3.75$, $SD_{\text{all}} = 1.88$). This effect emerged among Judge-Good participants, $M_{\text{HighSupport}} = 5.80$, $SD_{\text{HighSupport}} = 1.21$ vs. $M_{\text{LowSupport}} = 2.99$, $SD_{\text{LowSupport}} = 1.74$; pairwise $F(1, 597) = 300.42$,

$p < .001$, $\eta_p^2 = .34$, and Judge-Bad participants, $M_{\text{HighSupport}} = 6.50$, $SD_{\text{HighSupport}} = 0.77$ vs. $M_{\text{LowSupport}} = 4.58$, $SD_{\text{LowSupport}} = 1.66$; pairwise $F(1, 597) = 143.89$, $p < .001$, $\eta_p^2 = .19$; see OSF for remaining output, which is incidental.

Main Results: Threshold Behavior

Next, we conducted a repeated-measures Binary Logistic Regression via SPSS GEE entering participant as a subject variable; Judgment Type (Packed vs. Unpacked) as a within-subject factor; What's Unpacked (High Support vs. Low Support) and Valence of Judged Behavior (Good Behavior vs. Bad Behavior) as between-subject factors; and threshold behavior (yes-reward/punish vs. no-reward/punish) as the dependent variable.

As hypothesized, there was the critical two-way interaction between Judgment Type and What's Unpacked, Wald = 81.51, $df = 1$, $p < .001$—meaning participants changed their behavior upon moving from packed judgment to unpacked judgment, but *how* they did depended on the support provided by *what* was unpacked. This effect held across Valence, as indicated by null three-way interaction, Wald = .046, $df = 1$, $p = .830$ (see OSF for remaining output, which is incidental). Figure 2 plots the results across each Valence.

Among Judge-Good participants, pairwise comparisons reveal that although 47.14% of these High Support participants (66 of 140) initially established a packed threshold of five early arrivals, this increased to 65.71% (92 of those same 140 participants) when they judged an unpacked exemplar who hit this threshold in a *clearly good way*, $SE = .041$, $df = 1$, $p < .001$, 95% Wald CI_{diff} [−0.27, −0.11]; "no-judge" participants violated their threshold to become "yes-judge" participants ("quicker to judge"). However, this effect *flipped* among these Low Support participants. Although 47.13% of them (74 of 157) initially established a packed threshold of five early arrivals, this number decreased to 19.11% (30 of those same 157 participants) when they judged an unpacked exemplar who indeed also hit this same threshold, but did so in an *ambiguously good way*, $SE = .038$, $df = 1$, $p < .001$, 95% Wald CI_{diff} [0.21, 0.35]; "yes-judge" participants violated their threshold to become "no-judge" participants ("slower to judge").[5]

The same patterns emerged among Judge-Bad participants: Although 83.75% of these High Support participants (134 of 160) initially established a packed threshold of five late arrivals, this number increased to 92.50% (148 of those same 160 participants) when they judged an unpacked exemplar who hit this threshold in a *clearly bad way*, $SE = .027$, $df = 1$, $p = .001$, 95% Wald CI_{diff} [−0.14, −0.03] ("quicker to judge"). However, this effect *flipped* among these Low Support participants: Although 88.19% of them (127 of 144) initially established a packed threshold of five late arrivals, this number decreased to 70.83% (102 of those same 144 participants) when they judged an unpacked exemplar who indeed also hit this same threshold, but did so in an *ambiguously bad way*, $SE = .036$, $df = 1$, $p < .001$, 95% Wald CI_{diff} [0.10, 0.24] ("slower to judge").

Other Variables

Most participants passed the attention checks (valence: 97.00%, 583 of 601; unpacking: 89.35%, 537 of 601) and the honesty check (99.00%, 595 of 601), and reported we fairly set their expectations (High Support: 94.67%, 284 of 300; Low Support: 93.36%, 281 of 301). When rerunning our analyses while excluding participants who failed any of these checks, or reported unfair expectations (leaving $N = 496$), results are unchanged, Main effect of What's Unpacked on support: $F(1, 492) = 505.80$, $p < .001$, $\eta_p^2 = .51$; interaction between Judgment Type and What's Unpacked on threshold behavior, Wald = 87.76, $df = 1$, $p < .001$.

Experiment 4 extends our findings to include another input of support. Participants violated their own thresholds—becoming both "quicker to judge" and "slower to judge"—depending on the level of support of what was unpacked (here via how clearly good or bad the unpacked behaviors seemed).

Next, we assess a third input into support: the extent to which unpacked realities provide *consistent* or *inconsistent* evidence.

Experiments 5a−5b: Consistency As an Input Into Support

The order and structure of how unpacked events unfold should also bear on support (and thus on threshold asymmetries). For example, streaks of behaviors may sway judgment more than when those same behaviors are spread out over a longer mixed window of ups and downs. Consider a manager who gives an employee "three strikes this month" before reprimand; an employee who brazenly commits two strikes right at Day 1 may receive surprise punishment then and there, whereas one who behaves well for a full 28 of those days, with three scattered stumbles along the way, may enjoy surprise reprieve. Higher (vs. lower) *consistency*—the extent to which unpacked behaviors adhere together (Kelley, 1967)—should provide higher (vs. lower) support, because people tend to draw on perceived consistency as information for forming judgment. For example, various studies suggest people indeed care about streaks and weight them highly in judgment, even in cases when such streaks reflect random data (Ayton & Fischer, 2004; Carlson & Shu, 2007; Croson & Sundali, 2005; Gilovich et al., 1985).

Experiments 5a−5b tested whether the higher (vs. lower) support provided by higher (vs. lower) unpacked consistency affects judgment thresholds—with higher support producing "quicker" judgment but lower support producing "slower" judgment.

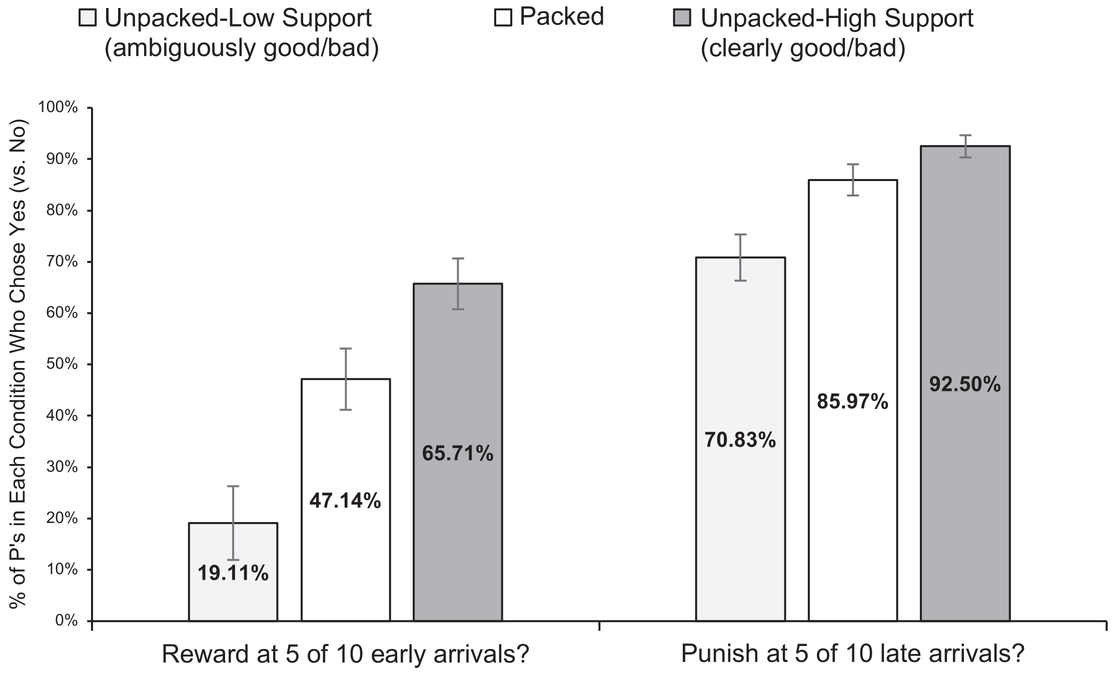
Experiment 5a: Issuing Reward and Punishment Despite Others Not Earning It ("Quicker to Judge")

In Experiment 5a, participants judged others' behavior across economic games involving helping versus harming a partner. First, all participants indicated their thresholds for the number of helpful versus harmful games others must play to elicit reward versus punishment. Then, all participants judged someone who *fell short* of this threshold. We hypothesized that participants would be compelled to act anyway ("quicker to judge") to the extent the target's

[5] By virtue of our fully randomized design, note that the two packed judgments within a given valence—here, for example, Judge-Good/High Support/Packed (47.14%) and Judge-Good/Low Support/Packed (47.13%)—are made *prior* to the randomly assigned unpacking manipulation that distinguishes those participants. In Figure 2, we average them for visual ease (resulting in the middle white Reward bar of "47.14%"). We also do this for the two Judge-Bad/Packed conditions (the middle white Punish bar of "85.97%" is the average of Judge-Bad/High Support/Packed and Judge-Bad/Low Support/Packed). We analyze each of these sets of conditions separately in the main text.

Figure 2

Experiment 4: Percentage of Participants in Each Condition Who Acted on the Threshold, as a Function of Support



Note. Error bars ±1 standard error.

behaviors in other games provided consistent support (in favor of acting early), despite knowing about those possibilities beforehand.

Method

Participants. We requested 600 participants from Prolific Academic, yielding 606 participants (67.99% women; 21.62% non-White; $M_{\text{age}} = 37.64$, $SD_{\text{age}} = 13.32$) who completed the experiment for $0.48.

Procedure. The experiment followed a 2 (what's unpacked, between subjects: high support vs. low support) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior) design.

Participants learned that we would be running another study involving putting workers (from their same population) into pairs and having them play economic games with each other—with each player getting 10 turns, and at each turn being able to "help" or "harm" their partner—and that we wanted to "motivate players to behave maximally helpful and minimally harmful." We therefore explained that we would try to motivate this behavior via bonuses and penalties from the study's set pay of $10: that we would "issue a $5 bonus if they end up being helpful across their 10 turns (making their total pay $15)" and "issue a −$5 penalty if they end up being harmful across their 10 turns (making their total pay $5)." We further explained that we were "trying to come up with a fair number of helpful/harmful behaviors that, if hit, should trigger the bonus/penalty (i.e., for how many of these 10 turns do they need to help/harm their partner to earn the bonus/penalty?)"—so we were recruiting the current participants to share their feedback on what this number should be, which would "inform our decision."

All participants then read the exact ways in which participants could treat each other in each of these 10 turns, falling into one of seven buckets: "Give $1.00 to their partner (help, big)"; "Give $0.50 to their partner (help, medium)"; "Give $0.10 to their partner (help, small)"; "Skip turn (neutral)"; "Take $0.10 from their partner (harm, small)"; "Take $0.50 from their partner (harm, medium)"; or "Take $1.00 from their partner (harm, big)."

We then randomly assigned participants to "Judge-Good" or "Judge-Bad" conditions—with participants in all conditions evaluating a target who *fell short* of their own threshold. For ease, we report the Valence conditions separately: We first report the Judge-Good conditions in full, followed by their "Judge-Bad" counterparts.

Judge-Good Conditions. Judge-Good participants evaluated the number of "big helps" needed for the bonus. Then, they made their *Packed* judgment:

> Out of their 10 turns, what's the minimum number of moves of "give $1.00 to their partner (help, big)" that should earn our workers the bonus? (assume the remaining turns can play out it any other combination of the other kinds of moves).

They indicated their threshold via 10-point scale from 1 (*If 1 of their 10 moves involves "give $1.00 to their partner (help, big)*" that should earn them the bonus) to 10 (*If 10 of their 10 moves involves "give $1.00 to their partner (help, big)*" that should earn them the bonus). In this packed task, note that participants must consider all the many possible combinations of those prestated other moves in order to figure out what number of "big helps" would make for a fair threshold for us to implement. As it turned out, these participants set this threshold at about five, such that workers get the reward if a minimum of $M = 5.25$ of their 10 turns were "big helps"; $SD = 2.31$, min = 1, max = 10. At least one participant chose each of the 10 thresholds.

Second, we put this threshold to the test. These Judge-Good participants then made their *Unpacked* judgment, as randomly assigned to one of two "What's Unpacked" conditions. We told participants that we would show them the actual behavior of a randomly selected pilot worker, "Worker H"—and for all participants, Worker H "ended up 'giving $1.00 to their partner (help, big)' for X of their 10 moves." For placeholder X, each participant saw one move *less* than their own chosen threshold; for example, if a participant stated that five of these "big helps" should get the bonus (regardless of how the remaining 5 play out), then that participant would learn here that Worker H ended up doing only four "big helps."

Critically, participants then moved to the *Unpacked* stage. We showed participants all 10 of Worker H's moves, listed in randomized order. The list always included the X number of "big helps"—one less than participants' threshold. However, the other moves on the list varied by condition. For these High Support participants, we drew each of their remaining moves—the number of which depended on participants' self-set threshold—from the other two kinds of *helpful* moves (randomly drawn from this pool of 2, with replacement); either "Give $0.50 to their partner (help, medium)" or "Give $0.10 to their partner (help, small)." For these Low Support participants, we drew each of their remaining moves—the number of which depended on participants' self-set threshold—from *all* of the other six kinds of moves (randomly drawn from this pool of 6, with replacement), which includes these two other helpful kinds but also includes playing neutrally and playing harmfully. We then asked all participants: "Do you think this should earn Worker H the bonus?" (forced choice: "yes" or "no").

Thus, all told, note our key test: We are testing whether participants indeed say "No" to bonusing Worker H, given that they all evaluated someone who fell short of their own preset threshold (by 1 move). However, we hypothesized that unpacking consistent versus inconsistent cases differentially affects support—swaying a *larger* number of High Support (i.e., consistent) participants to bonus Worker H relative to their own packed judgment ("quicker to judge"), but a *smaller* number of Low Support (i.e., inconsistent) participants to do so ("slower to judge").

Judge-Good participants then completed a manipulation check for support, regarding the consistency of a worker's 10 moves that now came to mind: We asked them to "bring to mind a few examples of a randomly selected worker making 10 moves in our games" and to indicate "how consistently helpful" this worker's 10 moves are, rated from 1 (*not consistently helpful*) to 7 (*very consistently helpful*; with 4 = *normal/average*).

Judge-Bad Conditions. Judge-Bad participants followed this same procedure, except evaluated the number of "big harms" needed for the penalty. They read:

> Out of their 10 turns, what's the minimum number of moves of "take $1.00 from their partner (harm, big)" that should earn our workers the penalty? (assume the remaining turns can play out it any other combination of the other kinds of moves)

As it turned out, these participants set this threshold at about three, such that workers get the penalty if a minimum of $M = 3.51$ of their 10 turns were "big harms"; $SD = 2.10$, min = 1, max = 10. At least one participant chose each of the 10 thresholds, except a threshold of nine big harms; no participant chose a threshold of nine.

In turn, following all the same (converse) phrasings, they then evaluated whether unpacked Worker H, who always fell short of their penalty-threshold by one move, should get the penalty (forced choice: "yes" or "no")—with Worker H's remaining moves either being drawn only from the other harmful moves (High Support; consistent) or from all moves, including playing neutrally and playing helpfully (Low Support; inconsistent).

Judge-Bad participants then completed the same (converse) manipulation check: We asked them to "bring to mind a few examples of a randomly selected worker making 10 moves in our games" and to indicate "how consistently harmful" this worker's 10 moves are, rated from 1 (*not consistently harmful*) to 7 (*very consistently harmful*; with 4 = *normal/average*).

Finally, all participants reported demographic information and completed an attention check regarding their judged thresholds (forced choice from 1 of 2 options corresponding to their Valence condition); an attention check regarding what information was unpacked (forced choice from 1 of 2 options, describing the consistent vs. inconsistent texts); and the same honesty check and "fairly previewed" check from prior experiments.

Results and Discussion

Manipulation Check: Support (Via Consistency). First, there was a significant effect of What's Unpacked on support (i.e., consistency), Univariate GLM $F(1, 602) = 249.75$, $p < .001$, $\eta_p^2 = .29$. High Support participants were more likely to have supportive (i.e., consistent) examples top of mind ($M_{\text{all}} = 5.89$, $SD_{\text{all}} = 1.23$) as compared with Low Support participants ($M_{\text{all}} = 4.12$, $SD_{\text{all}} = 1.52$). This effect emerged among Judge-Good participants, $M_{\text{HighSupport}} = 5.75$, $SD_{\text{HighSupport}} = 1.25$ vs. $M_{\text{LowSupport}} = 4.23$, $SD_{\text{LowSupport}} = 1.58$; pairwise $F(1, 602) = 93.53$, $p < .001$, $\eta_p^2 = .13$, and Judge-Bad participants, $M_{\text{HighSupport}} = 6.03$, $SD_{\text{HighSupport}} = 1.19$ vs. $M_{\text{LowSupport}} = 4.02$, $SD_{\text{LowSupport}} = 1.45$; pairwise $F(1, 602) = 160.61$, $p < .001$, $\eta_p^2 = .21$; see OSF for remaining output, which is incidental.

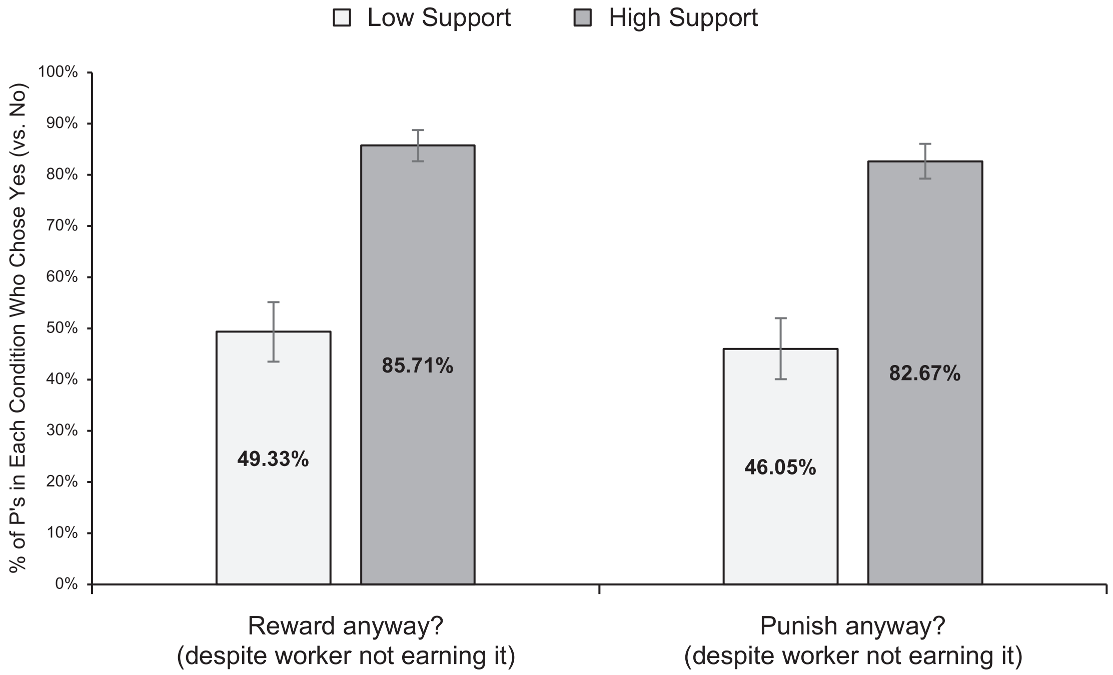
Main Results: Threshold Behavior. Next, we conducted a Binary Logistic Regression with What's Unpacked (High Support vs. Low Support) and Valence of Judged Behavior (Good Behavior vs. Bad Behavior) as between-subjects factors, and threshold behavior (yes-reward/punish vs. no-reward/punish) as the dependent variable.

As hypothesized, there was the critical main effect of What's Unpacked, $\beta = 1.62$, $SE = .61$, $p = .008$—which held across Valence (null interaction: $\beta = 0.10$, $SE = .39$, $p = .802$; main effect of Valence, $\beta = 0.03$, $SE = .56$, $p = .952$). Among Judge-Good conditions, more High Support participants (85.71%, 132 of 154) than Low Support participants (49.33%, 74 of 150) chose to reward Worker H anyway, despite them failing to meet the positive threshold, $\beta = 1.82$, $SE = .28$, $p < .001$; among Judge-Bad conditions, more High Support participants (82.67%, 124 of 150) than Low Support participants (46.05%, 70 of 152) chose to punish Worker H anyway, despite them avoiding the negative threshold, $\beta = 1.72$, $SE = .27$, $p < .001$. Figure 3 plots the results across each Valence.

Further Insights. First, readers may wonder why ~50% of Low Support participants *also* "acted anyway." This finding echoes Experiment 1 (and others) whereby simply experiencing unpacked cases can elicit basic support. Moreover, as we randomly drew from all six "moves" for these participants, some may have seen supportive

Figure 3

Experiment 5a: Percentage of Participants in Each Condition Who Acted on the Threshold, Despite the Worker Not Passing It—As a Function of Support



Note. Error bars ±1 standard error.

(i.e., consistent) unpacking (including identical draws to High Support participants). From this view, having some variety should elicit *higher* "act-anyway's" relative to obviously disqualifying exemplars (e.g., presumably, ~0% of judges would reward a worker who fails to earn it, *plus* is always harmful—less than the ~50% who reward when this worker's other behaviors are mixed, as we find here). As intended, the key test is the relative difference between these particular Low Support versus High Support conditions on average.

Second, readers may wonder how this effect varies across participants' self-set thresholds. Figures 4A–4B plot these results (Figure 4A, reward; Figure 4B, punishment). As can be seen, the effect appears to diminish among higher set thresholds; indeed, when entering Set Threshold (1 to 10) into the model as a factor, there is a significant two-way interaction between this factor and What's Unpacked, $\beta = 0.59$, $SE = .28$, $p = .039$ (null three-way interaction with Valence, $\beta = 0.28$, $SE = .18$, $p = .122$; see OSF for remaining output). This result further validates our manipulation: There is decreasing room for consistency as the number of spots for it decreases, and therefore the effect *should* decrease as thresholds increase. In any case, this analysis is exploratory, not preregistered, and should be interpreted with caution given small and unequal cell sizes (see also Figures 4A–4B caption).

Other Variables. Most participants passed the attention checks (valence: 96.86%, 587 of 606; unpacking: 87.95%, 533 of 606) and the honesty check (99.01%, 600 of 606), and reported we fairly set their expectations (High Support: 91.12%, 277 of 304; Low Support: 94.70%, 286 of 302). When rerunning our analyses while excluding participants who failed any of these checks, or reported unfair expectations (leaving $N = 480$), results are unchanged, Main effect of What's Unpacked on support: $F(1, 476) = 299.42$, $p < .001$, $\eta_p^2 = .39$; on threshold behavior, $\beta = 1.63$, $SE = .69$, $p = .018$.

Experiment 5a extends our findings to include another input of support: consistency. All participants evaluated a target who fell short of their self-set thresholds—yet they became more likely to reward or punish them anyway ("quicker to judge") when the target's unpacked behaviors provided higher support (here via perceived consistency).

Experiment 5b: Withholding Reward and Punishment Despite Others Earning It ("Slower to Judge")

Experiment 5b tested the converse of Experiment 5a. Participants may be swayed to *withhold* reward and punishment—despite a target meeting their preset threshold—again depending on the consistency of unpacked support (here, in favor of withholding), and again despite knowing those possibilities beforehand.

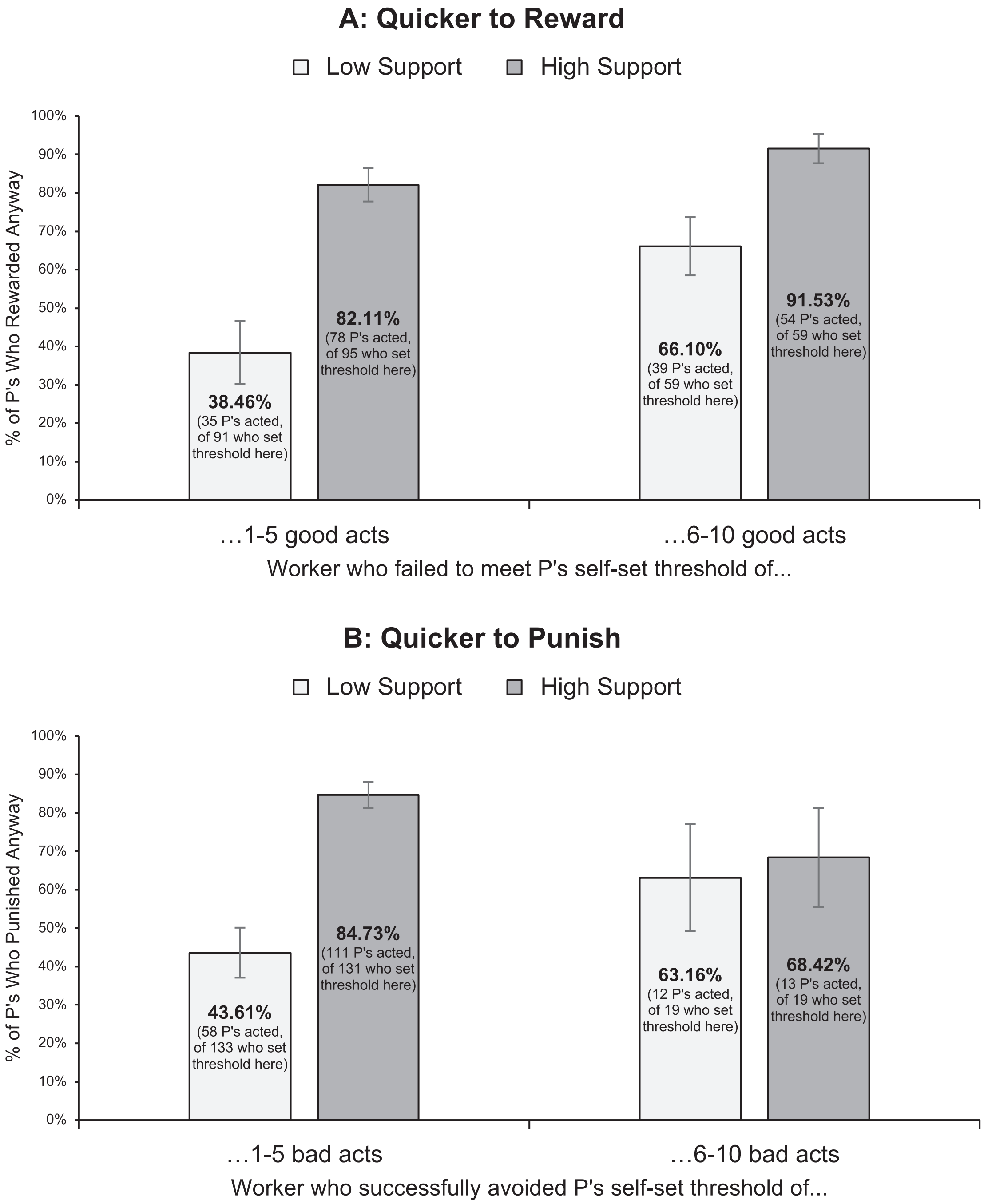
Method

Participants. We requested 600 participants from Prolific Academic, yielding 619 participants (71.57% women; 27.95% non-White; $M_{\text{age}} = 35.52$, $SD_{\text{age}} = 13.38$) who completed the experiment for $0.48.

Procedure. The experiment followed a 2 (what's unpacked, between subjects: high support vs. low support) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior) design.

Figure 4

(A, B) Experiment 5a: Same Results as Shown in Figure 3, Except Split by Participants' Self-Set Thresholds



Note. In the experiment, all participants first indicated their own self-set threshold from 1 to 10 acts; participants then learned that the worker fell short of this threshold by one act. Depicted is the percentage of participants who chose "yes" (vs. "no") to act on that threshold anyway, despite the worker missing it—split between those who set a threshold from 1–5 acts versus 6–10 acts (presented like this for visual ease; for each threshold level, see Supplemental Figures S1A–S1B). The key pattern to note is that the effect appears to diminish at higher (6–10 acts) versus lower (1–5 acts) thresholds. Error bars ±1 standard error.

The study was identical to Experiment 5a, with two key changes. First, upon seeing unpacked Worker H's 10 turns, Worker H always *met* participants' preset threshold. For example, if a participant indicated that five "big helps" (or 5 "big harms") should earn our workers the bonus (or the penalty), they then learned that Worker H exhibited exactly five of them—and so we tested whether they give the bonus (or penalty). As it turned out, Judge-Good participants set this reward-threshold at $M = 5.32$,

SD = 2.36, min = 1, max = 10; Judge-Bad participants set this punishment-threshold at M = 3.59, SD = 2.14, min = 1, max = 10. At least 1 participant chose each of the 10 thresholds, within each Valence.

Second, we flipped the random draw of the other behaviors—the composition of which comprises our consistent versus inconsistent manipulation—to account for this successful meeting of the threshold. For High Support participants, we drew each of their remaining moves—the number of which depended on participants' self-set threshold[6]—from *all* of the other six kinds of moves (randomly drawn from this pool of six, with replacement); but for Low Support participants, we drew each of their remaining moves—the number of which depended on participants' self-set threshold—from the two strongest moves from the *opposite* valence (randomly drawn from this pool of 2, with replacement). That is, for their remaining moves, Judge-Good/Low Support participants only saw "Take $1.00 from their partner (harm, big)" or "Take $0.50 from their partner (harm, medium)"; for their remaining moves, Judge-Bad/Low Support only saw "Give $1.00 to their partner (help, big)" or "Give $0.50 to their partner (help, medium)."

Note how this change allows us to test the input of consistency—the extent to which unpacked behaviors adhere together to provide support—in ways that go beyond literal high or low variance per se. Here, we compare the consequences of hitting a reward-threshold (for example) in a mix of good and bad behaviors versus hitting a reward-threshold while only doing bad behaviors—with the hypothesis that the latter should provide less support, swaying participants to *withhold* reward anyway.

Finally, all participants completed the same other checks as in Experiment 5a.

Results and Discussion

Manipulation Check: Support (Via Consistency). First, there was a significant effect of What's Unpacked on support (i.e., consistency), Univariate GLM $F(1, 615) = 46.76, p < .001, \eta_p^2 = .07$. Low Support participants were less likely to have supportive (i.e., consistent) examples top of mind ($M_{\text{all}} = 3.86, SD_{\text{all}} = 1.73$) as compared with High Support participants ($M_{\text{all}} = 4.74, SD_{\text{all}} = 1.47$). This effect emerged among Judge-Good participants, $M_{\text{LowSupport}} = 4.18, SD_{\text{LowSupport}} = 1.81$ vs. $M_{\text{HighSupport}} = 4.77, SD_{\text{HighSupport}} = 1.55$; pairwise $F(1, 615) = 10.74, p = .001, \eta_p^2 = .02$, and Judge-Bad participants, $M_{\text{LowSupport}} = 3.55, SD_{\text{LowSupport}} = 1.58$ vs. $M_{\text{HighSupport}} = 4.70, SD_{\text{HighSupport}} = 1.38$; pairwise $F(1, 615) = 41.10, p < .001, \eta_p^2 = .06$; see OSF for remaining output, which is incidental.

Main Results: Threshold Behavior. Next, we conducted a Binary Logistic Regression with What's Unpacked (High Support vs. Low Support) and Valence of Judged Behavior (Good Behavior vs. Bad Behavior) as between-subjects factors, and threshold behavior (yes-reward/punish vs. no-reward/punish) as the dependent variable.

As hypothesized, there again was the critical main effect of What's Unpacked, $\beta = 1.82, SE = .61, p = .003$—which again held across Valence (null interaction: $\beta = 0.36, SE = .39, p = .355$; main effect of Valence, $\beta = 0.66, SE = .56, p = .242$). Among Judge-Good conditions, more Low Support participants (36.84%, 56 of 152) than High Support participants (16.23%, 25 of 154) chose to withhold reward from Worker H, despite them meeting the positive threshold, $\beta = 1.10, SE = .28, p < .001$; among Judge-Bad conditions, more Low Support participants (43.95%, 69 of 157) than High Support participants (15.38%, 24 of 156) chose to withhold punishment from Worker H, despite them meeting the negative threshold, $\beta = 1.46, SE = .27, p < .001$. Figure 5 plots the results across each Valence.

Further Insights. First, as in Experiment 5a, readers may wonder why ~15% of High Support participants *also* "withheld anyway." Again, such a finding is consistent with our framework. From this view, having some variety should elicit *higher* "withhold-anyway's" relative to obviously qualifying exemplars (e.g., presumably, ~0% of judges would withhold reward from a worker who does enough to earn it, *plus* is always helpful—less than the ~15% who withhold when this worker's other behaviors are mixed [as we find here]). The key test is between these Low Support versus High Support conditions.

Second, as in Experiment 5a, we plotted these results across self-set thresholds (see Figures 6A–6B). The effect appears to diminish among higher set thresholds (two-way interaction between Set Threshold [1–10] and What's Unpacked, $\beta = 0.66, SE = .35, p = .062$; three-way interaction with Valence, $\beta = 0.30, SE = .23, p = .20$; see OSF for remaining output). Again, this analysis should be read with caution (see also Figures 6A–6B caption).

Other Variables. Most participants passed the attention checks (valence: 96.28%, 596 of 619; unpacking: 69.31%, 429 of 619) and the honesty check (98.55%, 610 of 619), and reported we fairly set their expectations (Low Support: 84.79%, 262 of 309; High Support: 90.00%, 279 of 310). When rerunning our analyses while excluding participants who failed any of these checks, or reported unfair expectations (leaving $N = 366$), results are unchanged, Main effect of What's Unpacked on support: $F(1, 362) = 91.45, p < .001, \eta_p^2 = .20$; on threshold behavior, $\beta = 2.57, SE = .88, p = .004$.

Experiment 5b complements our findings from Experiment 5a. All participants evaluated a target who met their preset threshold for earning reward or punishment—yet they became more likely to *withhold* reward or punishment ("slower to judge") when the target's unpacked behaviors provided lower support (here via perceived [in]consistency).

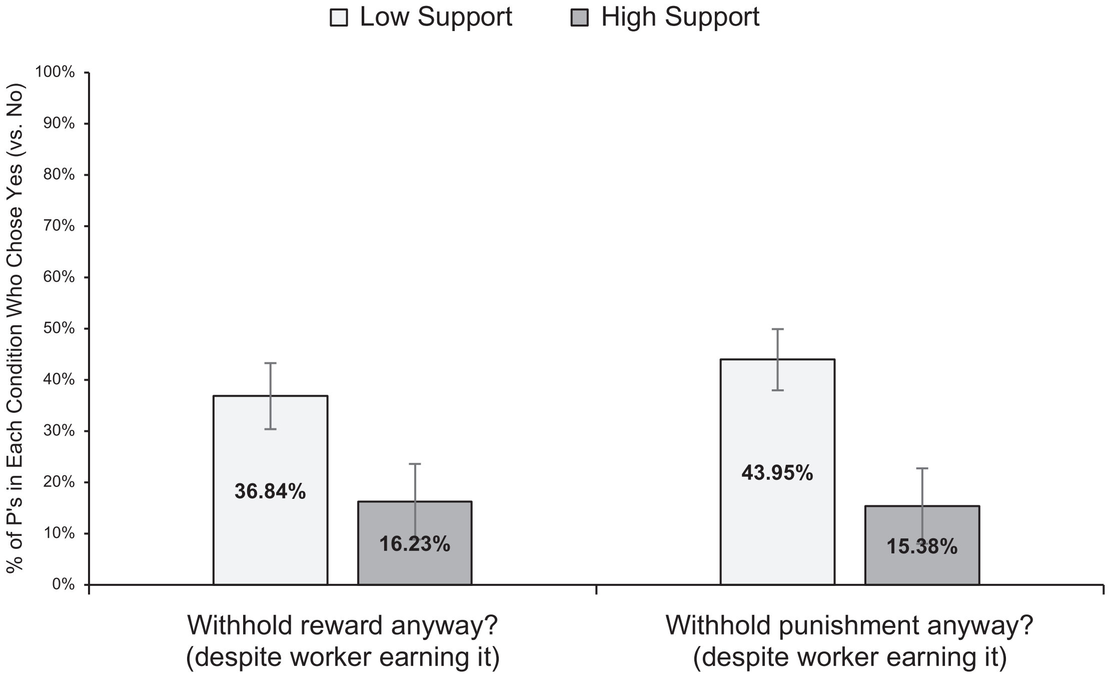
Experiments 6a–6b: So What?

Finally, Experiments 6a–6b shift to assessing downstream consequences. A discrepancy between preset versus followed thresholds could cause social conflict and miscommunication to the extent the other party lacks the unpacked knowledge to contextualize the violation. For example, when a manager gives an employee "three strikes" before reprimand, but then reprimands the employee after two strikes, both the employee and third-party observers might be tempted to conclude the manager is being unfair or hypocritical in

[6] This design creates a complication among participants who preset a threshold of 10, because the manipulation occurs at the remaining behaviors beyond one's threshold but these participants do not see any such behaviors; thus, they do not receive the manipulation. As it turned out, 6.62% of participants (41 of 619) preset a threshold of 10 (these 41 participants were similarly distributed across our key comparison conditions: Judge-Bad/Low Support, n = 5 of 157 vs. Judge-Bad/High Support, n = 6 of 156; Judge-Good/Low Support, n = 15 of 152 vs. Judge-Good/High Support, n = 15 of 154). In any case, note that—if anything—including these participants in our analyses (which we do, reported in the Results section of this study) provides a more *conservative* test of our hypothesis, as their inclusion dilutes any effect of the manipulation. To this point, results hold (and indeed are directionally stronger) when excluding these participants (see OSF).

Figure 5

Experiment 5b: Percentage of Participants in Each Condition Who Withheld Acting on the Threshold, Despite the Worker Passing It—As a Function of Support



Note. Error bars ±1 standard error.

their behavior and so cannot be trusted the next time—whereas the actor themselves (e.g., the manager) may disagree, assuming they have fuller access into their underlying rationale for the violation (Jones & Nisbett, 1971) and are more motivated to justify discrepancies between how they should behave versus how they actually behave (Batson et al., 1999; Valdesolo & DeSteno, 2007).

We tested such outcomes via people's actual experiences with threshold violations (Experiment 6a) and via more general evaluations in controlled settings (Experiment 6b).

Experiment 6a: Consequences of Threshold Violations (Actual Experiences)

First, Experiment 6a assessed people's own experiences, both in terms of being on the receiving end of threshold violations and also simply noticing them in the world. We hypothesized that such violations have had adverse effects (e.g., on violators' reputations).

Method

Participants. We requested 200 successfully screened participants (see Procedure section) from our university's subject pool, which is open to diverse populations beyond students (spanning, e.g., university staff, local community members, and peer participants from across the globe)—yielding 206 successfully screened participants (64.08% women; 56.80% non-White; $M_{\text{age}} = 31.99$, $SD_{\text{age}} = 11.57$) who completed the experiment for $1.00 (additional demographics: 37.38% were university students; 79.61% were from the United States; 78.64% had at least a college degree).

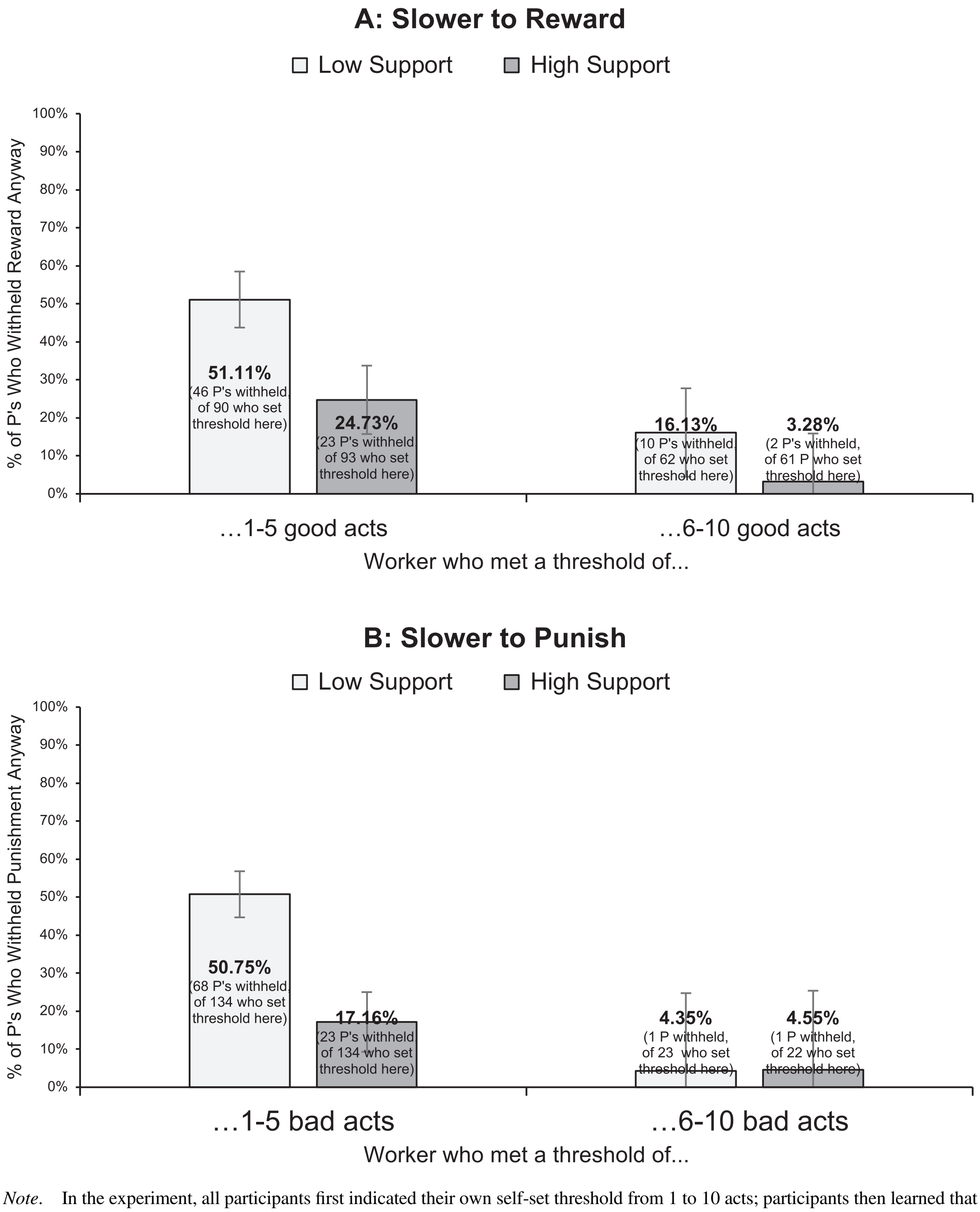
Procedure. The experiment followed a 2 (others' threshold behavior, within subjects: adhere vs. violate) × 2 (one's own role, within subjects: recipient vs. observer) design. Throughout the main text, we will refer to "violations"—but we did not use this word nor related ones in our actual study so as to avoid negative connotation.

First, all participants reported demographic information and completed our screening procedure: We asked whether they could bring to mind four unique experiences (random order; forced choice for each, "yes" vs. "no") involving people from their "actual lives/experiences (e.g., not celebrities/public figures/etc.)." For two of these four experiences—*self-adhere* and *self-violate*—we asked participants if *they themselves* have been on the receiving end of a person who "expected you to behave according to some 'threshold' (e.g., '3 more bad behaviors like that and something bad will happen!' or '3 more good behaviors like that and something good will happen!')." It could be anything along those lines (e.g., bad or good; beyond a literal "3"; involving "anyone, e.g., family, friends, peers, teachers, bosses, colleagues, policymakers")—so long as what transpired was that the person ended up following this threshold (they "acted on their mark"; *self-adhere*), or they ended up *not* following this threshold (they "acted sooner or later than their mark"; *self-violate*). The other two of these four experiences—*observer-violate* and *observer-adhere*—were identical, except they pertained to *someone else* being on the receiving end of each outcome (i.e., times when participants themselves were outside observers).

Only those participants who indicated "yes" to having had all four experiences proceeded to take the study, in which they evaluated each of them (one at a time in randomized order) via three critical dependent measures (also presented in randomized order). First,

Figure 6

(A, B) Experiment 5b: Same Results as Shown in Figure 5, Except Split by Participants' Self-Set Thresholds



Note. In the experiment, all participants first indicated their own self-set threshold from 1 to 10 acts; participants then learned that the worker indeed met this threshold. Depicted is the percentage of participants who withheld acting anyway (i.e., those who chose "no" [vs. "yes"]), despite the worker hitting it—Split between those who set a threshold from 1–5 acts versus 6–10 acts (again, presented like this for visual ease; for each threshold level, see Supplemental Figures S2A–S2B). The pattern to note is that the effect appears to diminish at higher (6–10 acts) versus lower (1–5 acts) thresholds. Error bars ±1 standard error.

participants reported how this experience "made them feel," rated from −5 (*very negative*) to 5 (*very positive*). Second, they reported how this experience "made them view the person" (the threshold setter), via same scale. Third, we informed participants of an unrelated upcoming study of ours that promised to be "fun and rewarding" for those who take it, and we asked: "Should we invite this person to participate? You wouldn't be identified in any way; we will coordinate details/logistics for doing this at the end of today's study" (forced choice: "yes" vs. "no"). This was true: At the end of the study, we asked any participant who said "yes" to report their target's contact information (email address, cell number, or social media profile). Lab staff then invited all listed targets to join our subject pool (which hosts studies that involve playing games and other enjoyable experiences).

After evaluating each experience, participants also provided more details (simply for our descriptive knowledge). They reported the *kind of person* involved (forced choice: "parent/family"; "teacher/instructor"; "manager/work colleague"; "policy/law-maker"; "peer/friend"; "other"); the *threshold's valence* (forced choice: "positive [e.g., Do X for a reward/good thing to happen]"; "negative [e.g., Do X for a punishment/bad thing to happen"]); how *long ago* it happened (forced choice: "recently"; "long time ago"; "somewhere in between"); and how *impactful* the outcome was (forced choice: "small impact [didn't really matter]"; "medium impact [mattered somewhat]"; "big impact [mattered a great deal])." In addition, for each of the two "violate" targets, participants reported the *violation's direction* (forced choice: "person was 'quicker' to act than expected [e.g., they said 3, and acted at 2]"; "person was 'slower' to act than expected [e.g., they said 3, and acted at 4])"; and the *degree* of this behavior (forced choice: "much quicker/slower to act than expected"; "a little quicker/slower to act than expected"; "somewhere in between").

Finally, all participants completed an attention check regarding what the study was about (forced choice from 1 of 3 options: "Rating people who set cut-offs for judging others"; "Rating people who ranked their favorite things"; "Rating people who competed in sports"), and the same honesty check from prior experiments.

Results and Discussion

Screening/Attrition. To yield our preregistered sample size of 200 successfully screened participants—participants who could recall all four kinds of threshold (in)consistencies—we needed to recruit 494 in total (i.e., we retained 41.70% of the sample, 206 of 494). This finding is itself noteworthy because it highlights that the threshold (in)consistencies we document in the current research are common in everyday life. Of the 288 participants who did not pass this screener, the majority (85.42%, 246 of 288) could bring to mind at least one of the four (in)consistencies (see OSF). There were no demographic differences between participants who screened-in versus screened-out (*ps* ≥ .125; see OSF).

Main Results: Violating Thresholds Makes Others Feel Worse. First, we analyzed effects on how participants felt. We conducted a repeated-measures Linear Regression via SPSS GEE entering participant as a subject variable; Others' Threshold Behavior (Adhere vs. Violate) and One's Own Role (Recipient vs. Observer) as within-subject factors; and own experienced feelings (continuous, −5 to 5; lower = worse) as the dependent variable.

As hypothesized, there was the critical main effect of Others' Threshold Behavior, Wald = 152.79, *df* = 1, *p* < .001, which held across One's Own Role (null interaction: Wald = 0.02, *df* = 1, *p* = .879; main effect of One's Own Role, Wald = 2.11, *df* = 1, *p* = .147). When participants were personally involved with a judge's threshold behavior, they felt worse upon experiencing threshold violations (*M* = −1.56, *SD* = 2.74) versus threshold adherences (*M* = 1.05, *SD* = 2.93), paired *t*(205) = 9.16, *p* < .001, *d* = 0.64; likewise, even when participants merely witnessed a judge's threshold behavior involving someone else, they *also* felt worse upon experiencing threshold violations (*M* = −1.39, *SD* = 2.36) versus threshold adherences (*M* = 1.27, *SD* = 2.53), paired *t*(205) = 10.76, *p* < .001, *d* = 0.75.

Main Results: Violating Thresholds Makes Threshold Setters Look Worse. Second, we analyzed effects on how participants viewed the threshold setter, via the same analysis.

The same hypothesized effects emerged: There was the critical main effect of Others' Threshold Behavior, Wald = 183.93, *df* = 1, *p* < .001, which again held across One's Own Role (null interaction: Wald = 0.14, *df* = 1, *p* = .708; main effect of One's Own Role, Wald < .001, *df* = 1, *p* = .984). When participants were personally involved with a judge's threshold behavior, they came away judging the threshold setter more negatively upon experiencing threshold violations (*M* = −1.58, *SD* = 2.71) versus threshold adherences (*M* = 1.35, *SD* = 2.77), paired *t*(205) = 10.58, *p* < .001, *d* = 0.74; likewise, even when participants merely witnessed a judge's threshold behavior involving someone else, they *also* judged them more negatively upon experiencing threshold violations (*M* = −1.63, *SD* = 2.38) versus threshold adherences (*M* = 1.41, *SD* = 2.60), paired *t*(205) = 11.88, *p* < .001, *d* = 0.83.

Main Results: Violating Thresholds Reduces Reward Invites. Third, we analyzed effects on participants' inviting others for our rewarding study, via the same analysis but using logistic regression (yes-invited vs. no-invited).

Again, there was the critical main effect of Others' Threshold Behavior, Wald = 58.10, *df* = 1, *p* < .001, which held across One's Own Role (null interaction: Wald = 0.67, *df* = 1, *p* = .412; main effect of One's Own Role, Wald = 0.003, *df* = 1, *p* = .957). When participants were personally involved with a judge's threshold behavior, they were less likely to actually invite this judge for our rewarding study upon experiencing threshold violations (31.07%, 64 of 206) versus threshold adherences (54.85%, 113 of 206), paired Wald = 43.11, *df* = 1, *p* < .001; likewise, even when participants merely witnessed a judge's threshold behavior involving someone else, they *too* were less likely to actually invite this judge for our rewarding study upon experiencing threshold violations (29.61%, 61 of 206) versus threshold adherences (56.31%, 116 of 206), paired Wald = 45.23, *df* = 1, *p* < .001.

Further Insights. Readers may be interested in exploring the additional variables we measured, such as the valence of the threshold (reward vs. punishment) and the direction of the violation (quicker to judge vs. slower to judge). There are many potential questions to test in our data file (see OSF), and we encourage researchers to test them. For example, readers may wonder about self-interested motivations to pursue pleasure and avoid pain (e.g., Thorndike, 1911); is it *really* the case, for instance, that people prefer being punished as promised as opposed to enjoying a reprieve?

We investigated this idea (via exploratory, nonpreregistered analyses) and found mixed support for it. On our "feelings" measure, for example, we found that participants felt better after observing a threshold violation versus a threshold consistency—when that violation entailed the person being *slower to punish* (two-way interaction between Others' Threshold Behavior and Violation's Direction, for negative thresholds: Wald = 37.92, *df* = 1, *p* < .001), and especially when the person was slower to punish *participants themselves* (three-way interaction with One's Own Role: Wald = 18.78, *df* = 1, *p* < .001; self-feelings after delayed punishment [$M = -0.76$, $SD = 2.62$] vs. on-time punishment [$M = -1.18$, $SD = 2.54$] vs. hastened punishment [$M = -2.72$, $SD = 2.11$]). However, we did not observe these patterns on our other measures, nor did we find the converse effect for positive thresholds: While participants felt worse when the person delayed their reward ($M = -1.84$, $SD = 2.67$) versus giving it as promised ($M = 2.41$, $SD = 2.23$), they also felt worse when the person *hastened* their reward ($M = 0.13$, $SD = 3.34$) versus giving it as promised (two-way interaction between Others' Threshold Behavior and Violation's Direction, for positive thresholds: Wald = 0.19, $df = 1$, $p = .666$; three-way interaction with One's Own Role: Wald = 1.25, $df = 1$, $p = .263$). We will return to these ideas in the General Discussion. Experiment 6b will manipulate some of these factors in specific contexts.

Other Variables. Participants brought to mind varied experiences (see OSF for full details). The relative majority entailed "parents/family" (involved in 32.65% of all experiences brought to mind, 269 of 824); entailed "negative/punishment-related" thresholds (involved in 56.80%, 356 of 824); happened "somewhere in between" recently and long ago (involved in 40.05%, 330 of 824); and had "medium" impact (involved in 46.36%, 382 of 824). For the two Violate targets, most violations entailed being "slower" to act (involved in 56.10%, 231 of 412), to a "much" slower degree (involved in 34.95%, 144 of 412). Overall, results hold when rerunning our analyses with these variables entered as covariates (Main effect of Others' Threshold Behavior on feelings, Wald = 11.38, $df = 1$, $p < .001$; on views, Wald = 11.40, $df = 1$, $p < .001$; on invites, Wald = 7.42, $df = 1$, $p = .006$).

Most participants passed the attention check (95.15%, 196 of 206) and the honesty check (89.81%, 185 of 206). When rerunning our analyses while excluding participants who failed either of these checks (leaving $N = 177$), results are unchanged (Main effect of Others' Threshold Behavior on feelings, Wald = 148.65, $df = 1$, $p < .001$; on views, Wald = 170.94, $df = 1$, $p < .001$; on invites, Wald = 51.94, $df = 1$, $p < .001$).

Experiment 6b: Consequences of Threshold Violations (Generalized Effects)

Experiment 6b tested more controlled settings that manipulated some of the free-varying features from Experiment 6a, thus assessing more general insights. We again hypothesized that threshold violations can sometimes elicit harmful reputational effects, at least in these common cases when the other party lacks the unpacked knowledge to contextualize the violation.

Method

Participants. We requested and yielded 100 "Cloud Approved" participants from Cloud Research (34.00% women; 25.00% non-White; $M_{\text{age}} = 37.56$, $SD_{\text{age}} = 10.05$) who completed the experiment for $5.00.

Procedure. The experiment was fully within subjects, allowing us to manipulate many parameters: It followed a 5 (target: parent, teacher, manager, policymaker, layperson) × 3 (target's threshold behavior: target was quicker to judge vs. slower to judge vs. accurate) × 2 (judge how: point when target viewed others differently vs. treated others differently) × 2 (valence of judged behavior: good behavior vs. bad behavior) design, yielding 60 judged events. As in Experiment 6a, here in the main text we refer to "violations" (and the like), but we used no such words in our actual study so as to avoid negative connotation.

All participants learned they would judge 60 unique events, one at a time in randomized order (divided across 10 pages [6 targets per page] to make it easier for participants to track progress). The description of each target followed the same format, except we rotated through our manipulations of interest. We varied (a) who committed the threshold-action (Target); (b) whether this action involved the target violating or adhering to their threshold (Threshold Behavior); (c) what the target did upon judgment (Judge How); and (d) whether the target was judging others' good or bad behavior (Valence). We included these various manipulations simply for generalizability; across them, we hypothesized that these violations may elicit worse evaluations of the threshold setter.

To take one concrete example, all participants evaluated the following target—which was a *manager* (Target); who was *quicker* to judge their employee (Threshold Behavior); in terms of *punishing* them (Judge How); for the employee's repeated *bad* behavior (Valence):

> Imagine the following. First: A manager establishes a clear cut-off for an employee's bad behavior ("3 bad acts and I'll punish you"). Then: As it turns out, the manager acts sooner than their established cut-off (They punish the employee after "2 bad acts").

For our dependent variable, participants then rated this target via a five-item Reputation block (items presented in randomized order), with each item rated from 1 (*disagree*) to 5 (*agree*; with 3 = *it depends*). The five items (phrased for this target, for example) were, (a) *This manager is a hypocrite*; (b) *This manager has hurt the employee's ability to accurately learn from experience*; (c) *This manager acted unfairly*; (d) *This manager held the employee to different standards than they hold other employees*; and (e) *This manager shouldn't be trusted for setting future rules like this*. Across these parameters, we hypothesized that Threshold Behavior—namely, threshold inconsistencies (quicker *or* slower to judge) versus threshold consistencies—may elicit more negative reputational assessments of the threshold setter.

All 60 events (and the phrasing of their dependent measures) followed this same format, except we manipulated each parameter where it appeared. For Target, we rotated through five kinds. Parents interacted with their children; teachers interacted with their students; managers interacted with their employees; policymakers interacted with citizens; and laypeople interacted with peers. For Threshold Behavior and Valence, all targets always set a threshold of "3 good acts" or "3 bad acts," and acted either sooner by one such act ("2 good acts"; "3 bad acts"); acted later by one such act ("4 good acts"; "4 bad acts"); or acting right at that threshold ("3 good acts"; "3 bad acts"; we made these inconsistencies always off by ±1 from 3 to allow for a more conservative test—the degree of these "violations" is small, and each target commits them just once). For our final factor (Judge How), targets were described as either *treating* the recipient differently at this point (rewarding or punishing them; e.g., the manager "punishes the employee" after only 2 bad acts), or *perceiving* the recipient differently at this point (e.g., the manager "views the employee negatively" after only 2 bad acts).

Finally, all participants reported demographic information and completed the same attention check from Experiment 6a plus the same honesty check from prior experiments.

Results and Discussion

Main Results: Judgments of Target's Reputation. We reverse-coded the five dependent measures such that lower scores reflect more negative reputational assessments, and collapsed them into a composite Reputation Scale (across each of the 60 events: α's ≥ .818). We then conducted a repeated-measures Linear Regression via SPSS GEE entering participant as a subject variable; Target (Parent, Teacher, Manager, Policymaker, Layperson), Threshold Behavior (Quicker to Judge vs. Slower to Judge vs. Accurate), Judge How (Perception vs. Treatment), and Valence of Judged Behavior (Good Behavior vs. Bad Behavior) as within-subject factors; and Reputation Scale (continuous, 1–5; lower = worse) as the dependent variable.

As hypothesized, there was the critical main effect of Threshold Behavior, Wald = 341.68, *df* = 2, $p < .001$—meaning participants differentially evaluated threshold setters depending on whether they violated versus adhered to their thresholds. This effect held across all other factors aside from Valence, as indicated by a two-way interaction between Threshold Behavior and Valence, Wald = 107.58, *df* = 2, $p < .001$ (no other interactions were significant, *p*s ≥ .321; see OSF for remaining output, which is incidental). Figure 7 plots the results across each Valence (collapsing across all other factors).

For *positive* thresholds (e.g., judges who rewarded others for good behavior): Pairwise comparisons reveal that being slow to positively judge others elicited worse reputations ($M_{\text{all}} = 2.39$, $SD_{\text{all}} = 0.77$) than being quick to positively judge others ($M_{\text{all}} = 3.27$, $SD_{\text{all}} = 1.03$), paired $t(99) = 9.50$, $p < .001$, $d = 0.95$. Critically, however, both kinds of violations (at least on average, in these contexts) elicited worse reputations than remaining consistent with one's threshold ($M_{\text{all}} = 4.61$, $SD_{\text{all}} = 0.81$): accurate versus slow, paired $t(99) = 19.76$, $p < .001$, $d = 1.98$; accurate versus quick, paired $t(99) = 11.85$, $p < .001$, $d = 1.19$.

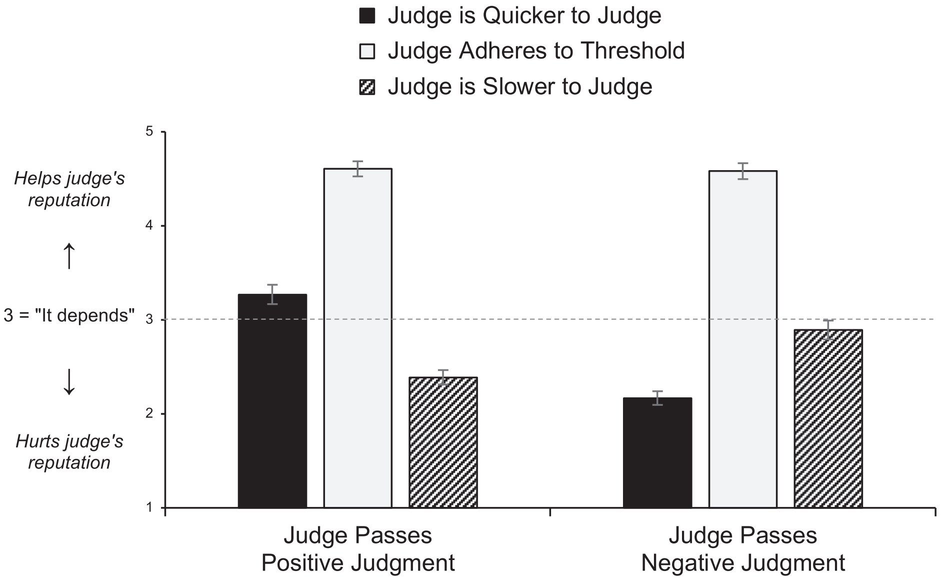
Likewise, for *negative* thresholds (e.g., judges who punished others for bad behavior): Pairwise comparisons first reveal the opposite effect, such that being *quick* to negatively judge others elicited worse reputations ($M_{\text{all}} = 2.17$, $SD_{\text{all}} = 0.73$) than being *slow* to negatively judge others ($M_{\text{all}} = 2.89$, $SD_{\text{all}} = 1.01$), paired $t(99) = 7.73$, $p < .001$, $d = 0.77$. But again, both kinds of violations (at least on average, in these contexts) elicited worse reputations than remaining consistent ($M_{\text{all}} = 4.58$, $SD_{\text{all}} = 0.85$): accurate versus quick, paired $t(99) = 20.36$, $p < .001$, $d = 2.04$; accurate versus slow, paired $t(99) = 14.03$, $p < .001$, $d = 1.20$.

Comparisons to "It Depends". As can be seen in Figure 7, the horizontal dotted line references how each of these effects compares to the scale midpoint—"it depends."

A few insights emerge from one-sample *t* tests comparing each bar to this point. First, adhering to one's threshold clearly helped in these specific contexts (the two "adhere" bars are the best overall, and both are significantly higher than "it depends": *t*s ≥ 18.58, *p*s ≤ .001, *d*s ≥ 1.86). Second, being slower to positively judge, or quicker to negatively judge, clearly *hurt* in these specific contexts ("slower positive" and "quicker negative" are both significantly lower than "it depends": *t*s ≥ 7.92, *p*s ≤ .001, *d*s ≥ 0.79). Third, we find less clear evidence for the converse violations: Participants

Figure 7

Experiment 6b: Mean Evaluations of the Judge's Reputation as a Function of Valence and the Judge's Threshold Behavior



Note. This figure collapses across all targets (parents judging children; teachers judging students; managers judging employees; policymakers judging citizens; laypeople judging peers) and judgment types (the judge *rewards/punishes* vs. *views positively/negatively*). This two-way interaction in the figure—between Threshold Behavior and Valence—was not qualified by further interactions with any other factor (all *p*s ≥ .321), nor did any of these other factors have their own two-way interactions with Threshold Behavior (all *p*s ≥ .534). That is: This same pattern depicted in Figure 7 emerges regardless of which targets and/or judgment types are inputted (see Supplemental Figures S3–S6). Error bars ±1 standard error.

seemed more likely to acknowledge that "it depends" for these outcomes than for all other outcomes, "slower negative" did not significantly differ from "it depends," $t(99) = 1.07$, $p = .286$, $d = 0.11$; and, although "quicker positive" was significantly higher than "it depends," $t(99) = 2.61$, $p = .010$, $d = 0.26$, this stayed closer to the midpoint than did all other effects.

Effects of the Other Factors. As noted, these patterns are identical regardless of *who* the judge was (e.g., parents vs. managers, and so on) and *what* their judgment threshold was (perception vs. treatment); if readers are wondering how the effect varies by these factors, it resembles Figure 7 (see Supplemental Figures S3–S6 for each of these other figures).

Other Variables. Most participants passed the attention check (96.00%, 96 of 100) and the honesty check (100.00%, 100 of 100). When rerunning our analyses while excluding participants who failed either of these checks (leaving $N = 96$), results are unchanged (Main effect of Threshold Behavior on evaluations: Wald = 407.12, $df = 2$, $p < .001$; interaction between Threshold Behavior and Valence on evaluations, Wald = 116.85, $df = 2$, $p < .001$).

Experiment 6b further highlights consequences. The same targets were made worse off by violating (vs. adhering to) their preset thresholds, at least in these contexts.

General Discussion

Did a friend speak out of turn when questioning one's character? Did an employee earn special praise for how they treated their coworkers? Such everyday evaluations highlight the role of people's underlying thresholds for passing social judgment. The results of 10 experiments (see Table 2 for a data summary) reveal that people often violate their preset thresholds, even after formally establishing them based on having full information about what might unfold. Moreover, we propose a framework to understand these threshold violations, which we test across many contexts, behaviors, and designs (including within-subjects designs, with people violating their own preset thresholds). All told, our central finding is that people are swayed to be "quicker" *and* "slower" to judge than they declare beforehand—both as a function of unpacked psychological support.

Theoretical Contributions

Our findings advance Klein and O'Brien (2018) initial claim that people are quicker to judge than they think. First, we examine many social judgment domains using experimental designs that match the measurement of the dependent variable across conditions, largely involving ecologically valid thresholds for real behavior (e.g., issuing/withholding rewards and punishments). Second, Klein and O'Brien (2018) neither tested nor discussed *why* people are "quicker to judge than they think," aside from broadly speculating that real-time judgment may be swayed by a "System 1 suite of affective responses designed to provide rapid online feedback about the current environment" (p. 13222) that is too subtle to appreciate in prospect. In contrast, we put forth a novel theoretical account that integrates the psychology underlying the time course and formation of these everyday social judgments. Using the logic of support theory (Tversky & Koehler, 1994) to model these dynamics, we reveal that Klein and O'Brien's (2018) account is incomplete, and instead is one part of a broader framework: People can be swayed to be "quicker" *and* "slower" to judge both as a function of *support* (which need not be constrained to "System 1" inputs).

Our findings raise fruitful insights into the psychological effects of thresholds writ large. Findings in other areas of psychology emphasize that reference points and comparison standards exert substantial influence over people's judgments (e.g., Kahneman & Tversky, 1979). The kinds of thresholds assessed in the current research—which essentially reflect how people think about standards for acceptable and unacceptable behavior—may represent a similarly important focus of research, especially since people apparently do not always adhere to them. Given that meaningful social judgments often comprise *repeated* interactions (e.g., involving forming conclusions based on recurring observations of a colleague or acquaintance, as opposed to one-shot judgments of an isolated stranger), our findings further highlight that moving beyond one-shot study designs to more dynamic paradigms might reveal unforeseen changes in preferences and behaviors as things unfold.

Our findings also contribute to research on reducing stereotyping and discrimination. A major theme from this research is that reducing such biases requires reducing ambiguity—not just through gaining more knowledge about the target of judgment but also through more clearly establishing one's judgment criteria beforehand (e.g., "Thus, discrimination typically occurs when socially appropriate behavior is not clearly defined …": Hodson et al., 2002, p. 461; "Having committed to unambiguous criteria, [evaluators] will be unable to define merit to the benefit of specific job candidates … Our research thus demonstrates the efficacy of a method to reduce job discrimination: the establishment of standards of merit prior to the review of candidates"; Uhlmann & Cohen, 2005, pp. 478–479). Our findings warn that, even after establishing a concrete system that predetermines judgment thresholds, people may nonetheless violate them as each good or bad behavior unfolds piece by piece. Evaluation systems that appear fairly improved on the surface, but still yield selectively quick or slow enforcement, will require a closer look.

Beyond social judgment, our findings also contribute to research on motivation, which has similarly drawn on the logic of support theory to understand people's failures to attain goals. People often set goals to finish tasks by a certain date, but then run late; unpacking individual hurdles can lead people to predict longer (and thus potentially more accurate) completion times (e.g., Connolly & Dean, 1997; Forsyth & Burt, 2008; Kruger & Evans, 2004). People often set goals to consume a certain number of calories, but then overeat; unpacking individual meals can lead people to predict bigger (and thus potentially more accurate) intake (e.g., Jia et al., 2020, Study 4). People often set goals to save a certain amount of money, but then overspend; unpacking individual purchases can lead people to predict bigger (and thus potentially more accurate) expenses (e.g., Howard et al., 2022; Peetz et al., 2015); and so on. Such examples can be more broadly understood through the lens of our framework—all essentially highlighting cases in which people violate self-set thresholds. Our "quicker to judge" versus "slower to judge" terminology may also help integrate competing effects in this research. For example, unpacking tasks indeed mitigate the planning fallacy when they unpack lengthy delays (essentially, "quicker to judge")—but *worsen* the planning fallacy when they unpack *atypically short* delays (essentially, "slower to judge": Hadjichristidis et al., 2014). Together, these findings suggest our framework should apply

Table 2
Summary of Data

Experiment	Goal of study	Sample	Setting	Study design	Key finding	Summary
Experiment 1	Test basic unpacking effect	N = 804 P's (45% women; 27% non-White; M = 40 years old)	Hypothetical behavior	2 (judgment type, between subjects: packed vs. unpacked) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior)	Effect of judgment type, $\beta = -0.31$, $SE = .07$, $p < .001$	Higher support = quicker to judge
Experiment 2a	Test Moderator 1: Manipulate unpacked *emotion* (positive)	N = 499 P's (41% women; 28% non-White; M = 40 years old)	Real behavior	2 (judgment type, between subjects: packed vs. unpacked)	Effect of judgment type, $\beta = 2.45$, $SE = .24$, $p < .001$	Higher support = quicker to judge
Experiment 2b	Test Moderator 1: Manipulate unpacked *emotion* (negative)	N = 491 P's (44% women; 28% non-White; M = 40 years old)	Real behavior	2 (judgment type, between subjects: packed vs. unpacked)	Effect of judgment type, $\beta = 0.73$, $SE = .21$, $p < .001$	Higher support = quicker to judge
Experiment 3a	Test Moderator 1: Manipulate unpacked *emotion* (positive)	N = 599 P's (45% women; 27% non-White; M = 40 years old)	Real behavior	3 (judgment type, between subjects: packed vs. unpacked vs. unpacked-blunted)	Effect of judgment type, $F(2, 596) = 21.25$, $p < .001$, $\eta_p^2 = .07$	Higher support = quicker to judge; lower support = slower to judge
Experiment 3b	Test Moderator 1: Manipulate unpacked *emotion* (negative)	N = 285 P's (61% women; 30% non-White; M = 42 years old)	Real behavior	3 (judgment type, between subjects: packed vs. unpacked vs. unpacked-weak)	Effect of judgment type, $F(2, 282) = 11.56$, $p < .001$, $\eta_p^2 = .08$	Higher support = quicker to judge; lower support = slower to judge
Experiment 4	Test Moderator 2: Manipulate unpacked *clearness of goodness/badness*	N = 601 P's (48% women; 28% non-White; M = 41 years old)	Real behavior	2 (judgment type, within subjects: packed vs. unpacked) × (what's unpacked, between subjects: high support vs. low support) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior)	Two-way interaction between judgment type and what's unpacked, wald = 81.51, $df = 1$, $p < .001$	Higher support = quicker to judge; lower support = Slower to judge
Experiment 5a	Test Moderator 3: Manipulate unpacked *consistency*	N = 606 P's (68% women; 22% non-White; M = 38 years old)	Real behavior	2 (what's unpacked, between subjects: high support vs. low support) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior)	Effect of what's unpacked, $\beta = 1.62$, $SE = .61$, $p = .008$	Higher support = quicker to judge
Experiment 5b	Test Moderator 3: Manipulate unpacked *consistency*	N = 619 P's (72% women; 28% non-White; M = 36 years old)	Real behavior	2 (what's unpacked, between subjects: high support vs. low support) × 2 (valence of judged behavior, between subjects: good behavior vs. bad behavior)	Effect of what's unpacked, $\beta = 1.82$, $SE = .61$, $p = .003$	Lower support = slower to judge
Experiment 6a	Test downstream consequences (real-world stimuli)	N = 206 P's (64% women; 57% non-White; M = 32 years old)	Real behavior	2 (others' threshold behavior, within subjects: adhere vs. violate) × 2 (one's own role, within subjects: recipient vs. observer)	Effect of others' threshold behavior, walds ≥ 58.10, dfs = 1, $ps \leq .001$	Threshold violations undermined violators' reputations

(*table continues*)

Table 2 *(continued)*

Experiment	Goal of study	Sample	Setting	Study design	Key finding	Summary
Experiment 6b	Test downstream consequences (controlled stimuli)	N = 100 P's (34% women; 25% non-White; M = 38 years old)	Hypothetical behavior	Fully within subjects: 5 (target: parent, teacher, manager, policymaker, layperson) × 3 (target's threshold behavior: target was quicker to judge vs. slower to judge vs. accurate) × 2 (judge how: point when target viewed others differently vs. treated others differently) × 2 (valence of judged behavior: good behavior vs. bad behavior)	Effect of target's threshold behavior, wald = 341.68, df = 2, $p < .001$	Threshold violations undermined violators' reputations

Note. SE = standard error.

beyond social judgment, making the same predictions across many judgment domains.

Practical Implications

That people violate social judgment thresholds bears on other real-world issues, as such thresholds often *are* set in advance in everyday life. In contrast to policymakers, the citizens affected by predetermined policies could feel differently once the relevant behaviors start to occur; difficulties in maintaining public support may partly reflect the discrepant mode of judgment between how people make policy (e.g., presetting a threshold that can be enforced across many possible outcomes) versus how people experience policy (e.g., reacting to the specific outcomes that happen to unfold, in real time). Such discrepancies may be at the root of many kinds of conflict and miscommunication, with different parties calling for reward or punishment at different times and contexts. One study found that people are especially likely to derogate others as hypocrites when they first proclaim the value of good behavior and then behave badly (vs. the other way around: Barden et al., 2005)—suggesting that the kinds of violations observed in our research (which conceptually resemble this "proclaim first, behave second" order of operations) may be especially likely to garner social stigma.

It is also interesting to consider the potential connection between our findings (i.e., a ready willingness among evaluators to violate their agreed-upon thresholds) and understanding the challenges people might face in navigating evolving social standards for appropriate behavior. In today's information age, people's present actions are widely documented (e.g., on social media) for future eyes to see and judge. If what counts as good or bad shifts over time (as it often does: Ronson, 2016), people could routinely run into unforeseen reputational challenges for having committed identical behavior. For example, as put in our terms, society at Time 1 might "preset" what counts as socially appropriate; yet society at Time 2 may have changed to lower this threshold for offense and thus chastise a past-appropriate-actor anyway—essentially reflecting our "quicker to judge" effect.

Future Directions

Other Inputs?

Another benefit of our model is that one can use it to make predictions about any variable so long as one knows how it bears on support: Psychological inputs that increase (vs. decrease) support should "hasten" (vs. "slow") judgment thresholds, even in cases when people have full information of those possibilities when establishing their thresholds beforehand.

One such input is diagnosticity (e.g., Skowronski & Carlston, 1987, 1989). In the ability domain, for instance, positive (vs. negative) behaviors are typically more diagnostic for discriminating between alternative trait categorizations and thus elicit positivity biases in social judgment (e.g., upon observing someone solve a complicated math problem, judges tend to infer the person must be a math whiz—because a nonwhiz cannot solve it; in contrast, observing someone struggle for a solution is weighted less heavily in ability judgments, because math whizzes can also struggle); but in the morality domain, negative (vs. positive) behaviors are typically the more diagnostic ones and thus elicit negativity biases (e.g., upon observing someone commit fraud, judges tend to infer the person must be a liar—because an honest person cannot be fraudulent; in contrast, observing someone tell the truth is weighted less heavily in morality judgments, because liars can also be honest).

This logic could be fruitfully understood through the lens of our framework to the extent that diagnosticity feeds into support: "Quicker to judge" effects should emerge for diagnostic unpacking whereas "slower to judge" effects should emerge for nondiagnostic unpacking. Going further, one might predict to find a negativity bias in our research (such that people are generally "quicker" to pass judgment when unpacked behaviors are negative vs. positive) given that our tested domains fall closer to the morality versus ability side of the equation. This is indeed what we find in our main effects of Valence in Experiments 1–4–6b. This also explains the valence asymmetry we have found in our previous research on threshold judgments (e.g., O'Brien & Klein, 2017). Future research on this front should ensure that unpacked negative versus positive behaviors are otherwise precisely matched on value (which we did not design our studies to do), and assess other signals of diagnosticity as they are learned in everyday life (e.g., individual pieces of negative vs. positive information tend to be more distinct from each other—which might also predict a negativity bias in the context of our research: Alves et al., 2017).

As for other potential inputs, outcomes that seem certain, unchangeable, or otherwise "real" should hasten judgment while those that seem uncertain, changeable, or otherwise "hypothetical" should slow it (Markman & Beike, 2012; Miller & Kahneman, 1986); outcomes that confirm one's prior expectations should hasten judgment while those that disconfirm them should slow it (Ditto & Lopez, 1992; Jonas et al., 2001; Nickerson, 1998); outcomes that bolster self-views should hasten judgment while those that threaten self-views should slow it (e.g., people may extend the thresholds at which they view *themselves* vs. others as bad actors: Chambers & Windschitl, 2004; Klein & Epley, 2016; Klein & O'Brien, 2017; O'Brien & Kardas, 2016); outcomes that seem more versus less extreme (Fiske, 1980), more versus less intentional (Ames & Fiske, 2013), and more versus less concrete, vivid, and immediate (Trope & Liberman, 2010) should hasten versus slow judgment; and outcomes that trigger other kinds of motivational or regulatory processes should operate accordingly (e.g., friends may think one nasty fight will render them foes, but in reality, friends work to stay friends: Cameron & Payne, 2011; Kawakami et al., 2009; Wilson & Gilbert, 2005). Incidental factors that also influence these inputs might wield similar effects (e.g., perceptions of the "clearness of goodness/badness" of an event might be distorted by mere processing fluency: Reber et al., 2004). Different presentations of information should also bear on support in this way (e.g., someone who falls short of a reward threshold but *ends strongly* might be praised more than someone who meets it but *ends weakly*: Loewenstein & Prelec, 1993; O'Brien & Ellsworth, 2012).

The broader point here is that our tested inputs are not exhaustive (see O'Brien, 2023 for a conceptual review of such inputs as they bear on threshold judgments); future research can and should explore many other possibilities, and also taxonomize them (e.g., by prevalence, effect size, and so on).

Calibrating Thresholds?

Another question for future research is when people set thresholds accurately. Perhaps prompting threshold setters to simulate possible outcomes more thoroughly (via, e.g., increasing their time to think or using vivid-imagery prompts similar to "de-focusing" techniques, which might help bring to mind fuller distributional information: Wilson et al., 2000) can help them focus on a wider range of eligible examples, and therefore help them set thresholds that will stick more broadly. Similarly, to the extent that personal experience with unpacked states exposes one to a diverse variety of them, then more versus less experienced judges may tend to set more accurate thresholds; acquiring personal experience has been cited as one strategy for closing "empathy gaps" (Van Boven et al., 2013). Discrepancies may be reduced when those who set thresholds (e.g., policymakers) once walked in the shoes of affected parties (e.g., constituents).

At the same time, support theory research warns that it is impractical to expect threshold setters to become calibrated on their own ("People can be encouraged to unpack a category into its components, but they cannot be expected to think of all relevant conjunctive unpackings or generate all relevant future scenarios": Tversky & Koehler, 1994, p. 565). An easier strategy may be to recruit intermediaries to enforce threshold adherence no matter how much one *wants* to act or delay as things unfold ("blind justice": Baron, 1995). Threshold setters may also benefit from setting *weighted* thresholds if they can do so, akin to the "points" system embedded in many driving laws (whereby drivers get punished after hitting a preset violation threshold, but larger infractions like drunk driving count more toward this threshold than do smaller infractions like running a red light).

Are Threshold Setters (Un)aware?

This support-based framework assumes presetters underappreciate specific fitting cases (e.g., those at distributional tails) that—if unpacked and accounted for—may lead them to set thresholds that better consider what all could fit and unfold under their threshold. An alternative explanation is that presetters fully appreciate such cases but choose to dismiss them. For example, perhaps their goal is to set a threshold around common cases but then adjust for unusual circumstances if those occur. Indeed, observing people violate their thresholds need not mean they planned to adhere to them. This possibility strikes us as unlikely to explain all our findings—and, even so, it may still invite problems (e.g., for the threshold setter's reputation) if such allowances are not spelled out beforehand (e.g., as in Experiments 6a–6b). In any case, this is an important question for future research to further tease apart. One way to do so would be to measure anticipated flexibility (e.g., by asking Packed participants to indicate how flexible their threshold is, separately from asking them to indicate the threshold itself). Another way would be to test whether people *re*calibrate their preset thresholds after an unpacking manipulation (echoing our proposed "personal experience" strategy for calibration); our support-based framework predicts people will recalibrate (e.g., "Oh, right—my threshold should account for this") while this alternative explanation predicts people will not recalibrate (e.g., "I already know this—I just don't think it should affect my preset threshold").

When Are Threshold Violations Adaptive?

Finally, future research should further tease apart the downstream consequences of threshold violations—both bad *and good*. Although

Experiments 6a–6b highlighted some problems for how threshold violators are judged, these emerged under particular conditions.

First, Experiments 6a–6b did not ask or inform participants about specific reasons for why others violated their thresholds. Observers may adjust their judgments depending on their own unpacked knowledge, consistent with our framework. Note, however, that Experiment 6b did provide participants the option of indicating "it depends" yet they still passed judgment on the violator—suggesting observers may not fully appreciate situational exceptions unless such exceptions are made explicit, which they often are not (Jones & Nisbett, 1971). Second, Experiment 6a found no effects of participants being on the receiving end of threshold violations themselves versus observing threshold violations happen to others, but higher stakes contexts may reveal different patterns. Desires to punish rulebreakers (Tyler & Boeckmann, 1997) are presumably less strong when others break rules to benefit *oneself*—perhaps explaining why we did not find positive effects of threshold violations in Experiment 6b (which assessed outside observers only). In general, one can imagine that self-serving motivations should lead people to be especially appreciative of others who are slower to punish them and quicker to reward them (e.g., Thorndike, 1911).

However, as we had noted in our Experiment 6a discussion, we found mixed results for such effects; if anything, the results suggested people might appreciate slow self-punishment but not necessarily quick self-reward. Such an asymmetry is consistent with theories of achievement motivation, which propose that people do not pursue pleasurable outcomes mindlessly; they also derive utility from earning them (e.g., Higgins, 1997). In any case, more research is needed to better establish the bad *and* good of threshold violations—including cases when violations help and adherences hurt (e.g., perhaps a teacher's reputation takes a *hit* for dutifully sticking to one's deadline for a student who experiences a unique tragedy that prevents them from fairly earning it, as opposed to granting the student an extension).

These ideas emphasize a broader point about the benefits versus costs of threshold violations. On the one hand, one may view threshold violators as simply rationally adapting to new information—following the Bayesian axiom "When the facts change, I change my mind." On the other hand, we believe there are at least two critical shortcomings to such a view. First, even in cases when threshold violations reflect rational updating, violators may still encounter other costs (e.g., people often derogate others for updating social judgments based on Bayesian reasoning alone: Cao et al., 2019). Second, in principle (and as we have emphasized throughout in our experiments), threshold presetters have all the same information that threshold-followers do. That is, the threshold violations in our research do not occur because of truly new information; rather, they occur because presetters fail account for certain aspects of the information they already have. A more fitting summary of our findings may therefore be that people follow the axiom "When I'm reminded of a known fact, I change my mind." This is not the same as learning something truly new. Threshold violations that occur in reaction to ostensibly agreed-upon (vs. truly new) information may seem more unfair than fair.

Concluding Thoughts

Everyday social life often entails evaluating others against some threshold for their behavior. Managers tell employees that, if they land a certain number of clients, they will get a bonus; parents tell children that, if they miss a certain number of chores, they will be grounded; and so on. In less formal settings, too, people constantly hold others to certain expectations of good and bad behavior, guiding their reputations and relationships.

The current research asks: When do people pass such judgments? Our findings suggest the answer depends on the point at which people are asked, raising novel insights into theory, policy, and everyday interactions. People can be swayed to be both "quicker" *and* "slower" to pass social judgment depending on what unfolds, even when they knew those possibilities beforehand. When it comes to treating others, making exceptions to the rule may often be the rule—for better or worse.

References

Alves, H., Koch, A., & Unkelbach, C. (2017). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences*, *21*(2), 69–79. https://doi.org/10.1016/j.tics.2016.12.006

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science*, *24*(9), 1755–1762. https://doi.org/10.1177/0956797613480507

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400. https://doi.org/10.1037/h0022280

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369–1378. https://doi.org/10.3758/BF03206327

Barden, J., Rucker, D. D., & Petty, R. E. (2005). "Saying one thing and doing another": Examining the impact of event order on hypocrisy judgments of others. *Personality and Social Psychology Bulletin*, *31*(11), 1463–1474. https://doi.org/10.1177/0146167205276430

Baron, J. (1995). Blind justice: Fairness to groups and the do-no-harm principle. *Journal of Behavioral Decision Making*, *8*(2), 71–83. https://doi.org/10.1002/bdm.3960080202

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525–537. https://doi.org/10.1037/0022-3514.77.3.525

Birnbaum, M. H. (1972). Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, *93*(1), 35–42. https://doi.org/10.1037/h0032589

Cameron, C. D., & Payne, B. K. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *Journal of Personality and Social Psychology*, *100*(1), 1–15. https://doi.org/10.1037/a0021643

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same Bayesian judgment they criticize in others. *Psychological Science*, *30*(1), 20–31. https://doi.org/10.1177/0956797618805750

Carlson, K. A., & Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes*, *104*(1), 113–121. https://doi.org/10.1016/j.obhdp.2007.03.004

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin*, *130*(5), 813–838. https://doi.org/10.1037/0033-2909.130.5.813

Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science*, *43*(7), 1029–1045. https://doi.org/10.1287/mnsc.43.7.1029

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2012). How quick decisions illuminate moral character. *Social Psychological & Personality Science*, *4*(3), 308–315. https://doi.org/10.1177/1948550612457688

Croson, R., & Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. *Journal of Risk and Uncertainty*, *30*(3), 195–209. https://doi.org/10.1007/s11166-005-1153-2

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 330–344. https://doi.org/10.1037/0096-1523.4.2.330

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889

Forsyth, D. K., & Burt, C. D. B. (2008). Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory & Cognition*, *36*(4), 791–798. https://doi.org/10.3758/MC.36.4.791

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

Hadjichristidis, C., Summers, B., & Thomas, K. (2014). Unpacking estimates of task duration: The role of typicality and temporality. *Journal of Experimental Social Psychology*, *51*, 45–50. https://doi.org/10.1016/j.jesp.2013.10.009

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280–1300. https://doi.org/10.1037/0003-066X.52.12.1280

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, *28*(4), 460–471. https://doi.org/10.1177/0146167202287004

Howard, R. C., Hardisty, D. J., Sussman, A. B., & Lukas, M. F. (2022). Understanding and neutralizing the expense prediction bias: The role of accessibility, typicality, and skewness. *JMR, Journal of Marketing Research*, *59*(2), 435–452. https://doi.org/10.1177/00222437211068025

Howlett, M. (2022). Looking at the 'field' through a Zoom lens: Methodological reflections on conducting online research during a global pandemic. *Qualitative Research*, *22*(3), 387–402. https://doi.org/10.1177/1468794120985691

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, *100*(4), 719–737. https://doi.org/10.1037/a0022408

Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*(3), 235–257. https://doi.org/10.1023/A:1007740225484

Jia, M. L., Li, X., & Krishna, A. (2020). Contraction with unpacking: When unpacking leads to lower calorie budgets. *The Journal of Consumer Research*, *46*(5), 853–870. https://doi.org/10.1093/jcr/ucz036

Johnson, E. J., Hershey, J., Meszaros, J., & Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty*, *7*(1), 35–51. https://doi.org/10.1007/BF01065313

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, *80*(4), 557–571. https://doi.org/10.1037/0022-3514.80.4.557

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. *Advances in Experimental Social Psychology*, *2*, 219–266. https://doi.org/10.1016/S0065-2601(08)60107-0

Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior*. General Learning Press.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, *76*, 728–741. https://www.jstor.org/stable/1806070

Kahneman, D., & Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*(5911), 276–278. https://doi.org/10.1126/science.1164951

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–238). University of Nebraska Press.

Keltner, D., Ellsworth, P. C., & Edwards, K. (1993). Beyond simple pessimism: Effects of sadness and anger on social perception. *Journal of Personality and Social Psychology*, *64*(5), 740–752. https://doi.org/10.1037/0022-3514.64.5.740

Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology*, *110*(5), 660–674. https://doi.org/10.1037/pspa0000050

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*(2), 149–166. https://doi.org/10.1521/soco.2016.34.2.149

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*(2), 210–229. https://doi.org/10.1037/pspa0000088

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 of the United States of America*, *115*(52), 13222–13227. https://doi.org/10.1073/pnas.1805327115

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, *40*(5), 586–598. https://doi.org/10.1016/j.jesp.2003.11.001

Krull, D. S., Seger, C. R., & Silvera, D. H. (2008). Smile when you say that: Effects of willingness on dispositional inferences. *Journal of Experimental Social Psychology*, *44*(3), 735–742. https://doi.org/10.1016/j.jesp.2007.05.004

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, *14*(4), 473–493. https://doi.org/10.1080/026999300402763

Loewenstein, G. F., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review*, *100*(1), 91–108. https://doi.org/10.1037/0033-295X.100.1.91

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review*, *106*(1), 210–214. https://doi.org/10.1037/0033-295X.106.1.210

Markman, K. D., & Beike, D. R. (2012). Regret. consistency. and choice: An Opportunity × Mitigation framework. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 305–325). Guilford Press.

Miller, D. T., & Kahneman, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Mulford, M., & Dawes, R. M. (1999). Subadditivity in memory for personal events. *Psychological Science*, *10*(1), 47–51. https://doi.org/10.1111/1467-9280.00105

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

O'Brien, E. (2023). *A flexible threshold theory of change perception in self, others, and the world*. PsyArXiv. https://doi.org/10.31234/osf.io/bg496

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. (2022). Losing sight of piecemeal progress: People lump and dismiss improvement efforts that fall short of categorical change—Despite improving. *Psychological Science*, *33*, 1278–1299. https://doi.org/10.1177/09567976221075302

O'Brien, E., & Ellsworth, P. C. (2012). Saving the last for best: A positivity bias for end experiences. *Psychological Science*, *23*(2), 163–165. https://doi.org/10.1177/0956797611427408

O'Brien, E., & Kardas, M. (2016). The implicit meaning of (my) change. *Journal of Personality and Social Psychology*, *111*(6), 882–894. https://doi.org/10.1037/pspi0000073

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*(2), 161–185. https://doi.org/10.1037/pspa0000070

Peetz, J., Buehler, R., Koehler, D. J., & Moher, E. (2015). Bigger not better: Unpacking future expenses inflates spending predictions. *Basic and Applied Social Psychology*, *37*(1), 19–30. https://doi.org/10.1080/01973533.2014.973109

Perry, J., & Hume, T. (2016, September 4). *Mother Teresa declared a saint before huge crowds in the Vatican*. CNN. https://www.cnn.com/2016/09/04/europe/mother-teresa-canonization/index.html

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Reddit. (2021). https://www.reddit.com/r/toastme/comments/ocw01r/just_cleared_my_first_year_of_post_graduate_in/

Redelmeier, D. A., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgement in medicine: Discounting unspecified possibilities. *Medical Decision Making*, *15*(3), 227–230. https://doi.org/10.1177/0272989X9501500305

Ronson, J. (2016). *So you've been publicly shamed*. Riverhead Books.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, *104*(2), 406–415. https://doi.org/10.1037/0033-295X.104.2.406

Russo, J. E., & Kolzow, K. J. (1994). Where is the fault in fault trees? *Journal of Experimental Psychology: Human Perception and Performance*, *20*(1), 17–32. https://doi.org/10.1037/0096-1523.20.1.17

Schwarz, N., & Clore, G. L. (2007). Feelings and phenomenal experiences. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 385–407). Guilford Press.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699. https://doi.org/10.1037/0022-3514.52.4.689

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 573–582. https://doi.org/10.1037/0278-7393.30.3.573

Taylor, A. (2016, September 1). Why Mother Teresa is still no saint to many of her critics. *Washington Post*. https://www.washingtonpost.com/news/worldviews/wp/2015/02/25/why-to-many-critics-mother-teresa-is-still-no-saint/

Thorndike, E. L. (1911). *Animal intelligence*. Macmillan.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. https://doi.org/10.1037/a0018963

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567. https://doi.org/10.1037/0033-295X.101.4.547

Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law & Society Review*, *31*(2), 237–266. https://doi.org/10.2307/3053926

Uhlmann, E., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*(6), 474–480. https://doi.org/10.1111/j.0956-7976.2005.01559.x

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689–690. https://doi.org/10.1111/j.1467-9280.2007.01961.x

Van Boven, L., & Epley, N. (2003). The unpacking effect in evaluative judgments. *Journal of Experimental Social Psychology*, *39*(3), 263–269. https://doi.org/10.1016/S0022-1031(02)00516-4

Van Boven, L., Loewenstein, G., Dunning, D., & Nordgren, L. F. (2013). Changing places: A dual judgment model of empathy gaps in emotional perspective taking. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 48, pp. 117–171). Academic Press.

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, *14*(3), 131–134. https://doi.org/10.1111/j.0963-7214.2005.00355.x

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *78*(5), 821–836. https://doi.org/10.1037/0022-3514.78.5.821

Received September 18, 2021
Revision received December 3, 2022
Accepted December 13, 2022 ■