Journal of Experimental Social Psychology 103 (2022) 104381

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Repeated exposure to success harshens reactions to failure☆

Kristina A. Wald *, Ed O'Brien

University of Chicago, USA

ARTICLE INFO

Keywords:
Repeated exposure
Change perception
Performance
Criticism
Social judgment

ABSTRACT

Eight experiments reveal that observing success wields adverse effects on how people judge others' failures. We find that repeated exposure to successful performances leads people to perceive a task as not so difficult or complicated after all—and so they more harshly evaluate those who struggle to do it. Across various tasks—from completing a motor-skills test, to doing a dance move, to sketching a drawing—repeated exposure to success led people to expect others to perform better on their first-time attempts (Experiments 1–2) and criticize others for failing (Experiments 3–4), even when the performer was unequipped to succeed (and when harsh critics themselves were mistakenly overconfident). This effect was not explained by incidentally negative effects of repeated exposure (e.g., annoyance, tiredness: Experiment 5); instead, it indeed depended on people's (mis) perceptions of learning from mere watching, such that it was moderated by the observability of successful execution (Experiment 6) and whether one judged another person's poor attempt (which entails mere watching) vs. one's own poor attempt (which entails watching *plus doing*: Experiment 7). Accordingly, harsh critics became kinder to others after attempting the task themselves (Experiment 8). These findings reveal when and why observing success risks callousing people toward failure. This blinding effect is especially consequential in today's information age, which offers unprecedented access to success and skilled performances (e.g., via social media): Such access may not only be inflating people's confidence for recreating what they see, but also deflating people's empathy and understanding of those who try (and fail).

Word count: 250.

"Don't criticize what you can't understand."
–Bob Dylan (1964).

Failure abounds. Chefs burn soufflés, actors miss marks, receivers drop passes, and students bomb exams. Everyone, from every walk of life, has at some point tried and failed.

And yet, despite this universal basis for shared understanding, everyday life has no shortage of critics—from television viewers who readily shout at expert performers from the comforts of their couch, to inexperienced audience members who readily jeer at the flubs of invited presenters. Such examples are especially psychologically interesting because the critics in question would presumably perform just as poorly—or far worse—if they were to attempt the same task that they readily condemn others for fumbling. Why are people so quick to criticize?

Of course, there could be valid reasons to criticize others' failures (e.g., couch-shouters may simply be demanding better value from highly-paid professionals), and doing so could serve other goals beyond trying to harm the target (e.g., engaging in nasty gossip to build social bonds: Dunbar, 2004). Other motivational factors may similarly spur criticism, such as being motivated to signal identity (Berendt & Uhrich, 2016; Hornsey, Oppes, & Svensson, 2002; Packer, 2008; Sedikides, 1993; Wilson & Ross, 2001) and to gain psychological leverage (Hareli & Weiner, 2002; Lyubomirsky & Ross, 1997; van de Ven, Zeelenberg, & Pieters, 2009; Wills, 1981).

We explore a yet-untested factor that could harshen reactions to failure over and above such explicit motivations: People may turn into harsher critics of others' failures as a function of observing *success*. Building on findings from Kardas and O'Brien (2018)—who documented an "illusion of skill acquisition" whereby overexposure to instructional videos led participants to overestimate *their own* performance abilities—we advance this idea by exploring downstream dynamics for *social* judgment. Even when performers have no basis to do well, and even when judges could do no better themselves if prompted to try, we propose (and find) that repeated exposure to success can blind people to the underlying difficulties of a task and thereby lead people to

☆ This paper has been recommended for acceptance by Dr Vanessa Bohns

* Corresponding author at: Behavioral Science, University of Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, USA.
E-mail address: kristina.wald@chicagobooth.edu (K.A. Wald).

https://doi.org/10.1016/j.jesp.2022.104381
Received 24 November 2021; Received in revised form 4 July 2022; Accepted 5 July 2022
Available online 26 August 2022

unfairly admonish others' failed attempts. In other words, people may sometimes criticize others simply because they have come to *truly believe* that the task is not so difficult after all, and so view their attacks as justified—even when others were doomed to fail all along.

1. (Over)exposure to success

The first part of our theorizing revolves around the effects of observing success. We focus primarily on motor-based skills, as they allow for concrete operationalizations of our constructs of interest and represent a large, consequential domain. Indeed, people live in a time of unprecedented access to others' skilled performances, from the millions of online instructional videos now just a click away, to streaming and on-demand services that have been programmed to showcase the world's top athletic feats and creative minds, to curated social-media profiles that are disproportionately populated by users' best moments (Chou & Edge, 2012; Jordan et al., 2011; O'Brien, Kristal, Ellsworth, & Schwarz, 2018). As Eskreis-Winkler and Fishbach (2020) concretely highlight: "For every two 'success' videos uploaded to YouTube (~25 million), there is about one 'failure' (~10.9 million)" (p. 57).

In principle, today's widespread access to success should come with many psychological benefits: A long history of research highlights the benefits of observing successful exemplars for one's own learning (e.g., novices who get to watch an expert instructor model a complicated task are more likely to try the task themselves and stick with subsequent practice: Andrieux & Proteau, 2016; Bandura, 1977; Scully & Newell, 1985). This learning has even been shown to occur at more implicit levels (e.g., implicit procedural learning and "mirror neuron" mimicry: Heyes, 2001; Heyes & Foster, 2002; Lyons, Young, & Keil, 2007; Mattar & Gribble, 2005).

In practice, however, there may also be unintended costs for viewers to the extent that they grow *over*exposed to success—which, given today's widespread access, is presumably an increasingly common (yet under-studied) state that people find themselves in. This is precisely the concern raised by the "illusion of skill acquisition" (Kardas & O'Brien, 2018), the finding that overexposure to success (e.g., watching an instructional video once vs. 20 times repeatedly) can promote overconfidence (e.g., repeated watchers come away expecting to perform better themselves—even for tasks in which overexposure has no effect on improving their attempts).

The current research will seek to advance these initial findings on *self*-judgment by exploring overexposure effects on *social* judgment. How and why might overexposure to success affect how people judge *others'* failures (even if others have no basis to do well to begin with)? Going further, we also uniquely test more specific mechanisms underlying the effects on both social and self-judgment.

2. Judging others' failures

A large literature suggests that watching success over and over again, just like doing anything over and over again, risks eliciting desensitization (Frederick & Loewenstein, 1999; Galak & Redden, 2018; Groves & Thompson, 1970; Wilson & Gilbert, 2008; Wolpe, 1982): Extensive watching may lead people to grow dull to the genuine difficulty of a complex skill. This problem is likely further exacerbated when people who judge others are *merely* watching, without ever trying the task themselves—which, again, is presumably a highly common state. By definition, repeated watching gives people more time to fluently track and memorize what steps to take—but without corresponding physical feedback, *merely* watching cannot as precisely convey how those steps feel upon taking them (e.g., consider all of the nerves, bodily sensations, and other bottom-up internal states that can be evoked in the act of doing and are critical for the actual execution of complex skills and behaviors: Ericsson, Krampe, & Tesch-Römer, 1993; Kolb, 2014). Put another way: Repeated exposure to success may create an "empathy gap" (Bohns, 2016; Campbell, O'Brien, Van Boven, Schwarz, & Ubel, 2014; Van Boven, Loewenstein, Dunning, & Nordgren, 2013) in people's ability to appreciate how difficult that skill would be for others to perform, even if they would be performing it without any requisite practice.[1]

Indeed, the degree to which people empathize with others' struggles largely depends on their own situational appraisals (Ellsworth & Scherer, 2003; Smith & Ellsworth, 1985; Wondra & Ellsworth, 2015), which reflects a broader psychological tendency for people to egocentrically anchor their estimations of another person's thoughts and feelings on their own thoughts and feelings (Ames, 2004; Epley, Keysar, Van Boven, & Gilovich, 2004; Katz & Allport, 1931; Nickerson, 1999; O'Brien & Ellsworth, 2012; Ross, Greene, & House, 1977; Van Boven et al., 2013). If repeated exposure to success dulls perceived difficulty and leads people to view the skill as not so complicated or impressive after all, then their subsequent reactions to failed attempts may be amplified in kind. Not only might people raise unfair expectations in their own abilities to perform the skill, but they may also raise unfair expectations in *anybody's* ability to do so—and thus, upon observing others try (and fail), people may be unfairly quick to criticize them for fumbling something so "easy." Other research highlights that people struggle to backtrack from personal change when recalling past states (e.g., experts who try to take the perspective of novices often fail to account for how little those novices actually know: Camerer, Loewenstein, & Weber, 1989; Ross, 1989).

These ideas suggest an intriguing possibility: Exposure to success may not only make people worse judges of themselves (Kardas & O'Brien, 2018), but also of *others* (the current research); not only might people grow (over)confident in their abilities to perform the skill on their own first attempts, but also (over)confident in others' abilities to perform well, too—even in cases when *no* party (neither failed performer nor harsh critic) is equipped to succeed. Today's unprecedented access to success may not only be inflating people's self-confidence (as primarily assessed in past research), but also *deflating* people's empathy and understanding of others' struggles (as primarily and uniquely assessed in the current research).

Throughout our research, we also sought to uniquely disentangle this *exposure hypothesis* from something closer to an *inference hypothesis*. What we mean by this is as follows: Suppose that we overexpose participants to success and then ask them to judge a fellow watcher's poor attempt—and indeed find that they are harsher to criticize it. From this set-up alone, it is unclear whether judges were harsher to criticize because (i) they themselves grew desensitized to seeing success many times—what we call an *exposure hypothesis*; or instead because (ii) they merely know that this other person had also seen success many times themselves before their attempt, and so believe they should have done better because of it—what we call an *inference hypothesis*.

That is, a second (and not mutually exclusive) mechanism may be that people believe that watching success repeatedly actually helps train one to do the task better (even if it does not in reality); therefore, people may hold performers to higher standards when they know that the *performer* has watched success repeatedly (regardless of whether the judge has watched success themselves). In other words, rather than requiring any desensitization dynamics, people may simply hold overgeneralized lay beliefs about the benefits of observational learning. The notion that "practice makes perfect" is now widely popularized (Duckworth & Gross, 2014; Dweck, 2008), with people believing even minimal practice should be effective (Sanchez & Dunning, 2018). Thus, people may be harsher to judge when they know that others were able to

[1] This framework thus depends on whether a task *looks* easy/hard vs. whether it *is* easy/hard. By design, we focus on the combination of "looks easy" plus "is hard" (as this represents a prevalent and consequential domain: Kardas & O'Brien, 2018), but note that others also exist in everyday life. We expand upon this idea in the General Discussion. Note also that Experiments 5–8 will assess boundary conditions of "looks easy" (holding success and task constant).

repeatedly observe success prior to their attempt, even when doing so does not help; people may hold others to higher standards simply when they know that others had first "trained" merely by observation—independent of whether judges themselves watch too.

This distinction is important because it suggests differences in how generalizable the effects may be. Exposure effects suggest wide generalizability; in the act of watching, people may come to perceive the task as easier and so believe *any*one should perform it better (even non-watchers and naïve first-timers). Inferences effects suggest narrower generalizability; people may specifically believe any *extensive watcher* (but not *any*one in general) should do it better.

We will uniquely test inference effects (i.e., when judges are told that *the other person* was overexposed to success) as they occur independently from exposure effects (i.e., when judges *themselves* are overexposed to success). Both mechanisms, however, make the same prediction: Exposure to success may risk *unfairly undermining* how people come to judge others.

3. The current research

Eight experiments (total $N = 6370$) explored the hypothesis that overexposure to success may lead people to more harshly evaluate others who try (and fail). Experiments 1–2 assessed people's expectations about first-time performance attempts as a function of their exposure to success. Experiments 3–4 (and all subsequent experiments) assessed whether this effect fosters interpersonal criticism. Experiments 5–8 assessed theory-driven boundaries of this effect. For example, Experiment 8 tested a strategy to soften people's criticism, derived from this process: After attempting the skill themselves, people may better appreciate its underlying complexities (as physical feedback helps close this "empathy gap") and so treat others' failures more kindly.

4. Experiment 1: Repeated exposure to success inflates expectations of first-time attempts

In Experiment 1, participants repeatedly watched a successful motor-skills performance. We hypothesized that overexposure to success would inflate people's expectations of how well *others* should do—even in cases when others would be attempting the task for the first time, and when merely watching success could not have helped *any* party perform any better to begin with.

In this and all experiments, we predetermined sample sizes of 150 participants per cell (or more, depending on resources), which provides ample power for detecting the typical effect sizes found in research using similar designs (Simmons, Nelson, & Simonsohn, 2018). We report all manipulations, measures, and exclusions (if any). All experiments were preregistered. For each experiment, we report a sensitivity analysis of the minimum effect size each of our samples had power to detect (via G*Power [Faul, Erdfelder, Lang, & Buchner, 2007]). Data, materials, and preregistrations are on the Open Science Framework (OSF): https://osf.io/m623y/.

4.1. Method

4.1.1. Participants

We requested 1000 participants from Amazon's Mechanical Turk, yielding 1006 ($M_{age} = 36.53$, $SD_{age} = 11.03$; 48.51% female; 29.03% non-White) who completed the study for $1.00. An additional 193 participants started the survey but did not finish it (49 in the low exposure conditions, 56 in the high exposure conditions, and 88 who dropped out before being assigned to a condition).

4.1.2. Procedure

Participants were randomly assigned to a 2 (Exposure to Success, between-subjects: low vs. high) × 3 (Prediction Target, between-subjects: self vs. similar other vs. naïve other) × 2 (Phase, within-subjects: predicted scores vs. own actual scores) mixed-factor design.

All participants were informed that they would learn about a motor-skills test, answer some questions about it, and then actually attempt it. First, they clicked through a visual tutorial that thoroughly explained this test, which was validated for online use (Cusack, Vezenkova, Gottschalk, & Calin-Jageman, 2015; see OSF for a copy). In the test, test-takers must accurately trace the shape of a maze using a computer trackpad. The test assesses fine-grained motor skills by requiring test-takers to trace in reverse (e.g., needing to move their finger downward in order to trace upward), as quickly as possible. Scores are automatically and objectively recorded by the testing software, ranging from 0% (the entirety of their trace fell off the correct path) to 100% (the entirety of their trace fell on the correct path). This particular stimulus is ideal for the current research because Kardas and O'Brien (2018, Experiment 4) found that it indeed fits our targeted combination of "looks easy" plus "is hard"—such that overexposure to a successful exemplar does *not* actually improve people's own performance, despite their expectations to the contrary.

All participants then watched a video of another person successfully completing the test, earning the high score of 94%. We manipulated the number of times that the video looped: Low-exposure participants watched it once (about 10 s); high-exposure participants watched it 20 times repeatedly (about 3 min). All participants were instructed to do nothing else except passively watch the video (e.g., not to mimic the person's movements with their own hands).

After watching, participants made predictions about a randomly-assigned target who would attempt the motor-skills test themselves, with no additional training. Self-participants were asked: "What score do you think you will earn?" Responses were recorded from 0% to 100% corresponding to the actual scale of the motor-skills test. This item conceptually replicates what Kardas and O'Brien (2018) tested in the "illusion of skill acquisition"; higher (vs. lower) exposure to success should thus lead our participants to be more confident in *their own* abilities.

Novel to our research, remaining participants judged *others*. Similar-other participants were asked: "Based on their video watching experience (they watch the same video you did, the same number of times as you), what score do you think a randomly-selected other participant like this will earn?" This allows us to test our inference hypothesis (and perhaps our exposure hypothesis). Naïve-other participants were asked: "Based on their video watching experience (they watch the same video you did, the same number of times as you, except the video cuts out a bit), what score do you think a randomly-selected other participant like this will earn?" This uniquely tests our exposure hypothesis, as participants were asked to judge another participant who had *not* had extensive exposure to success themselves.

After making predictions, all participants then attempted the motor-skills test themselves and their scores were automatically recorded (thus allowing us to calculate [over]confidence).

4.1.2.1. Other variables. To end the study, all participants reported their age, ethnicity, and gender; how they watched the video (forced-choice: watched passively vs. actively); how quickly they made their attempt (forced-choice: went as fast as I could vs. did not go as fast as I could); the extent to which they practiced the hand motions before their attempt (forced-choice: not at all, a little bit, moderate, quite a bit, a lot); their device (forced-choice: a track-point or trackpad vs. a mouse vs. other [explain]); any technical difficulties (forced-choice: no vs. yes [explain]); and whether they had ever taken a similar study before (forced-choice: yes vs. no).

Finally, all participants completed 2 attention checks: one regarding how many times they themselves watched the expert video (forced-choice: 1× vs. 20×) and one regarding the target of their predictions (forced-choice: 3 options corresponding to each Target condition).

4.2. Results and discussion

The motor-skills testing software automatically recorded each participants' completion time in tracing the maze. As indicated in our preregistration, we excluded participants for whom the output showed a time of 0 s, meaning that they did not complete the test at all. We excluded 43 participants on this basis (24 from the low exposure conditions, 19 from the high exposure conditions), leaving 963 participants included in our analyses below. Sensitivity analyses (using correlations among the repeated measures; $\alpha = 0.05$ at 80% power) suggest that this sample size can detect an effect size of $\eta_p^2 = 0.005$ for a 3-way interaction.

We conducted a Repeated-Measures GLM with Exposure (judges watched success: 1× vs. 20×) and Target (judges then evaluated: themselves vs. similar others vs. naïve others) as between-subjects factors, and Phase (judge's predicted score of the target vs. judge's own actual score) as a within-subjects factor, with motor-skills score (0%–100%) as the dependent variable.

In terms of key output, there was a main effect of Exposure, $F(1, 957) = 20.55$, $p < .001$, $\eta_p^2 = 0.02$, such that watching success 20× repeatedly resulted in higher scores as compared to watching just 1×. Critically, however, this effect was qualified by an Exposure × Phase interaction, $F(1, 957) = 16.17$, $p < .001$, $\eta_p^2 = 0.02$, suggesting that watching success differentially influenced expected performances vs. actual performances. Also critically, this 2-way interaction was *not* further qualified by a 3-way interaction with Target, $F(2, 957) = 0.34$, $p = .714$, $\eta_p^2 = 0.001$, suggesting these differential effects do not differ by target (see Supplemental Material for all remaining main effects and interactions, which we assume are incidental to our hypothesis regardless of significance). Below we unpack these results via separate pairwise comparisons.

4.2.1. Inflated expectations of first-time attempts

Observing success 20× (vs. 1×) indeed inflated participants' *expected* performance of first-time attempts—even among those participants who predicted *others'* scores (see Fig. 1).

4.2.1.1. Exposure to success mistakenly inflates self-expectations (replicates Kardas & O'Brien, 2018). First, overexposure to success led people to grow overconfident in *their own* skills: Viewers who watched the successful performance 20× repeatedly—without engaging in any other training or practice themselves—then predicted they would earn a higher score on their own very first attempt at the same motor-skills task ($M = 65.82$ points, $SD = 19.14$) as compared to viewers who watched just 1× ($M = 58.67$ points, $SD = 24.67$), $F(1, 957) = 10.14$, $p = .002$, $\eta_p^2 = 0.01$ ($d = 0.32$). Yet higher expectations did not translate to reality: 20× participants performed no better ($M = 36.56$ points, $SD = 23.87$) than 1× participants ($M = 35.40$ points, $SD = 24.28$), $F(1, 957) = 0.18$, $p = .673$, $\eta_p^2 < 0.001$ ($d = 0.05$).[2]

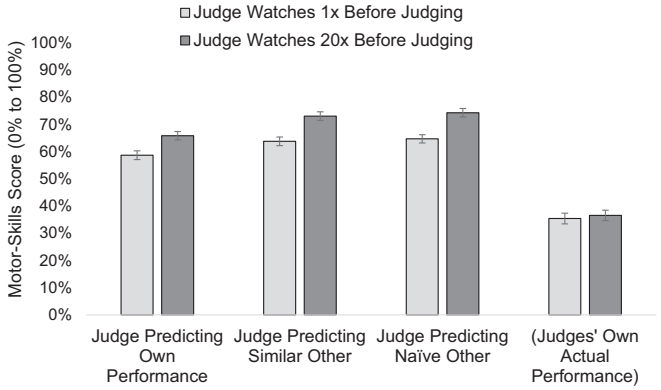


Fig. 1. Experiment 1: Predicted motor-skills performance as a function of exposure to success. The right-most pair of bars shows judges' actual performance. Error bars show ±1 *SE*.

4.2.1.2. Exposure to success also inflates expectations of others (unique to the current research). More critically for the current research, we also found a corresponding effect in terms of inflated expectations of *others*. Among participants who predicted the scores of similar others, participants who watched the video 20× repeatedly expected other watchers to perform better on the *others' own* first attempt, without any other training or practice ($M = 73.03$ points, $SD = 16.95$), as compared to participants who watched just 1× ($M = 63.77$ points, $SD = 20.79$), $F(1, 957) = 16.89$, $p < .001$, $\eta_p^2 = 0.02$ ($d = 0.49$)—suggesting support for our inference hypothesis. This effect held even among participants who predicted the scores of *naïve* others: Judges who watched success 20× repeatedly expected these naïve others to perform better ($M = 74.27$ points, $SD = 17.37$) as compared to the judges who watched just 1× ($M = 64.69$ points, $SD = 19.98$), $F(1, 957) = 19.24$, $p < .001$, $\eta_p^2 = 0.02$ ($d = 0.51$)—supporting our exposure hypothesis.

4.2.2. Other variables

Finally, most participants passed the attention checks (Exposure: 98.23%, 946 of 963; Target: 75.91%, 731 of 963); watched passively (80.58%, 776 of 963); traced quickly in their attempt (77.57%, 747 of 963); did not practice beforehand (60.33% "not at all," 581 of 963); did use a track-point or trackpad (88.47%, 852 of 963); reported no technical difficulties (89.82%, 865 of 963); and reported never taking a similar study (96.78%, 932 of 963). All key patterns hold when excluding participants based on these other variables (see Supplemental Material).

Experiment 1 provides initial support for our hypothesis. Exposure to success inflated expectations even of *others'* first-time attempts, despite yielding no actual performance benefits.

5. Experiment 2: A replication with judgments of fully naïve others

Experiment 2 sought to directly replicate Experiment 1 for robustness: Most critically, we used a different phrasing in the "naïve other" condition so as to ensure that judges knew they were predicting the performance of fully naïve others who had missed out on watching success.

5.1. Method

5.1.1. Participants

We requested and yielded 1000 participants from Amazon's Mechanical Turk ($M_{age} = 35.17$, $SD_{age} = 11.00$; 56.40% female; 26.00% non-White), who completed the study for $1.00. An additional 259

[2] Incidentally, the participants in our remaining conditions *also* did not exhibit any self-performance differences: similar-other participants, 20× ($M = 35.64$, $SD = 25.21$) vs. 1× ($M = 35.28$, $SD = 24.06$), $F(1, 957) = 0.02$, $p = .896$, $\eta_p^2 < 0.001$ ($d = 0.01$); naïve-other participants, 20× ($M = 37.34$, $SD = 25.14$) vs. 1× ($M = 37.76$, $SD = 23.35$), $F(1, 957) = 0.03$, $p = .874$, $\eta_p^2 < 0.001$ ($d = -0.02$). The right-most pair of bars in Fig. 1 thus collapses across prediction target. Overall, more Watch 20× participants [84.17%, 404 of 480] than Watch 1× participants [79.30%, 383 of 483] tended to be overconfident, $\chi^2(1) = 3.82$, $p = .051$, $w = 0.06$, meaning their predicted score was higher than their actual score. Note, however, two limitations to this finding. First, the result does not surpass the standard $p = .05$ cutoff for significance—but, it is marginal, in the hypothesized direction, and the same result does indeed surpass this cutoff in Experiment 2 (see next footnote). Second, these results become statistically weaker (both here and in Experiment 2) after excluding participants based on attention-check failures and so on—but, this is presumably due to the decreased power (see Supplemental Material).

participants started the survey but did not finish it (55 in the low exposure conditions, 76 in the high exposure conditions, and 128 who dropped out before being assigned to a condition).

5.1.2. Procedure

Procedures were identical to Experiment 1, except naïve-other participants read: "Based on their video watching experience (they're assigned to watch the same video you did, the same number of times as you – but the video freezes after the first 3 seconds, and the survey goes right to the game without showing the rest of the video), what score do you think a randomly-selected other participant like this will earn?" Thus, it was made even more explicitly clear that these other targets in question were fully naïve to having seen the successful product.

Simply for clarity, we also adapted the Target attention check such that participants first reported which score they predicted (forced-choice: self vs. other); then, those who indicated "other" were prompted to specify this other target (forced-choice: similar-other vs. naïve-other).

5.2. Results and discussion

Again, as indicated in our preregistration, we excluded 38 participants whose timer showed 0 s (19 from the low exposure conditions, 19 from the high exposure conditions), leaving 962 participants included in our analyses below. Sensitivity analyses (using correlations among the repeated measures; $\alpha = 0.05$ at 80% power) suggest that this sample size can detect an effect size of $\eta_p^2 = 0.005$ for a 3-way interaction.

5.2.1. Key results

All results replicated (see Fig. 2): the main effect of Exposure, $F(1,956) = 15.41$, $p < .001$, $\eta_p^2 = 0.020$, the critical Exposure × Phase interaction, $F(1,956) = 9.56$, $p = .002$, $\eta_p^2 = 0.010$, and *no* 3-way interaction with Target, $F(1,956) = 0.26$, $p = .774$, $\eta_p^2 = 0.001$ (again, see Supplemental Material for remaining output, which is incidental to our hypothesis regardless of significance). Exposure to success not only inflated how participants thought *they* would do ($M_{Watch20x} = 66.19$ points, $SD = 19.98$ vs. $M_{Watch1x} = 57.72$ points, $SD = 21.39$), $F(1,956) = 13.19$, $p < .001$, $\eta_p^2 = 0.014$ ($d = 0.41$), but also inflated how they thought *similar others* would do ($M_{Watch20x} = 75.02$ points, $SD = 16.14$ vs. $M_{Watch1x} = 66.52$ points, $SD = 18.92$), $F(1,956) = 14.02$, $p < .001$, $\eta_p^2 = 0.014$ ($d = 0.48$), and how they thought *naïve others* would do ($M_{Watch20x} = 61.66$ points, $SD = 21.51$ vs. $M_{Watch1x} = 56.92$ points, $SD = 23.80$), $F(1,956) = 4.41$, $p = .036$, $\eta_p^2 = 0.010$ ($d = 0.21$).[3]

5.2.2. Other variables

Again, most participants passed the attention checks (Exposure: 98.44%, 947 of 962; Target: 92.83%, 893 of 962); watched passively (82.64%, 795 of 962); traced quickly in their attempt: 75.88%, 730 of 962); did not practice beforehand (62.68% "not at all," 603 of 962); did use a track-point or trackpad (90.96%, 875 of 962); reported no technical difficulties (90.12%, 867 of 962); and reported never taking a similar study (98.02%, 943 of 962). All key patterns hold when excluding participants based on these other variables (see Supplemental Material).

[3] Again, as in Experiment 1, exposure to success did not improve *actual* performances: self-participants, 20× ($M = 38.14$, $SD = 24.22$) vs. 1× ($M = 36.62$, $SD = 24.43$), $F(1,956) = 0.30$, $p = .584$, $\eta_p^2 < 0.001$ ($d = 0.06$); similar-other participants, 20× ($M = 38.06$, $SD = 24.56$) vs. 1× ($M = 33.89$, $SD = 24.08$), $F(1,956) = 2.35$, $p = .126$, $\eta_p^2 = 0.002$ ($d = 0.17$); naïve-other participants, 20× ($M = 34.52$, $SD = 24.91$) vs. 1× ($M = 37.52$, $SD = 26.07$), $F(1,956) = 1.23$, $p = .268$, $\eta_p^2 = 0.001$ ($d = -0.12$). The right-most pair of bars in Fig. 2 thus collapses across prediction target. Overall, more Watch 20× participants [83.23%, 392 of 471] than Watch 1× participants [76.37%, 375 of 491] were overconfident, $\chi^2(1) = 6.98$, $p = .008$, $w = 0.09$, meaning their predicted score was higher than their actual score.

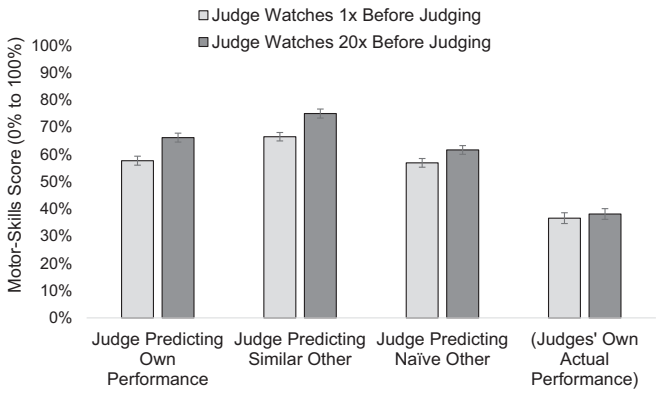


Fig. 2. Experiment 2: Motor-skills results replicating Experiment 1. Error bars show ±1 *SE*.

Experiments 1–2 show that overexposure to success indeed inflates people's expectations even of *others'* first-time attempts, and even when mere watching does not actually help anyone.

Next, Experiments 3–4 (and all subsequent experiments) assessed the implications of these inflated expectations of others' performances for spurring criticism of their failed attempts.

6. Experiment 3: Repeated exposure to success harshens criticism of failure (motor-skills test)

Experiment 3 tested this hypothesis in the same context of Experiments 1–2: the motor-skills test. We hypothesized that overexposure to success may inflate criticism of others' failures. In addition, we assessed whether this effect holds even among participants who are objectively overconfident—thus going beyond the possibility that harsher critics *really are* better at the test.

6.1. Method

6.1.1. Participants

We requested 600 "Cloud Approved" participants from Cloud Research, yielding 603 ($M_{age} = 39.44$, $SD_{age} = 11.91$; 46.10% female; 25.70% non-White) who completed the study for $1.00.[4] An additional 116 participants started the survey but did not finish it (28 in the low exposure condition, 43 in the high exposure condition, and 45 who dropped out before being assigned to a condition).

6.1.2. Procedure

Participants were randomly assigned to a 2 (Exposure to Success, between-subjects: low vs. high) × 2 (Prediction Target, within-subjects: self vs. other) × 2 (Phase, within-subjects: predicted self-score vs. actual self-score) mixed-factor design.

Procedures were generally similar to Experiment 1, but with the following key changes. First: Given that we found no differences across *types* of others in Experiments 1–2, here we only included the "similar other" condition (simply for ease of comparison). Also, after watching, all participants predicted both their own score and a similar other's score, one at a time in random order (as opposed to manipulating this factor *between*-subjects, like we did in Experiments 1–2).

Second, and more critically: Participants were informed of this other participant's actual score and reported their criticism of it, comprising our key dependent variable. This single item was rated after participants made both predictions and before they made their own attempt. The

[4] We initially preregistered to request half this initial number of participants, but we then preregistered (via a separate post-hoc preregistration) to double this number to ensure higher power. See OSF for both preregistrations.

other participant always scored poorly—we programmed the item such that they scored half of whatever judge-participants had expected them to score. We measured criticism by asking participants to respond to the following measure: "Now that I think about it: They deserve some criticism for this score" (from 1 = *don't agree at all*, 10 = *completely agree*).

6.1.2.1. Other variables. Finally, participants made their own attempt—allowing us to directly test the extent to which (over)confidence is associated with criticizing the other participant's poor attempt. They also completed the same end-of-study items from Experiment 1; the attention check for "target" referred to how well the other participant did (forced-choice: well vs. poorly).

6.2. Results and discussion

Again, as indicated in our preregistration, we excluded 27 participants whose timer showed 0 s (14 from the low exposure condition, 13 from the high exposure condition), leaving 576 participants included in our analyses below. Sensitivity analyses ($\alpha = 0.05$ at 80% power) suggest that our sample size can detect an effect size of $d = 0.23$ for testing differences between two independent means.

6.2.1. Key results

We conducted two separate independent-samples *t*-tests with Exposure (judges first watched success: 1× vs. 20×) as the independent variable: In the first, judges' expectations of how well the *other participant* should do (0%–100%) was the dependent variable, and in the second, judges' subsequent *criticism* of the other participant's poor attempt (1−10) was the dependent variable.

6.2.1.1. Exposure to success inflates expectations of others' performance. First, replicating Experiments 1–2, participants who watched success 20× then expected the other participant to perform better ($M = 72.16$ points, $SD = 16.46$) than participants who watched success just 1× ($M = 64.18$ points, $SD = 18.86$), $t(574) = -5.40$, $p < .001$, $d = -0.45$.

6.2.1.2. Exposure to success harshens criticism of others' failure. Second, participants who watched success 20× also gave *harsher criticism* of the other participant's poor attempt ($M = 4.35$, $SD = 2.58$) than participants who watched success just 1× ($M = 3.89$, $SD = 2.58$), $t(574) = -2.16$, $p = .032$, $d = -0.18$.

6.2.1.3. Harshened criticism holds among overconfident judges—if anything, it is stronger. Third, we assessed participants' *own* predicted and actual performance. As in Experiments 1–2, we replicated Kardas and O'Brien's (2018) "illusion of skill acquisition" (via the same analyses there; again, see Supplemental Material for full output), as reflected in a main effect of Exposure ($F(1, 574) = 11.47$, $p < .001$, $\eta_p^2 = 0.02$) that was qualified by an Exposure × Phase interaction ($F(1, 574) = 12.35$, $p < .001$, $\eta_p^2 = 0.02$): Repeated exposure to success improved *predicted* scores (20×, $M = 67.45$ points, $SD = 20.28$ vs. 1×, $M = 58.30$ points, $SD = 22.69$; $F(1, 574) = 25.93$, $p < .001$, $\eta_p{}^2 = 0.04$ ($d = 0.43$) but not *actual* scores 20×, $M = 35.87$ points, $SD = 23.59$ vs. 1×, $M = 35.52$ points, $SD = 23.40$; $F(1, 574) = 0.03$, $p = .858$, $\eta_p{}^2 < 0.001$ ($d = 0.01$).

For each participant, we subtracted their actual score from their predicted score, yielding a self-overconfidence index. Overall, more Watch 20× participants (86.17%, 243 of 282) than Watch 1× participants (74.83%, 220 of 294) were overconfident, $\chi^2(1) = 11.74$, $p < .001$, $w = 0.14$. When re-running our criticism analyses while including overconfidence as a factor (via linear regression), the effect of Exposure remained significant,$b = .12$, $p = .031$) and was robust to overconfidence (Exposure × Overconfidence interaction: $b = -.23$, $p = .088$)—if anything, the *more* overconfident participants were, the *more* they criticized others, indicated by a positive-coefficient main effect of overconfidence ($b = .44$, $p < .001$ (and likewise by a positive correlation between overconfidence and criticism ($r = 0.242$, $p < .001$).

6.2.1.4. Exploratory mediation. Putting these effects together, we also decided (after the fact; not preregistered) to explore their relative strengths in driving criticism. We ran a mediation analysis (SPSS PROCESS Model 4, 5000 bootstrapped iterations) using Exposure as the independent variable, criticism as the dependent variable, and self-overconfidence (judges' own self-predicted score minus self-actual score) and social predictions (judges' predictions of the other participant's score) as simultaneous mediators. The effect of repeated exposure to success on one's criticism of failure was in fact mediated by self-overconfidence, indicated by a significant indirect effect of this variable, $b = 0.15$, $SE = 0.06$, 95% CI_{boot} [0.06, 0.28]—and, interestingly, this driving role of self-overconfidence was stronger than that of judges' *explicit social predictions* about the other person, which did *not* have a significant indirect effect, $b = 0.07$, $SE = 0.06$, 95% CI_{boot} [−0.04, 0.19].

6.2.2. Other variables

Again, most participants passed the attention checks (Exposure: 98.44%, 567 of 576; Target: 91.32%, 526 of 576); watched passively (86.28%, 497 of 576); traced quickly in their attempt (73.26%, 422 of 576); did not practice beforehand (69.10% "not at all," 398 of 576); did use a track-point or trackpad (89.93%, 518 of 576); reported no technical difficulties (94.27%, 543 of 576); and reported never taking a similar study (93.75%, 540 of 576). All key patterns hold when excluding participants based on these other variables (some output fell beyond the $p < .05$ significance threshold, presumably due to the decreased power; see Supplemental Material).

Experiment 3 extends our evidence thus far by showing that inflated expectations for others—as a result of overexposure to success—indeed translates into harshened criticism (even, and perhaps especially, among extensive viewers who grew overconfident in their own abilities).

7. Experiment 4: Repeated exposure to success harshens criticism of failure (dance move)

Next, Experiment 4 tested this idea via a new domain (judging a dance move—the moonwalk) and new measures. In addition, we utilized a new design to further tease apart an *exposure hypothesis* (as people repeatedly watch success, this exposure may harshen their criticism of failure) from an *inference hypothesis* (right from the start—regardless of their own exposure—people may be harsher to criticize anyone who has repeatedly watched success).

7.1. Method

7.1.1. Participants

We requested 1000 participants from Amazon's Mechanical Turk, yielding 999 ($M_{age} = 39.30$, $SD_{age} = 12.77$; 52.55% female; 24.32% non-White) who completed the study for $1.00. An additional 35 participants started the survey but did not finish it (16 in the low judge exposure conditions, 12 in the high judge exposure conditions, and 7 who dropped out before being assigned to a condition).

7.1.2. Procedure

Participants were randomly assigned to a 2 (Judge's Exposure to Success: low vs. high) × 2 (Performer's Exposure to Success: low vs. high) fully between-subjects design.

All participants were informed that they would judge the performance of a real participant who attempted the "moonwalk" dance move. First, they watched a video of an expert dance instructor performing the

moonwalk, which played either once in full (about 5 s) or 20 times repeatedly (about 2 min) based on random assignment (see OSF for these videos).

Next, participants were informed that participants in another study also watched this exact training video and were then instructed to immediately attempt the moonwalk themselves. This was indeed true (see next paragraph). Our participants then watched a video of one of these real attempts and reported their reactions via the following 5-item Criticism scale, comprising our key dependent variables (items were shown in randomized order and each was rated from 1 = *don't agree at all*, 10 = *completely agree*): "This attempt was worse than I thought it'd be"; "This participant might be a bit incompetent"; "This participant should've done better"; "I'd avoid recruiting this participant on my own dance team"; and "This attempt deserves some criticism."

As noted, this set-up was real. The other purported study was Kardas and O'Brien's (2018) Experiment 3, whereby those authors showed the training video to participants either 1× or 20× (as we did) and then recorded their attempts—and found that overexposure improved *predicted* dancing abilities but not *actual* dancing abilities (as judged by a blind outside panel). We thus showed *our* participants one of these recorded attempts and assessed their criticism of it. Like our motor-skills task from Experiments 1–3, this particular stimulus is ideal for the current research because Kardas and O'Brien (2018) established that it fits our targeted combination of "looks easy" plus "is hard"—thereby allowing us to assess our novel focus on *social* judgment. Because overexposure in this context has already been shown to inflate predicted abilities but not actual abilities, we can assume that any harshened criticism following from overexposure must be unfounded, at least on average, and that many of the judges showing this effect are overconfident.

We showed all participants the same (poor) attempt. To select it, we sorted through Kardas and O'Brien's (2018) open-access battery of 100 recorded attempts. In their study, each attempt was rated by hypothesis-blind judges from 1 = *pretty bad attempt*, to 10 = *pretty good attempt*. We sorted the 100 videos by average rating and sought to select one that was mostly bad—clearly a bad attempt, but one that still counts as an attempt to begin with (e.g., in some recordings, the dancer simply walks across the room without much moonwalking). To meet this criterion, we settled on a video in the 60th percentile with a mean rating of 3.31 (see our OSF for the video).

Thus, we tested whether overexposure to the expert video leads participants to become harsher judges of the same bad attempt. This uniquely tests our exposure hypothesis.

We also (independently) manipulated judges' knowledge about *this performer's exposure* to success prior to making their attempt. Based on random assignment, judges were informed that the performer had watched the expert video either once in full or 20 times repeatedly. Thus, we also tested whether participants become harsher to judge when they believe that the poor performer was overexposed to success beforehand. This uniquely tests our inference hypothesis.

7.1.2.1. Other variables. Finally, all participants reported their age, ethnicity, gender, and any technical difficulties (forced-choice: no vs. yes [explain]). They also completed an honesty check regarding whether they truly watched all videos (forced-choice: yes vs. no), plus 2 attention checks: one regarding how many times they themselves watched the expert video (forced-choice: 1× vs. 20×) and one regarding how many times the amateur performer watched the expert video (forced-choice: 1× vs. 20×).

7.2. Results and discussion

We combined the 5-item Criticism scale into a composite measure (α = 0.89). We conducted a Univariate GLM with Judge's Exposure (judges watched success: 1× vs. 20×) and Performer's Exposure (judges were informed that the performer had watched success: 1× vs. 20×) as between-subjects factors, with criticism of the attempt (1–10) as the dependent variable. Sensitivity analyses (α = 0.05 at 80% power) suggest that our sample size can detect an effect size of $\eta_p^2 = 0.008$ for a 2-way interaction.

7.2.1. Exposure to success harshens criticism of others' failure

There was a main effect of Judge's Exposure, $F(1, 995) = 6.23$, $p = .013$, $\eta_p{}^2 = 0.01$, such that watching success 20 times repeatedly indeed instigated harsher criticism of the same exact failed attempt as compared to watching success just once—supporting an exposure hypothesis. In addition, there was also a main effect of Performer's Exposure, $F(1, 995) = 40.17$, $p < .001$, $\eta_p{}^2 = 0.04$, such that judges gave harsher criticism when they believed the performer had first been repeatedly exposed to success—supporting an inference hypothesis. Moreover, there was *no* interaction, $F(1, 995) = 0.38$, $p = .537$, $\eta_p{}^2 < 0.01$—revealing support for two independent effects of observing success that similarly inflate people's criticism of failure (see Fig. 3).

That is, put in terms of pairwise comparisons: Judges who themselves were overexposed to success gave harsher criticism of the performer's failed attempt regardless of the performer's own exposure to success (when the performer watched 20 times: $M_{\text{JudgeWatch20x}} = 7.40$, $SD = 2.15$ vs. $M_{\text{JudgeWatch1x}} = 6.96$, SD = 2.05), $F(1, 995) = 4.88$, $p = .027$, $\eta_p{}^2 = 0.01$ ($d = 0.21$); when the performer watched once: $M_{\text{JudgeWatch20x}} = 6.41$, $SD = 2.48$ vs. $M_{\text{JudgeWatch1x}} = 6.15$, SD = 2.23), $F(1, 995) = 1.75$, $p = .186$, $\eta_p{}^2 = 0.002$ ($d = 0.11$)). This supports the effect of self-exposure. However, analyzed the other way, and also of interest: Performers who were overexposed to success were judged more harshly regardless of the judge's own exposure to success (when the judge watched 20 times: $M_{\text{PerformerWatch20x}} = 7.40$, $SD = 2.15$ vs. $M_{\text{PerformerWatch1x}} = 6.41$, SD = 2.48, $F(1, 995) = 24.27$, $p < .001$, $\eta_p{}^2 = 0.02$ ($d = 0.43$); when the judge watched once: $M_{\text{PerformerWatch20x}} = 6.96$, $SD = 2.05$ vs. $M_{\text{PerformerWatch1x}} = 6.15$, SD = 2.23, $F(1, 995) = 16.31$, $p < .001$, $\eta_p{}^2 = 0.02$ ($d = 0.38$). This supports a separate inference effect that operates alongside the self-exposure effect.

7.2.2. Other variables

Most participants passed the attention checks (Judge Exposure: 98.90%, 988 of 999; Performer Exposure: 91.59%, 915 of 999); passed the honesty check (98.30%, 982 of 999); and reported no technical difficulties (96.70%, 966 of 999). All key patterns hold when excluding participants based on these other variables (see Supplemental Material).

Experiment 4 extends further support for our hypotheses. These patterns emerged in a context in which research has established that

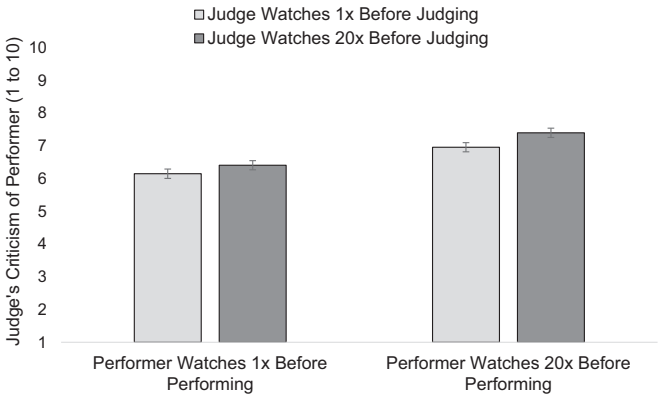


Fig. 3. Experiment 4: Criticism (5-item scale; scale means are plotted) as a function of the judge's exposure to success and the performer's exposure to success. Error bars show ±1 *SE*.

overexposure to success does not actually make any party better off, despite judges' beliefs—and, as revealed here, their criticism—otherwise.

Next, for all remaining experiments, we sought to dive deeper into our central exposure hypothesis. Experiments 5–7 continue to assess harshened criticism while also testing theory-driven boundaries: This effect should *not* simply reflect incidentally negative effects of repeated exposure (e.g., annoyance, tiredness) that could lead to harsher criticism (Experiment 5); but it *should* reflect judges' (mis)perceived learning from mere watching (i.e., coming to view the task as easier to perform than is warranted for the performer), such that the effect should be moderated by the observability of successful execution (Experiment 6) and whether people judge another person's poor attempt (which entails mere watching) or their own poor attempt (which entails watching *plus doing* (Experiment 7).

8. Experiment 5: Ruling out incidentally negative effects of repeated exposure

In Experiment 5, we sought to rule out an alternative explanation: Having to watch the successful video 20 times in a row may be a more negative experience than watching just once (e.g., by being more annoying or fatiguing), leading those participants to then "lash out" in the form of harsher criticism. This possibility cannot explain why we did not also observe more negative reactions on our measures of predicted performance; nonetheless, definitively ruling it out would more clearly suggest that something like desensitization to the task's difficulty indeed better explains people's inflated criticism (as we have proposed). Experiment 5 served this goal.

In addition, in this experiment (and in Experiment 8), we do not mention any information about the performer's exposure to success and instead only manipulate the judge's exposure. This change helps center the study on our exposure hypothesis while also expanding ecological validity. In everyday life, presumably it is more common for overexposed viewers to then come across countless other stimuli (e.g., flipping to other content; coming offline to then observe other real-world events) for which they have no knowledge of the person's training history and underlying psychology. This change in our study design better captures these dynamics, and so would suggest that the criticism effect (if it holds) may be quite prevalent in everyday life.

8.1. Method

8.1.1. Participants

We requested 500 participants from Cloud Research, yielding 502 (M_{age} = 40.23, SD_{age} = 12.68; 47.17% female; 23.06% non-White) who completed the study for $1.00. An additional 26 participants started the survey but did not finish it (12 in the low exposure condition, 9 in the high exposure condition, and 5 who dropped out before being assigned to a condition).

8.1.2. Procedure

Participants were randomly assigned to a 2 (Judge's Exposure to Success: low vs. high) between-subjects design.

Procedures were very similar to Experiment 4—whereby all participants first watched the same expert moonwalker (either once or 20 times) and then all evaluated the same poor attempt of the laboratory participant—except we made 2 key changes for current purposes (in addition to eliminating information about the performer's exposure, as noted above).

First and most important: Across conditions, we equated the length of the expert video that participants first watched before judging the performer. Recall from Experiment 4 that a single loop of the expert moonwalker ran about 5 s, with participants being randomly assigned to watch this loop once in full (thus watching for about 5 s in total) or 20 times repeatedly (thus watching for about 2 min in total). Here, we exactly replicated this design for participants who were assigned to the high-exposure condition: They watched that same ~2 min video, involving 20 consecutive ~5 s loops of the expert moonwalker. Unique to the current study, however, participants assigned to the low-exposure condition *also* saw a ~ 2 min video, involving 20 consecutive ~5 s loops—thus identically matching the video length of the high-exposure condition. For these participants, the first 19 loops showed a ~ 5 s clip of an unrelated custom-made video of moving plain circles; then, for their twentieth loop, they saw the expert moonwalker clip (see OSF for these videos). In this way, we isolated exposure *to success* in the same way as in all other studies—with some participants gaining low exposure (watching the ~5 s expert clip once in full) and others gaining high exposure (watching this same ~5 s expert clip 20 times in a row)—while equating overall video length.

Second: After watching this ~2 min video (which contained low or high exposure to the expert), all participants then watched the same poor performance attempt from Experiment 4 and indicated their criticism via the same 5-item scale from Experiment 4. Following these measures and our demographic questions, we added several questions to measure participants' psychological state when they had watched the initial ~2 min video beforehand. Participants rated 3 items (each from 1 = *not at all*, 7 = *extremely*): "how annoyed/frustrated" they had felt from watching, "how tired" they had felt from watching, and "how long" the 2 min had felt.

As preregistered, we hypothesized to find null effects on these items, given that we intentionally designed the videos to be matched on overall length and repetitiveness across conditions. In turn, we hypothesized that Experiment 4's spiked criticism in response to high exposure to *success* would nonetheless still emerge under these conditions. In contrast, if extensive watching makes people harsher critics simply because it causes a more negative experience—as opposed to, for example, desensitizing people to the skill's difficulty—then Experiment 4's results should *not* replicate in this study.

8.1.2.1. Other variables. Finally, all participants reported their age, ethnicity, gender, and any technical difficulties (forced-choice: no vs. yes [explain]). They also completed the same honesty check and "exposure" attention check from Experiment 4.

8.2. Results and discussion

We combined the 5-item Criticism scale into a composite measure (α = 0.89). We conducted an independent-samples *t*-test with Exposure (judges watched success: 1× vs. 20×) as the independent variable, and criticism of the attempt (1–10) as the dependent variable. We then ran this same analysis again using each of the 3 psychological-state measures as dependent variables instead. Sensitivity analyses (α = 0.05 at 80% power) suggest that our sample size can detect an effect size of d = 0.25 for testing differences between two independent means.

8.2.1. The criticism effect replicates…

First, the key criticism effect from Experiment 4 indeed replicated. Participants more harshly criticized the (poor) attempt of the laboratory participant if they first watched the expert clip loop 20 times in a row (M = 6.70, SD = 2.48) than if they first watched that expert clip just once (M = 6.16, SD = 2.19), $t(500) = 2.59$, $p = .010$, $d = 0.23$—even though participants in both conditions spent the same amount of time (~ 2 min) having to first watch the initial video.

8.2.2. …Despite opposite effects on participants' psychological state

Next, recall that we hypothesized to find a null effect between conditions in eliciting a negative state. Unexpectedly, we found a significant *opposite* effect[5]: Participants who watched a ~ 2 min video containing 20 loops of the expert moonwalker felt *less* "annoyed/frustrated" (M = 2.57, SD = 1.55), *less* "tired" (M = 2.71, SD = 1.63), and felt like those 2 min were *less* "long" (M = 3.83, SD = 1.71) than did participants who watched a ~ 2 min video containing 19 loops of moving plain circles plus 1 loop of the expert moonwalker ("annoyed/frustrated": M = 3.33, SD = 1.77, $t(500) = -5.13$, $p < .001$, $d = -0.46$; "tired": M = 3.19, SD = 1.88, $t(500) = -3.07$, $p = .002$, $d = -0.27$; "long": M = 4.35, SD = 1.77, $t(500) = -3.34$, $p < .001$, $d = -0.30$). These 3 items were also correlated as a scale ($\alpha = 0.88$), and analyses of this composite similarly revealed that high exposure to the expert moonwalker put participants in a *less negative* state (M = 3.03, SD = 1.45) as compared to low exposure to the expert moonwalker, (M = 3.62, SD = 1.62), $t(500) = -4.28$, $p < .001$, $d = -0.38$—yet it was these former participants who were *harsher* to judge the poor attempt. Finally (not preregistered), the criticism effect holds when including this composite measure as a control variable in the analyses, $F(1,499) = 8.44$, $p = .004$, $\eta_p{}^2 = 0.02$.

8.2.3. Other variables

Finally, most participants passed the attention check for their Exposure condition (80.08% passed, 402 of 502); passed the honesty check (98.01%, 492 of 502); and reported no technical difficulties (96.81%, 486 of 502). Consistent with our preregistration,[6] all key patterns hold when excluding participants based on these other variables (see Supplemental Material).

Experiment 5 suggests that the harshening effect of repeated exposure to success cannot be attributed to a more generally negative psychological state, ruling out this potential alternative explanation (and in turn increasing support for our proposed desensitization-related dynamics).

9. Experiment 6: Ruling in effects of success type

In Experiment 6, we sought to rule in one theory-driven boundary to the effect, via a moderation-based approach: the role of the *type* of successful information to which participants are repeatedly exposed. According to our theorizing, overexposure to success leads viewers to view the skill as not so difficult or complicated after all, thus driving their increased criticism of others who do it poorly. If this rationale is correct, then repeatedly exposing people to the same amount of success—but in ways that do not clearly convey the actual steps of execution (which presumably makes it less likely for people to grow accustomed to the skill's complexity)—should attenuate the effect. Experiment 6 tested this possibility. In doing so, note that we hold the *amount* of exposure to success constant—all repeatedly-exposed participants encounter the same exposure to the successful outcome itself—and so a moderation here would further highlight the specific role of exposure in dulling participants' beliefs about the task's difficulty (beyond repeated exposure to *any* depiction of success changing *any* success-related thinking).

9.1. Method

9.1.1. Participants

We requested 800 participants from Prolific Academic, yielding 800 (M_{age} = 36.04, SD_{age} = 13.74; 48.50% female; 21.75% non-White) who completed the study for $0.30 or $0.60 (depending on their exposure condition). An additional 73 participants started the survey but did not finish it (27 in the low exposure conditions, 46 in the high exposure conditions).

9.1.2. Procedure

Participants were randomly assigned to a 2 (Judge's Exposure to Success: low vs. high) × 2 (Type of Success: with steps vs. without steps) fully between-subjects design.

We returned to the motor-skills test from Experiments 1–3. Procedures were generally similar as before: All participants first watched a successful exemplar (either repeatedly or not), then criticized another participant's poor attempt. However, we made the following key changes.

First: Participants saw a recording of the actual failed attempt (rather than reading a description of their score, as in Experiment 3). We made the recording ourselves (i.e., we pretended to be an unskilled participant) and showed this same recording to all participants; the alleged player (actually us) does poorly, scoring 15%. Comprising our dependent variable, we used the fuller scale of criticism items that we had used in Experiments 4–5 (for the dancing stimulus), except that we included only the 3 items (of those original 5 items) that made the most sense in this context: "This attempt was worse than I thought it'd be"; "This participant should've done better"; and "This attempt deserves some criticism." Items were shown in randomized order and each was rated from 1 = *don't agree at all*, 10 = *completely agree*.

Second, and more critically: We also manipulated the *type of successful information* to which participants were (or were not) repeatedly exposed before they judged the failed attempt of the performer (whom we also described as having had the same exposure as the participant). Two of the four conditions—the "with steps" conditions—were identical to our typical set-up: Participants watched the successful expert video either once (low exposure; ~10 s) or 20 times in a row (high exposure; ~3 min). Here we hypothesized to again replicate the effect, such that high exposure to success harshens criticism of failure. For the other two conditions—the "without steps" conditions—we exposed participants to a mere *screenshot* of that successful expert video (specifically: it was the same still image of the final successful product), matched in timing such that it displayed for ~10 s (low exposure) or for ~3 min (high exposure). Here we uniquely hypothesized that the effect should be *attenuated*, because repeated exposure to success precludes these participants from being able to track (and thus perceive to "learn" and grow accustomed to) the actual steps of execution—despite these participants otherwise having the same exposure to the successful outcome itself as their "with steps" counterparts.

9.1.2.1. Other variables. Finally, all participants reported their age, ethnicity, gender, and any technical difficulties (forced-choice: no vs. yes [explain]). They also completed the same honesty check from prior studies, plus 2 attentions checks: one regarding their expert stimulus (forced-choice: video vs. image) and one regarding how long it lasted

[5] Though unexpected, in hindsight this opposite effect on psychological state may further highlight how our findings relate to overconfidence (as we have discussed and tested, e.g., in Experiment 3). A number of studies find that positive affect (which is conceptually similar to our measures here) can increase overconfidence (e.g., Ifcher & Zarghamee, 2014; Koellinger & Treffers, 2015; Prinz, Bergmann, & Wittwer, 2018; Sidi, Ackerman, & Erez, 2017). From this view, one may have predicted our "opposite" results here, whereby participants who were more (vs. less) exposed to success—who were led to feel less negative, and so perhaps also more overconfident—were those who gave harsher criticism of others' failures (in line with our overconfidence-criticism results elsewhere).

[6] For this experiment only, we preregistered to exclude participants based on these variables—which, if we did so, yields *stronger* support than what we report in the main text (criticism effect changes from $d = 0.23$ to $d = 0.29$; all psychological-state effects remain significantly opposite, $ds \leq$ -0.22; see Supplemental Material). We opted to report the full sample here for two reasons: (i) simply for stylistic consistency with all other experiments; (ii) in hindsight, we worried that excluding attention-check failures for these particular stimuli may unfairly exclude participants who answered correctly, but were instead thinking about overall exposure (rather than exposure to success specifically).

(forced-choice: 10 s vs. 3 min).

9.2. Results and discussion

We combined the 3-item Criticism scale into a composite measure (α = 0.88). We conducted a Univariate GLM with Judge's Exposure (judges watched success: 1× vs. 20×) and Type of Success (with steps vs. without steps) as between-subjects factors, with criticism of the attempt (1–10) as the dependent variable. Sensitivity analyses (α = 0.05 at 80% power) suggest that our sample size can detect an effect size of $\eta_p^2 = 0.01$ for a 2-way interaction.

9.2.1. Exposure to success harshens criticism of others' failure—only if it contains the steps

As hypothesized, there was the key Judge's Exposure × Type of Success interaction, $F(1, 796) = 8.21$, $p = .004$, $\eta_p^2 = 0.01$ (main effect of Judge's Exposure: $F(1, 796) = 4.21$, $p = .041$, $\eta_p^2 < 0.01$; main effect of Type of Success, $F(1, 796) = 21.85$, $p < .001$, $\eta_p^2 = 0.03$). See Fig. 4.

Pairwise comparisons reveal that, among the "with steps" conditions that simply reflect those from prior experiments (i.e., participants had low vs. high exposure to the full successful video), the criticism effect again replicated: Participants who watched success for 3 min (20×) gave harsher criticism of the other participant's poor attempt ($M = 6.58$, $SD = 2.54$) than participants who watched success for only 10 s (just 1×) ($M = 5.74$, $SD = 2.30$), $t(796) = 3.53$, $p < .001$, 95% CI = [0.38, 1.31], $d = 0.35$. Critically, however, this effect was attenuated among the "without steps" conditions: Regardless of whether participants had watched success for 3 min ($M = 5.29$, $SD = 2.54$) or 10 s ($M = 5.43$, $SD = 2.33$), they gave *similarly low* criticism, $t(796) = -0.57$, $p = .571$, 95% CI = [−0.62, 0.34], $d = -0.06$.

9.2.2. Other variables

Most participants passed the attention checks (Judge Exposure: 95.38%, 763 of 800; Success Type: 92.75%, 742 of 800); passed the honesty check (97.75%, 782 of 800); and reported no technical difficulties (96.75%, 774 of 800). All key patterns hold when excluding participants based on these other variables (see Supplemental Material).

Experiment 6 adds theory-bearing moderation of our findings thus far. The harshening effect of repeated exposure to success at least partly depends on the *kind* of success judges are exposed to (holding its *amount* constant): Only when judges could track successful execution did they raise their criticism of others' failures, suggesting that the effect may require more dulling experiential exposure (beyond repeated exposure to *any* "mere thinking/beliefs" about success).

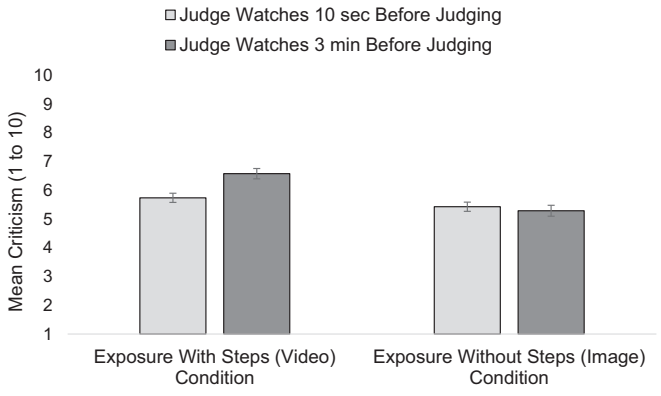


Fig. 4. Experiment 6: Criticism (3-item scale; scale means are plotted) as a function of the judge's exposure to success and the type of success within this exposure. Error bars show ±1 *SE*.

10. Experiment 7: Ruling in effects of self/other judgment

In Experiment 7, we sought to rule in another such boundary: whether repeated exposure to success differentially affects judges' criticism of *others*' failed attempts vs. *their own* failed attempts. If our rationale is correct that merely watching success dulls viewers to the complexity of a task (thereby harshening them to failure), then this effect should be most pronounced when people judge *others'* failed attempts—as, in such cases, mere watching is the sole psychological input into their judgment. In contrast, when people judge *their own* failed attempts, they have also gained input from *actually doing it*—which should better highlight the task's complexity and so temper their criticism of doing it poorly (despite also having repeatedly watched success beforehand). Experiment 7 tested this possibility. Note also that, while there could be incidental main effects between social criticism and self-criticism (e.g., perhaps people are kinder to others for social-desirability reasons; perhaps people are kinder to themselves due to self-enhancement), our hypothesis entails *within*-target differences from overexposure (i.e., the interaction: whether high vs. low exposure fosters more criticism *within* other-judgment than *within* self-judgment).

10.1. Method

10.1.1. Participants

First, we recruited "judge-self" participants. We requested 300 participants from Prolific Academic, yielding 300 ($M_{age} = 34.83$, $SD_{age} = 13.40$; 79.00% female; 30.33% non-White) who completed the study for $0.50 or $0.90 (depending on their exposure condition). An additional 111 participants started the survey but did not finish it (48 in the low exposure condition, 63 in the high exposure condition). The "judge-self" participants attempted a target task and reported their criticism of their own attempt (hence our use of the phrase "judge-self").

Next, we recruited 300 "judge-other" participants from the same population ($M_{age} = 38.33$, $SD_{age} = 14.22$; 73.00% female; 19.33% non-White; $0.30 or $0.70 [depending on their exposure condition]; 25 others started the survey but did not finish, 9 low/16 high), who were randomly yoked to view one of these 300 attempts (from our self-performers above) and report their criticism of this other participant's attempt (hence, "judge-other").

10.1.2. Procedure

Participants were randomly assigned to a 2 (Judge's Exposure to Success, between-subjects: low vs. high) × 2 (Target, yoked within-subjects such that the same attempt is judged twice, by each of two judges: judge-self vs. judge-other) mixed-factor design.

10.1.2.1. Judge-self conditions. For further generalizability, we assessed a new task: sketching a drawing. To begin, all "judge-self" participants were informed that they would learn about this drawing task, answer some questions about it, and then actually attempt it. They completed a practice round using the same drawing technology to be used in the main task, involving writing the word "OK" in an empty digital canvas using their finger on their trackpad. Their drawing strokes were processed and observable in real time as if they were drawing with pen and paper.

For the main task, these participants were informed that they would copy an image of a human hand as accurately as possible, using the same technology as in the practice round, in a time limit of 30 s (i.e., after 30 s, the survey would automatically proceed to a new screen and their drawing—in whatever state it was in after 30 s—would be automatically uploaded to our data repository). This task was inspired by a viral Internet video that instructed viewers how to easily draw a realistic human hand in a few quick steps—which went viral as people posted their comically mangled attempts upon trying to follow the video (Brooks, 2019). We used the actual viral video and hand image as our

stimuli (see Appendix A and OSF). Participants were not informed about the story or the background of the stimulus. Again, as for our other stimuli (motor-skills stimulus; dancing stimulus), this particular stimulus is ideal for the current research because it fits our targeted combination of "looks easy" plus "is hard."

First, however, these participants watched an instructional video (the video from the news story) showing how to expertly draw the hand. Based on random assignment, some participants watched this video 1× in full (about 10 s) while others watched the video 20× repeatedly (about 4 min). They were instructed to passively watch the video (as in all our experiments).

After watching, they then made their attempt at actually drawing the hand, exactly as described (see Appendix A for examples of these actual attempts). They then completed a 5-item Criticism scale of their own attempt, comprising our key dependent variables (items were shown in randomized order and each was rated from 1 = *don't agree at all*, 10 = *completely agree*). The items were new to this study (again, simply for generalizability): "I have little reason for performing this way"; "It's my fault for performing this way"; "It seems like I deserve some slack"; "My performance is understandable"; and "It seems like I was set up to fail."

10.1.2.2. Judge-other conditions. Next, we recruited our "judge-other" participants to also judge these same attempts—except from the perspective of observers rather than the actors themselves.

We described the "self-judge" conditions to them and explained that they would evaluate one of these participants' attempted drawings, selected at random (without replacement[7]). First, however, these "judge-other" participants also watched the instructional video showing how to expertly draw the hand. We yoked these participants such that they watched this video the same number of times as their drawer had done before making their attempt. In other words, criticism was effectively measured in within-subjects fashion: For each of our 300 (poorly) attempted drawings, two participants judged it after having identical exposures to success, with one participant being the drawer themselves ("judge-self" participants) and the other participant being an outside observer ("judge-other" participants). These "judge-other" participants rated the drawing via the same 5-item Criticism scale, simply adapted for target (e.g., using phrasings such as "They have little reason…" as opposed to "I have little reason").

10.1.2.3. Other variables (both sets of participants). Finally, all participants reported their age, ethnicity, gender, and any technical difficulties (forced-choice: no vs. yes [explain]). They also completed the same honesty check from prior studies, plus an attention check regarding how many times they watched the expert instructional video (forced-choice: 1× vs. 20×).

We also included an exploratory item for "self-judge" participants only: We directly asked them (after they completed the drawing) whether they thought they had been (over)confident in the drawing task (forced-choice: I was underconfident vs. I was accurately confident vs. I was overconfident).

10.2. Results and discussion

We combined the 5-item Criticism scale into a composite measure, reverse-coding where needed such that higher scores indicate higher criticism (α = 0.76). We conducted a Repeated-Measures GLM with Judge's Exposure (judges watched success: 1× vs. 20×) as a between-subjects factor and Target (self-judge vs. other-judge) as a within-subjects factor, with criticism of the attempt (1–10) as the dependent variable. Sensitivity analyses (using correlations among the repeated measures; α = 0.05 at 80% power) suggest that our sample size can detect an effect size of $\eta_p^2 = 0.006$ for a 2-way interaction.

10.2.1. Exposure to success harshens criticism of others' failure—more than one's own failure

As hypothesized, there was the key Judge's Exposure × Target interaction, $F(1,298) = 6.25$, $p = .013$, $\eta_p^2 = 0.02$ (main effect of Judge's Exposure: $F(1,298) = 35.41$, $p < .001$, $\eta_p^2 = 0.11$; main effect of Target, $F(1,298) = 9.21$, $p = .003$, $\eta_p^2 = 0.03$). See Fig. 5.

Pairwise comparisons reveal that, among the "other-judge" conditions (a simple replication of prior experiments), the criticism effect again replicated: Participants who watched success 20× gave harsher criticism of the other participant's poor attempt ($M = 5.23$, $SD = 2.15$) than participants who watched success just 1× ($M = 3.98$, $SD = 1.82$), $t(298) = -5.43$, $p < .001$, 95% CI = [−1.70, −0.79], $d = -0.70$. Critically, however, this effect was attenuated among the "self-judge" conditions: Although participants who watched success 20× also gave harsher criticism to this same poor attempt when it was *their own* poor attempt ($M = 4.47$, $SD = 1.60$) as compared to after they watched success just 1× ($M = 3.91$, $SD = 1.43$), $t(298) = -3.23$, $p = .001$, 95% CI = [−0.91, −0.22], $d = -0.32$, the 2-way interaction indicates that this relative boost in *self*-criticism was significantly smaller.

10.2.2. Other variables

Most participants passed the attention check (99.67%, 598 of 600); passed the honesty check (98.67%, 592 of 600); and reported no technical difficulties (97.67%, 586 of 600). All key patterns hold when excluding participants based on these other variables (see Supplemental Material). Regarding our exploratory overconfidence item ("self-judge" only), more Watch 20× participants explicitly reported that they had been overconfident (46.31% overconfident, 69 of 149; 8.05% underconfident, 12 of 149; 45.64% accurately confident, 68 of 149) relative to Watch 1× participants (30.46% overconfident, 46 of 151; 15.89% underconfident, 24 of 151; 53.64% accurately confident, 81 of 151), $\chi^2(2, N = 300) = 9.72$, $p = .008$, $w = 0.18$—which echoes our prior results about overexposure generally inflating (over)confidence.

Experiment 7 adds further theory-bearing moderation evidence: Overexposed judges were especially harsh to others' vs. *their own* attempt, despite holding exposure and the quality of that attempt constant. Direct experience may help people better appreciate the task's underlying complexity, which is otherwise dulled by repeated exposure.

11. Experiment 8: Consequences and debiasing

Finally, Experiment 8 had two goals. First, we tested behavioral

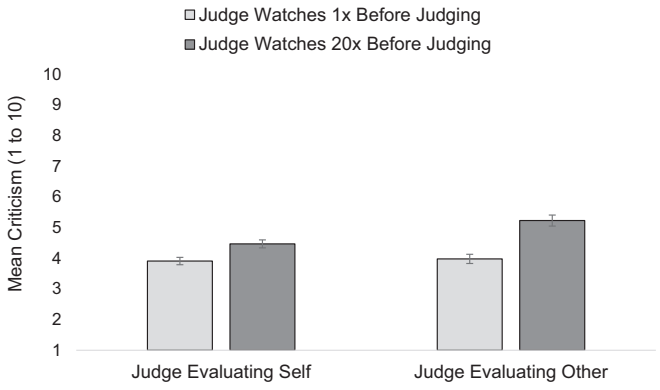


Fig. 5. Experiment 7: Criticism (5-item scale; scale means are plotted) as a function of the judge's exposure to success and their target of judgment (self vs. other). Error bars show ±1 *SE*.

[7] We preregistered that, if these random draws ended up showing the same attempt to more than one "judge-other" participant, we would analyze only the first rater (and then recruit more participants as needed to rate whatever was skipped). This occurred for 7.41% of the attempts (24 of 324). All ratings are retained in the data file (see OSF).

consequences: People who are repeatedly exposed to success may issue *smaller rewards* to others who attempt a task and perform poorly at it. Second, we tested a potential debiasing strategy derived from our proposed framework (and following from our self/other findings in Experiment 7): *Direct experience* may motivate judges to be kinder to struggling others (here, in the form of raising their initially-smaller rewards), if they judge others *after* actually attempting the task on their own—despite their repeated exposure to success beforehand. Indeed, direct experience with difficulty has been shown to close the so-called "empathy gaps" (Van Boven et al., 2013) like those that may emerge here. Experiment 8 tested these possibilities.

11.1. Method

11.1.1. Participants

We requested 500 participants from Cloud Research, yielding 484 (M_{age} = 34.85, SD_{age} = 9.81; 36.36% female; 38.22% non-White) who completed the study for $1.50. An additional 88 participants started the survey but did not finish it (12 in the low exposure condition, 19 in the high exposure condition, and 57 who dropped out before being assigned to a condition).

11.1.2. Procedure

Participants were randomly assigned to a 2 (Judge's Exposure to Success, between-subjects: low vs. high) × 2 (Phase, within-subjects: Reward before vs. after making own attempt) mixed-factor design.

Procedures were generally similar to Experiment 7, using the same "drawing task" stimulus in the same way (one minor change was that here we gave all participants a second practice round before attempting the main drawing; in addition to writing "OK," they were instructed to copy a "smiley face" image).

The substantive changes were in our design: First, all participants were exposed to the expert video (either 1× or 20×). Then, they saw the same single image of another participant's alleged (poor) attempt (see Appendix B), along with a randomized participant ID number (in reality, this attempt was created by us, like in Experiment 6). In this experiment, as in Experiment 5, they were not given any information about this participant's own exposure. To assess our key dependent variable, participants then indicated how much of a bonus payment this participant should receive based on the quality of their attempt, ranging from $0.00 to $0.10 in 1-cent increments. We tested whether overexposure to the expert video might lead participants to issue a smaller reward to the same poor attempt made by another participant.

Finally, we also assessed debiasing. All participants then actually attempted drawing the hand themselves, exactly as described (see Appendix B for examples of these actual attempts). After making their own attempt, they were re-shown the other participant's attempt and again indicated how much of a bonus payment this participant should receive for it (they were free to re-report their original response if they wanted). Thus, all told: We tested whether participants raise their initially-smaller rewards after having experienced this (difficult) task themselves.

11.1.2.1. Other variables. Finally, all participants reported their age, ethnicity, gender, and any technical difficulties (forced-choice: no vs. yes [explain]) and whether they had successfully used a trackpad throughout the study (forced-choice: no vs. yes). They were also informed that we took our stimuli from the viral Internet story as described and were asked whether they had ever seen the hand-drawing video (forced-choice: no vs. yes). They also completed an attention check regarding whether they had been informed about the specific training history of the other participant whose drawing they judged (forced-choice: no vs. yes).

Lastly, in addition to our key bonus-reward measures, we also sought to assess participants' beliefs about their own attempt (consistent with our prior experiments). We did so in two ways:

First, we assessed judges' explicit beliefs. After issuing their initial reward (and before attempting the task), all participants reported whether they believed their attempt would be better than this other participant's attempt (forced-choice: will do better, will do as bad/good, will do worse)—allowing us to assess whether overexposed participants are more likely to believe they will outperform the other participant. We then re-asked this question after participants made their attempt (forced-choice: did better, did as bad/good, did worse)—allowing us to assess whether overexposed participants reduce these inflated expectations after attempting the skill themselves.

Second, we sought outside confirmation by recruiting unique participants from the same population (on Cloud Research; N = 484; M_{age} = 37.85, SD_{age} = 11.87; 49.38% female; 26.86% non-White; study pay $0.10; 4 started but did not finish the survey) to compare the drawings as blind judges. They were yoked to rate one attempt, selected at random without replacement (thus, we recruited 484 judges to pair with our battery of 484 attempts). Each judge was shown the expert target image (from the video), followed by the alleged participant's attempt and an actual participant's attempt (presented in randomized order, each labeled with a randomized ID number). Their task was to choose which attempt was better (plus a third option to indicate the two attempts were equally bad/good). Thus, we tested whether our real participants did worse than the comparison participant (despite their predictions otherwise) even as decreed by people who were purely evaluating the output alone.

After making their choice, these outside judges reported their age, ethnicity, gender, as well as any technical difficulties (forced-choice: no vs. yes [explain]) and whether they had ever seen the viral hand-drawing video (forced-choice: no vs. yes). They also completed an attention check regarding what their task was about (forced-choice: evaluating hands vs. cars vs. animals).

11.2. Results and discussion

For our main analyses, we conducted a Repeated-Measures GLM with Exposure (judges watched success: 1× vs. 20×) as a between-subjects factor and Phase (judges issued reward before vs. after attempting the task themselves) as a within-subjects factor, with issued-reward (bonus amount from $0.00 to $0.10) as the dependent variable. Sensitivity analyses (using correlations among the repeated measures; $\alpha = 0.05$ at 80% power) suggest that our sample size can detect an effect size of $\eta_p^2 = 0.003$ for a 2-way interaction.

11.2.1. Exposure to success harshens behavioral treatment of others' failure—but personal experience with the task combats this effect

There was a main effect of Exposure, $F(1, 482) = 4.02$, $p = .045$, $\eta_p^2 = 0.01$, such that participants who watched success 20 times repeatedly issued a lower bonus as compared to participants who watched success just once (there was also an incidental main effect of Phase, $F(1, 482) = 80.17$, $p < .001$, $\eta_p^2 = 0.14$). Critically, however—and as hypothesized—this effect was qualified by an Exposure × Phase interaction, $F(1, 482) = 11.83$, $p < .001$, $\eta_p^2 = 0.02$: Watching success indeed differentially affected the bonus that participants issued before vs. after they made their own attempts (see Fig. 6).

Pairwise comparisons reveal that *before* attempting the task themselves, participants issued lower bonuses to the same failed attempt by another participant when they themselves first watched success 20 times repeatedly (M = 5.84 cents, SD = 3.29) vs. just once (M = 6.79 cents, SD = 3.29), $F(1, 482) = 10.05$, $p = .002$, $\eta_p{}^2 = 0.02$ ($d = -0.29$). However, this effect was attenuated *after* participants attempted the task: They issued similar bonuses regardless of whether they had first watched success 20 times repeatedly (M = 7.36 cents, SD = 3.01) vs. just once (M = 7.47 cents, SD = 3.16), $F(1, 482) = 0.14$, $p = .705$, $\eta_p{}^2 < 0.001$ ($d = -0.04$).

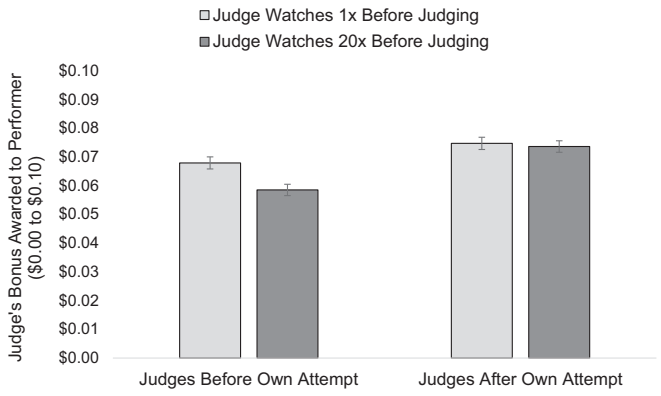


Fig. 6. Experiment 8: Mean amount-of-bonus issued by judges as a function of their own exposure to success, before and after their own actual attempt at the task. Error bars show ±1 *SE*.

11.2.2. Other variables

Most participants in the drawing study passed the attention check (66.32%, 321 of 484); reported no technical difficulties (96.90%, 469 of 484); used a track-point or trackpad (95.25%, 461 of 484); and had never seen the hand-drawing video (72.73%, 352 of 484). Likewise, among our separate sample of blind third-party judges, most passed their own attention check (99.79%, 483 of 484); reported no technical difficulties (99.59%, 482 of 484); and had never seen the hand-drawing video (95.04%, 460 of 484). All key patterns hold when excluding participants, from either of these study phases, based on these other variables (see Supplemental Material).

In addition, we found parallel patterns on our "task belief" measures. Before attempting the task, judges were more likely to believe they would outperform the other participant after they themselves watched success 20× (46.64% will do better, 111 of 238; 44.12% will do as bad/good, 105 of 238; 9.24% will do worse, 22 of 238) vs. 1× (39.43% will do better, 97 of 246; 43.90% will do as bad/good, 108 of 246: 16.67% will do worse, 41 of 246), $\chi^2(2, N = 484) = 6.58$, $p = .037$. And yet, after attempting the task, judges who watched success 20× were just as doubtful that they outperformed the other participant (18.07% thought they did better, 43 of 238; 21.43% thought they did as bad/good, 51 of 238: 60.50% thought they did worse, 144 of 238) as those who watched success 1× (19.11% thought they did better, 47 of 246; 20.33% thought they did as bad/good, 50 of 246: 60.57% thought they did worse, 149 of 246), $\chi^2(2, N = 484) = 0.14$, $p = .932$. When re-running our analyses on only those participants who seemed to be overconfident about being able to outperform the other participant (i.e., who predicted doing better, but did as bad/good or worse; $N = 264$), the key effects hold (main effect of Exposure, $F(1, 260) = 0.001$, $p = .977$, $\eta_p^2 < 0.001$; main effect of Phase: $F(1, 260) = 48.46$, $p < .001$, $\eta_p^2 = 0.27$; Exposure × Phase interaction, $F(1, 260) = 4.76$, $p = .031$, $\eta_p^2 = 0.04$).

The same pattern emerged among blind third-party judges: The majority of these judges found the attempts to be *worse* than the comparison attempt (11.16% thought the original judge did better, 54 of 484; 11.98% thought the original judge did as bad/good, 58 of 484; 76.86% thought the original judge did worse, 372 of 484), $\chi^2(2, N = 484) = 412.68$, $p < .001$. This result held regardless of whether blind third-party judges evaluated those who watched success 20× (10.92% thought the original judge did better, 26 of 238; 10.50% thought the original judge did as bad/good, 25 of 238; 78.57% thought the original judge did worse, 187 of 238), $\chi^2(2, N = 238) = 219.18$, $p < .001$) or evaluated those who watched success 1× (11.38% thought the original judge did better, 28 of 246; 13.41% thought the original judge did as bad/good, 33 of 246; 75.20% thought the original judge did worse, 185 of 246), $\chi^2(2, N = 246) = 194.22$, $p < .001$).

Experiment 8 extends our findings in two further ways. First, these results show that repeated exposure to success can have tangible consequences beyond "mere" expectations and perceptions, here in terms of people issuing smaller rewards for someone else's poor attempt. Second, getting people to attempt the task themselves combats this effect—consistent with our theorizing. Despite participants being fully informed about the task before making their attempt, gaining direct hands-on experience may be a surer route to developing a deeper appreciation for a task's underlying complexities and debiasing people's dulled beliefs (Van Boven et al., 2013).

12. General discussion

Our successes are often made in the public eye—and alas, so too are our failures. Eight experiments reveal that observing the former can distort people's responses to the latter. Across numerous contexts, measures, and stimuli, we uniquely found that repeated exposure to success inflated people's expectations of others' performances and led them to more harshly criticize and treat others for their failed attempts—even when the harsh critics would have done just as badly (and worse than what they confidently imagined) if they were prompted to try themselves. This effect emerged independently from incidentally negative states caused by repeated exposure; instead, it depended on (mis)perceptions of learning from repeated exposure to success. In turn, participants treated others' stumbles more kindly after they were exposed to the task themselves.

12.1. Insights and implications

First, these findings build on and significantly advance Kardas and O'Brien's (2018) "illusion of skill acquisition." Those authors found that repeated exposure to success inflates people's self-assessments of their own performance abilities, which they interpreted strictly as a self-overconfidence effect—thus implying that it should not emerge for other-oriented judgment ("I now think that *I* can do this impressive skill—but not you"). Our findings replicate this effect while also uniquely and centrally focusing on other-oriented outcomes: We find that repeated exposure to success also inflates people's *social judgments* of *others*' performance abilities, leading to unfairly harshened criticism of others' poor attempts. By doing so, we reveal the need for a re-interpretation of their findings: The "illusion of skill acquisition" is more general than mere self-perceptions or self-overconfidence per se, and instead reflects a broader changing appraisal (Wondra & Ellsworth, 2015) of the skill itself ("This isn't so impressive after all; *anyone* could do it"). Moreover, Experiment 7 uniquely finds that inflated criticism is even *more* pronounced for evaluations of others than of one-self—highlighting further unique outcomes that would not have been predicted by Kardas and O'Brien (2018). Our theorizing can explain this finding, as it fits with our proposal that firsthand experience with the task can serve to attenuate the inflated criticism from overexposure. When evaluating others, however, one lacks this direct experience, meaning that overexposure is especially impactful on criticism of *others'* attempts. Put another way: The current research uniquely reveals that repeated exposure to success may affect social judgment *even more* than self-judgment; the "illusion of skill acquisition" may be more valuably seen as a social effect applied to social settings rather than as a self-oriented issue.

Second, we nonetheless do find some evidence that our effects are especially pronounced among truly-overconfident participants. Thus, a second contribution of our unique focus on other-oriented outcomes is to research on overconfidence more generally, which traditionally focuses on self-oriented outcomes (e.g., estimating one's own knowledge: Kruger & Dunning, 1999; Rozenblit & Keil, 2002). Related research similarly documents how acquiring knowledge can disrupt one's ability to imagine ever not having it (e.g., the "curse of knowledge" in recalling one's own past naïve state: Camerer et al., 1989; the "illusion of competence" in forecasting one's own future naïve state: Koriat & Bjork, 2005). Few (if any) such studies, however, have examined how *self-*

overconfidence may bleed into one's judgments of *others*, such as criticizing others' failures. Problems of overconfidence may spread beyond the overconfident judge alone.

Third, by uniquely focusing on other-oriented outcomes, the current research also bears on additional other-oriented literatures. For example, while prior research on social judgment often suggests that people criticize others for motivated reasons (see Introduction), we highlight that criticizing *any*thing may start with (or, at least, be additionally influenced by) people's basic beliefs about the task or behavior itself. To date, for example, psychological attempts to explain why people punish others come in the form of intricate motivated models involving game theory and identity signaling (for a review, see Cushman, Sarin, & Ho, 2021). Our findings suggest a simpler answer: If something seems easy to do, those who fail to do it will seem more negatively noteworthy by comparison. From this perspective, understanding what affects people's beliefs about ease-of-execution (like exposure to success) may be just as critical as understanding more explicit motivations; existing research may *downplay* the pervasiveness of criticism in everyday life. Even merely believing that others have acquired extensive experience—but "experience" of the wrong kind (i.e., mere watching, without doing)—may sometimes widen empathy gaps rather than close them. More research in general should assess potential costs to lay beliefs of "practice makes perfect," which are popularly depicted in a uniformly positive light (Duckworth & Gross, 2014; Dweck, 2008). The same people who believe that "practice makes perfect" may (unfairly) condemn others for stumbling, especially if they do not distinguish between kinds of "practice."

Finally, our findings also raise practical implications in today's informational landscape, whereby (over)exposure to success is likely widespread (e.g., via YouTube, social media, and on-demand services). In principle, access to success should promote diverse learning. However, such widespread access to success may not only be inflating people's own self-overconfidence, but also simultaneously *deflating* people's understanding of others who try (and fail). One need only briefly skim any online forum to notice the pervasive criticism readily shared by users—many of whom, as our findings suggest, could do no better themselves if they were prompted to actually try. As such, our findings suggest a causal clue—i.e., (over)exposure to success—into rising narcissism (Twenge, Konrath, Foster, Campbell, & Bushman, 2008) and falling empathy (Konrath, O'Brien, & Hsing, 2011). Moreover, note that our findings reflect *mere* watching (i.e., the idea that observational learning may backfire if not paired with hands-on experience)—and it is this combination that may be growing over time in today's information age (e.g., getting to observe others perform feats and enjoy experiences from across the globe, with zero access to those things oneself). Experiment 8 hints at debiasing tactics that are scalable in everyday life: To promote better understandings and fairer treatment, feedback outlets could require people to *fully experience* the entity before critiquing others—both in online settings (e.g., word-of-mouth review websites can validate that reviewers have interacted with the entity before allowing them to comment) and in offline settings (e.g., in-person seminars can begin by requiring attendees to complete a practice run). By the same logic, our findings join others (Klein & O'Brien, 2017; O'Brien, 2022) warning that people should more closely consider the wisdom of advertising their failures to others, which some recent research advocates people do in order to build connections with others (Brooks et al., 2019; Steinmetz, 2018); receptive audiences may quickly turn damning depending on their previous exposure to success.

12.2. Future directions

One outstanding question entails constraints on generality. We tested immediate and repetitive exposure, for conceptual precision—but are people actually exposed in this way in everyday life? On the one hand, a growing number of real-world examples suggest yes, from the YouTube statistics cited in the Introduction (with the tens of millions of instructional videos that users likely watch repeatedly as they try to learn), to the short video clips that repeat themselves during the playing of a song on Spotify (which, according the company, boosts user engagement by up to 145%: Spotify, 2022), to whole platforms like Streamable and TikTok that are expressly designed to loop short video content (as one article reviews: "Generation Z are recorded to have an attention span of just eight seconds, meaning *shorter*, *recurring* content is key for tapping into the younger market": Dolan, 2020, italics added). TikTok's algorithm for selecting which videos to display on a user's front page is suspected to precisely select videos that have high *repeat viewing* potential ("They have to watch it, all the way, and then want to watch it again": Vala Marketing, 2020)—with *billions* of such videos being watched on users' front pages *every day* (Statista, 2021). Thus, our research may help speak to growing (but under-studied) consumption dynamics. On the other hand, future research should still test less immediate and less repetitive exposures, as these are presumably also common, yet should lead any exposure effects to decay.

Relatedly, and extending this last point, future research should assess further boundary conditions. What are the psychological factors that should amplify vs. attenuate the basic effect? Experiments 5–8 highlight a few. More broadly, we expect our findings to hold when (i) repeated exposure conveys enough information about the steps of execution (promoting desensitization dynamics), but (ii) repeated exposure (alone) does not actually improve performance. By design, we therefore assessed tasks that fit this combination of *looking easy* but *being hard*—which represent a prevalent and consequential domain (Kardas & O'Brien, 2018)—and in turn we assessed moderation by manipulating participants' knowledge of these task features. In any case, this combination is just a subset of what exists in everyday life. Some tasks, for example, might look increasingly hard or intimidating across repeated exposure to an expert—but, to one's relief, prove relatively straightforward upon finally doing them oneself (e.g., a simple public-speaking task). We suspect that such tasks are at least partly reflective of our first precondition, such that people cannot easily track the steps of execution from mere watching alone. Here, we predict the opposite effect such that watching many (vs. few) times should *increase* empathy and *decrease* criticism of others' failed attempts, relative to how viewers respond after they attempt the task. Direct experience may "debias" people in how they treat first-timers—meaning, in this case, callous them. These ideas suggest related boundaries, like individual differences across judges (e.g., baseline differences in viewers' expertise in knowing what to look for to begin with).

Still another related question entails other underlying mechanisms. For example, perhaps *any* repeated exposure to *any* success elicits effects that contribute to increased criticism and the like—such as by raising people's inferred base rates of success, or simply by giving people more time to imagine and get excited about the mere idea of success. First, this former possibility—while interesting—is unlikely in our studies, as we exposed participants to the same clip of the same successful performer (whereas watching many different successful performers seems more relevant for raising inferred base rates of success). Moreover, if our effect purely reflected such a difference, then we should have observed a baseline self-other difference in Experiment 7, to the extent that people infer a higher base rate of success for others vs. for themselves (e.g., Davidai & Gilovich, 2016)—yet we did not find such a difference (Fig. 5, light grey bars). Second, this latter possibility is also unlikely, given Experiment 6. Either way, another way to gain traction on this question is to ask: What happens when people are repeatedly exposed to *failed* exemplars? Our theorizing predicts that it should depend on the extent to which those failed attempts clearly (vs. unclearly) convey the correct steps (e.g., mis-stepping a dance move in a way that does [vs. does not] imply what viewers should have done instead). Future research should test these ideas.

Until these ideas are tested, the current research warrants a closer look at the value of observing success. "Everyone's a critic," it seems, which apparently holds true even among the (naïvely) unskilled.

Although genuine experts can provide invaluable feedback for self-improvement, mere observers may readily perceive themselves so—leaving none the wiser.

Open practices

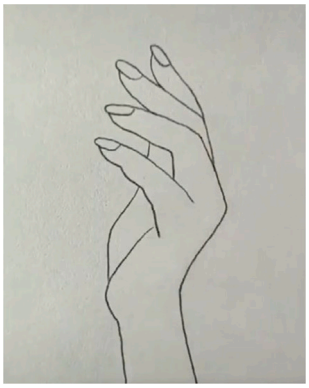
All data, materials, and preregistrations are on the Open Science Framework (OSF): https://osf.io/m623y/.

Author note

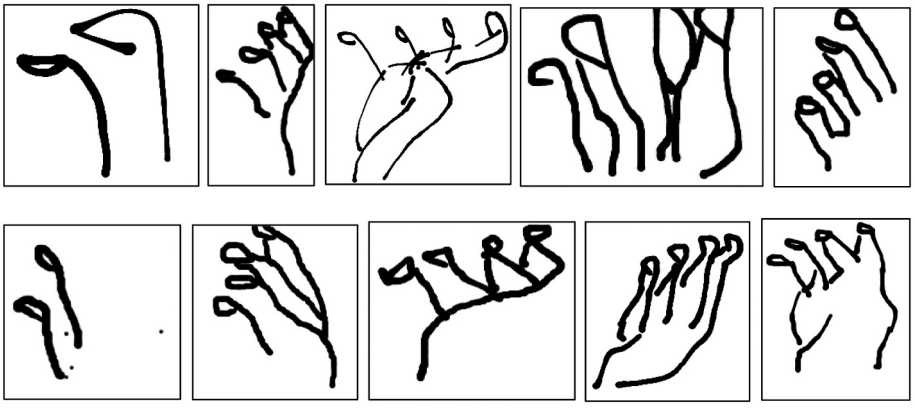
All data, materials, and preregistrations are on the Open Science Framework (OSF): https://osf.io/m623y/.

Appendix A. Drawings from Experiment 7

Expert drawing (to copy):



Actual attempts of participants, 10 randomly-chosen examples (see OSF for all 300 attempts):

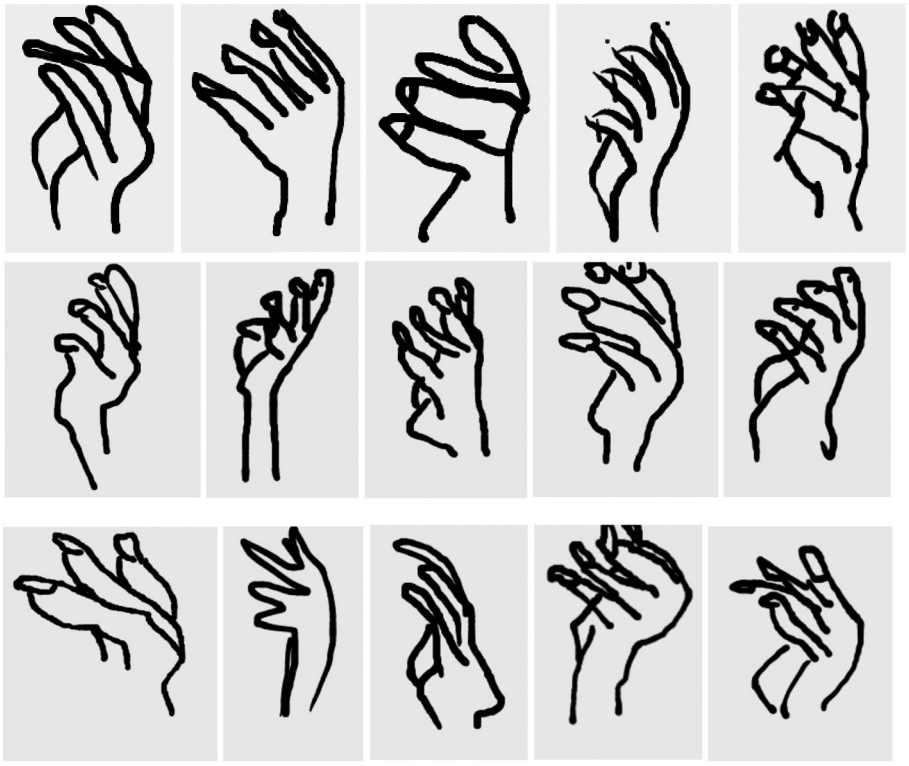


Appendix B. Drawings from Experiment 8

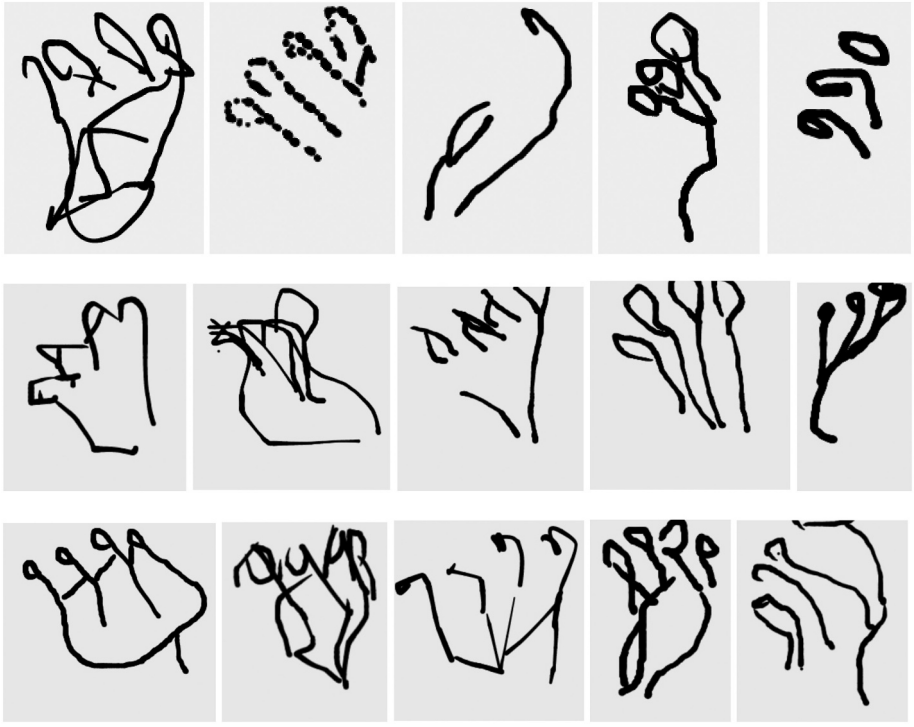
Expert drawing (to copy):

Competitor attempt (allegedly made by another participant):

Actual attempts of participants, 10 judged-good examples (see OSF for all 484 attempts):



Actual attempts of participants, 10 judged-bad examples (see OSF for all 484 attempts):



Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104381.

References

Ames, D. R. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*, 340–353.

Andrieux, M., & Proteau, L. (2016). Observational learning: Tell beginners what they are about to watch and they will learn better. *Frontiers in Psychology, 7*. Article 51.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215.

Berendt, J., & Uhrich, S. (2016). Enemies with benefits: The dual role of rivalry in shaping sports fans' identity. *European Sport Management Quarterly, 16*, 613–634.

Bohns, V. K. (2016). (Mis)understanding our influence over others: A review of the underestimation-of-compliance effect. *Current Directions in Psychological Science, 25*, 119–123.

Brooks, A. W., Huang, K., Abi-Esber, N., Buell, R. W., Huang, L., & Hall, B. (2019). Mitigating malicious envy: Why successful individuals should reveal their failures. *Journal of Experimental Psychology: General, 148*, 667–687.

Brooks, D. (2019). *The pleasure of watching others confront their own incompetence. New York Times*. (15 May). Available at https://www.nytimes.com/2019/05/15/magazine/pleasures-of-instructional-videos-drawing.html.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97*, 1232–1254.

Campbell, T., O'Brien, E., Van Boven, L., Schwarz, N., & Ubel, P. (2014). Too much experience: A desensitization bias in emotional perspective taking. *Journal of Personality and Social Psychology, 106*, 272–285.

Chou, H. T. G., & Edge, N. (2012). "They are happier and having better lives than I am": The impact of using Facebook on perceptions of others' lives. *Cyberpsychology, Behavior and Social Networking, 15*, 117–121.

Cusack, M., Vezenkova, N., Gottschalk, C., & Calin-Jageman, R. J. (2015). Direct and conceptual replications of Burgmer & Englich (2012): Power may have little to no effect on motor performance. *PLoS One, 10*, Article e0140806.

Cushman, F., Sarin, A., & Ho, M. (2021). Punishment as communication. In J. Doris, & M. Vargas (Eds.), *Oxford handbook of moral psychology*. Oxford, UK: Oxford Press.

Davidai, S., & Gilovich, T. (2016). The headwinds/tailwinds asymmetry: An availability bias in assessments of barriers and blessings. *Journal of Personality and Social Psychology, 111*, 835.

Dolan, L. (2020). From TikTok to Spotify: What the video loop says about Gen Z culture. https://medium.com/@leahdolan/from-gifs-to-tiktok-what-the-video-loop-says-about-gen-z-culture-dd5f4adc0755.

Duckworth, A., & Gross, J. J. (2014). Self-control and grit: Related but separable determinants of success. *Current Directions in Psychological Science, 23*, 319–325.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*, 100–110.

Dweck, C. S. (2008). *Mindset: The new psychology of success*. New York, NY: Random House.

Dylan, B. (1964). The times they are a-changin. Available at https://www.bobdylan.com/songs/times-they-are-changin/.

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572–595). Oxford, UK: Oxford University Press.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363–406.

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

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*, 175–191.

Frederick, S., & Loewenstein, G. (1999). Hedonic adaptation. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Wellbeing: The foundations of hedonic psychology* (pp. 302–329). New York, NY: Russell Sage Foundation.

Galak, J., & Redden, J. P. (2018). The properties and antecedents of hedonic decline. *Annual Review of Psychology, 69*, 1–25.

Groves, P. M., & Thompson, R. F. (1970). Habituation: A dual-process theory. *Psychological Review, 77*, 419–450.

Hareli, S., & Weiner, B. (2002). Dislike and envy as antecedents of pleasure at another's misfortune. *Motivation and Emotion, 26*, 257–277.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences, 5*, 253–261.

Heyes, C. M., & Foster, C. L. (2002). Motor learning by observation: Evidence from a serial reaction time task. *The Quarterly Journal of Experimental Psychology: Section A, 55*, 593–607.

Hornsey, M. J., Oppes, T., & Svensson, A. (2002). "It's OK if we say it, but you can't": Responses to intergroup and intragroup criticism. *European Journal of Social Psychology, 32*, 293–307.

Ifcher, J., & Zarghamee, H. (2014). Affect and overconfidence: A laboratory investigation. *Journal of Neuroscience, Psychology, and Economics, 7*, 125–150.

Jordan, A. H., Monin, B., Dweck, C. S., Lovett, B. J., John, O. P., & Gross, J. J. (2011). Misery has more company than people think: Underestimating the prevalence of others' negative emotions. *Personality and Social Psychology Bulletin, 37*, 120–135.

Kardas, M., & O'Brien, E. (2018). Easier seen than done: Merely watching others perform can foster an illusion of skill acquisition. *Psychological Science, 29*, 521–536.

Katz, D., & Allport, F. H. (1931). *Student attitudes*. Syracuse, NY: Craftsman Press.

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

Koellinger, P., & Treffers, T. (2015). Joy leads to overconfidence, and a simple countermeasure. *PLoS One, 10*, Article e0143263.

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

Konrath, S., O'Brien, E., & Hsing, C. (2011). Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review, 15*, 180–198.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*, 19751–19756.

Lyubomirsky, S., & Ross, L. (1997). Hedonic consequences of social comparison: A contrast of happy and unhappy people. *Journal of Personality and Social Psychology, 73*, 1141–1157.

Marketing, Vala (2020). https://valamarketing.com/tik-tok/.

Mattar, A. A., & Gribble, P. L. (2005). Motor learning by observing. *Neuron, 46*, 153–160.

Nickerson, R. S. (1999). How we know—And sometimes misjudge— What others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737–759.

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.

O'Brien, E., & Ellsworth, P. C. (2012). More than skin deep: Visceral states are not projected onto dissimilar others. *Psychological Science, 23*(4), 391–396. https://doi.org/10.1177/0956797611432179.

O'Brien, E., Kristal, A. C., Ellsworth, P. C., & Schwarz, N. (2018). (Mis)imagining the good life and the bad life: Envy and pity as a function of the focusing illusion. *Journal of Experimental Social Psychology, 75*, 41–53.

Packer, D. J. (2008). On being both with us and against us: A normative conflict model of dissent in social groups. *Personality and Social Psychology Review, 12*, 50–72.

Prinz, A., Bergmann, V., & Wittwer, J. (2018). Happy but overconfident: Positive affect leads to inaccurate metacomprehension. *Cognition and Emotion, 33*, 606–615.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Personality and Social Psychology, 13*, 279–301.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*, 341–357.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521–562.

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology, 114*, 10–28.

Scully, D. M., & Newell, K. M. (1985). Observational-learning and the acquisition of motor skills: Toward a visual perception perspective. *Journal of Human Movement Studies, 11*, 169–186.

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology, 65*, 317–338.

Sidi, Y., Ackerman, R., & Erez, A. (2017). Feeling happy and (over)confident: The role of positive affect in metacognitive processes. *Cognition and Emotion, 32*, 876–884.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*, 255–259.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*, 813–838.

Spotify. (2022). https://canvas.spotify.com/en-us.

Statista. (2021). https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/.

Steinmetz, J. (2018). Impression (mis)management when communicating success. *Basic and Applied Social Psychology, 40*, 320–328.

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the narcissistic personality inventory. *Journal of Personality, 76*, 875–902.

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.), *Vol. 48. Advances in experimental social psychology* (pp. 117–171). Burlington, VT: Academic Press.

van de Ven, N., Zeelenberg, M., & Pieters, R. (2009). Leveling up and down: The experiences of benign and malicious envy. *Emotion, 9*, 419–429.

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin, 90*, 245–271.
Wilson, A. E., & Ross, M. (2001). From chump to champ: People's appraisals of their earlier and present selves. *Journal of Personality and Social Psychology, 80*, 572–584.
Wilson, T. D., & Gilbert, D. T. (2008). Explain away: A model of affective adaptation. *Perspectives on Psychological Science, 3*, 370–386.
Wolpe, J. (1982). *The practice of behavior therapy* (3rd ed.). New York, NY: Pergamom Press.
Wondra, J. D., & Ellsworth, P. C. (2015). An appraisal theory of empathy and other vicarious emotional experiences. *Psychological Review, 122*, 411–428.