

The Counterfactual Detective: People Think About What Could Have Been to Uncover Past Events

Social Psychological and
Personality Science
1–9

© The Author(s) 2026



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/19485506261477807
journals.sagepub.com/home/spp



Jordan Bauman¹, Hailey Pawsey¹, Stephanie Denison¹,
and Ori Friedman¹

Abstract

We reveal a way that people use counterfactual thinking: They use it to infer motives and solve mysteries about what happened in the unseen past. In three preregistered experiments, participants (total $N = 1,242$) were asked to determine which of two competitors in a contest had sabotaged a third competitor. Participants often succeeded in identifying the culprit when counterfactual reasoning could be used to judge whether each suspect profited from the sabotage. Participants had more difficulty, though, when the mystery required using counterfactuals to judge how much each suspect profited. These findings show that people can use counterfactual reasoning to infer others' motives and to determine who was most likely to have produced past events. But they find this easier when counterfactuals support binary judgments rather than graded ones. Besides revealing a way that people use counterfactual reasoning, the findings are informative about how people learn about the unobserved past.

Keywords

counterfactuals, historical inferences, motive

Handling Editor: André Mata

Introduction

Sherlock Holmes and Watson are investigating a case. A small shop promised £10 bonuses to its two top salespeople. On the morning the bonuses were to be awarded, the shop owner received a telegram saying that Atkinson, the leading salesperson, had been stealing from the store for years. Atkinson was fired on the spot, and the bonuses were awarded to Bennett and Carter, the next best salespeople. Atkinson was later able to prove that the allegation in the telegram was false. It is up to Holmes and Watson, though, to determine who sent it. Watson immediately concludes the culprit is Bennett, since he was originally in second place and finished in first place with Atkinson fired. But Holmes tells Watson that he has pinned blame on the wrong salesperson. To solve the mystery, Watson must think counterfactually. He should consider how things could have turned out differently.

Watson is surprised. As an avid reader, he knows about the myriad functions, uses, and consequences of counterfactual reasoning (for reviews, see Byrne, 2016; De Brigard, 2025; Smallman & Summerville, 2018). People consider counterfactuals when thinking about how negative outcomes could have been prevented (e.g., Mandel & Lehman, 1996) and when preparing for the future (Markman et al., 1993; Roese & Epstude, 2017). Thinking about how things could

have gone better or worse affects how people feel about the events that did happen (e.g., Medvec et al., 1995; Roese, 1994) and the extent to which people see the events as lucky or unlucky (e.g., Teigen, 1996). It affects how events are remembered (De Brigard et al., 2018) and drives curiosity (e.g., FitzGibbon et al., 2021; Summerville, 2011). Imagining counterfactual alternatives to reality also influences and may even be required for attributing causality and moral responsibility (e.g., Gerstenberg et al., 2021; Wells & Gavanski, 1989). None of this, though, looks relevant for determining who sent the telegram. For example, when people consider counterfactuals to judge whether one event caused another, they know which events happened. So while simulating counterfactuals helps people determine causal relations between events they observed, it does not help them uncover which events happened in the unseen past.

Counterfactual reasoning is nonetheless the key to solving the mystery, and it points to Carter as the culprit: The shop owner had set out two bonuses, and Carter was

¹University of Waterloo, Ontario, Canada

Corresponding Author:

Ori Friedman, School of Psychology, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

Email: friedman@uwaterloo.ca

initially in third place. If the misleading telegram had not been sent (i.e., a counterfactual outcome), Carter would have gone empty-handed. Because Bennett was originally in second place, he was guaranteed a bonus either way. Here, determining who gained the most requires comparing the actual outcome with the counterfactual one that would have happened if the misleading telegram had not been sent. So, the *Case of the Misleading Telegram* reveals a potential use for counterfactual reasoning: Thinking about counterfactuals can be used to determine who was most motivated to produce some outcome and, by extension, to infer who did it. As such, it can be used to infer history and to deduce the nature of unseen past events.

In this paper, we investigate people's ability to use counterfactuals to uncover past events. The *Case of the Misleading Telegram* reveals that counterfactual reasoning can be used in this way. But are people besides Sherlock Holmes regularly able to do so? Besides expanding understanding of the functions and uses of counterfactual reasoning, investigating this question will also be informative about people's ability to infer history and reconstruct unseen past events. Existing work suggests that people draw these inferences in a variety of ways. They rely on statistical regularities, for example, recognizing physical similarities among objects formed in the same way (e.g., Fleming & Schmidt, 2019; Schmidt et al., 2019). They rely on core assumptions and constraints about how outcomes can be produced, such as the assumption that stacks are built from the ground up (Wong et al., 2025; also see Goulding et al., 2025). And they use their knowledge of causal forces together with their theory of mind to reconstruct or simulate different ways that outcomes could have been produced (Jara-Ettinger & Schachner, 2024; Lopez-Brau et al., 2022).¹ Counterfactual reasoning could be a further way for people to infer past events.

The Current Experiments

To investigate people's ability to use counterfactuals to uncover past events, we had participants read stories about contests for prizes. A competitor who was expected to win a prize was sabotaged and thereby removed from the contest. Participants then judged which of two other competitors was the saboteur.

We also explored two factors that might affect people's use of counterfactual reasoning to uncover the past. Often when someone is sabotaged, suspicion falls on their closest rival—for instance, the competitor who had been expected to finish just behind the victim. This tendency may be especially strong when a top performer is sabotaged, since the closest rival (i.e., the contestant originally in second place) ends up as the overall winner. A bias to blame the contestant who ends up winning, then, could prevent people from recognizing that some other competitor profited more. If so, people might be more likely to reason counterfactually when the victim is not a top contender—for instance, when a middling performer is sabotaged.

People's use of counterfactual reasoning to uncover the past could also depend on the kinds of comparisons judgments need to make. In the *Case of the Misleading Telegram*, Holmes had to determine whether each suspect profited from the telegram, yes or no. Counterfactual reasoning might be more difficult, though, if both suspects profited. Then counterfactual comparisons would be needed to determine which suspect profited more. This requires determining how much each gained, a graded judgment rather than a binary one.

General Methods

Preregistrations, materials, data, and code for all experiments are available on OSF at <https://osf.io/hg6rp/overview>. The Supplemental Materials report an additional experiment, conducted before those reported here. We disclose all measures, manipulations, and exclusions.

In each experiment, we sought to recruit at least 100 participants per between-subjects condition. This goal was not based on power calculations; we chose it because we observed it sufficed to reveal significant effects in many earlier experiments using varied designs. Even so, a sensitivity power analysis conducted using *g*power* 3.1 (Faul et al., 2007) suggested that this provided 80% power to detect significant interactions in each experiment with a small-to-medium effect size (i.e., Cohen's $f = 0.18$).

We recruited participants using Prolific and required that they were in the United States and were using a laptop or a tablet. The experiments included multiple-choice attention-check questions (see the materials at OSF), and participants were excluded if they answered at least one of these questions incorrectly or if they skipped questions in the main task. After the attention checks, participants were asked for their age and gender.

We analyzed the data in R, using *afex* (Singmann et al., 2024) for analysis of variance (ANOVA) and *emmeans* (Lenth, 2023) for post hoc tests. This paper was written using *apaquarto* (Schneider, 2026) and *apa_print* from *papaja* (Aust & Barth, 2024).

Experiment 1

Participants read a scenario where counterfactual reasoning supported binary judgments: Would each suspect have been worse off if not for the sabotage? To examine whether a tendency to pin the blame on the eventual winner interferes with this use of counterfactual reasoning, participants either read stories about top-performing contenders or middling ones.

Methods

Participants. The experiment was successfully completed by 433 participants (mean age = 36 years; 211 women, 212 men, and 10 indicating other or preferring not to answer); an additional eight participants were excluded for failing comprehension checks.

Each year, a company gives out \$9,000 bonuses to the salespeople who sell the most. A leaderboard is updated every day and is available for all employees to see. Here are this year's final results.

Rank	Employee
1	Cory
2	Robert
3	Angela
4	Pat
5	Sam
6	Remy
7	Jordan
8	Dalton
9	Karen
10	Jesse

Another salesperson, Martin, was set to finish in fourth place but was fired the night before the bonuses were awarded. An anonymous email was sent to the boss with false information about him.

Figure 1. Vignette from Experiment 1.

Note. This figure shows the script when Martin was set to finish in fourth place and five bonuses were available.

Procedure. Participants read about a company that gives \$9,000 bonuses each year to each of its top salespeople; see Figure 1. Martin was set to win a bonus but was instead fired because someone sent the company boss an email with false information about him. Participants saw a leaderboard showing the final rankings of 10 employees and judged which of the two employees, Pat or Sam, had sent the email. Both employees ended up winning bonuses. Pat was the stronger performer and ended up finishing in Martin's original position; Sam was the weaker performer and finished one spot below. Participants responded using a 6-point scale ranging from "very likely Pat" (coded 1) to "very likely Sam" (coded 6). That is, higher scores indicated greater belief that the *weaker* performer was the saboteur.

The experiment used a 2×2 between-subjects design. One factor was whether the characters were highly ranked or middling. When the characters were highly ranked, Pat took first place over from Martin, and Sam ended up in second place. When the characters had middling rankings, Pat took over fourth place from Martin, and Sam ended up in fifth place.

The other factor was who benefited from the sabotage. We manipulated this by varying how many bonuses were awarded. When Martin was originally in first place, there was one bonus or two; when he was originally in fourth place, there were four bonuses or five. Either way, this affected which suspect profited. For example, when Martin was originally in fourth place and four bonuses were awarded, his firing allowed Pat but not Sam to win a bonus. That is, the sabotage benefited only the stronger performer. By contrast, with five bonuses awarded, Pat was already guaranteed a bonus, but the firing allowed Sam to win one. So here, the sabotage benefited only the weaker performer.

Results

The left panel of Figure 2 shows participants' mean ratings. Higher ratings indicate more confidence that the saboteur was the weaker performer; lower ratings indicate more confidence that it was the stronger performer.

We entered ratings into a 2 (who profited: weaker performer, stronger performer) $\times$ 2 (area of leaderboard: high, middling) ANOVA. It revealed main effects of both who profited, $F(1, 429) = 174.22, p < .001, \eta_p^2 = .289$, and area of leaderboard, $F(1, 429) = 15.43, p < .001, \eta_p^2 = .035$, but no interaction, $F(1, 429) = 0.52, p = .470, \eta_p^2 = .001$. Participants more strongly believed the weaker performer was the culprit when she profited from the misleading telegram, and when the characters had middling rankings rather than high ones.

Single-sample tests against the scale midpoint (3.5) showed that when the weaker performer benefited more, participants mostly agreed she was the culprit *only* when the contestants had middling ranks ($M = 4.21$, 95% CI [3.96, 4.46], $t(109) = 5.57, p < .001$), but not when they had high rankings ($M = 3.65$, 95% CI [3.39, 3.92], $t(114) = 1.14, p = .257$). By contrast, when the stronger performer benefited more, participants mostly agreed she was the culprit regardless of whether the candidates' rankings were middling ($M = 2.54$, 95% CI [2.30, 2.78], $t(106) = -8.00, p < .001$) or high ($M = 2.16$, 95% CI [1.99, 2.33], $t(100) = -15.73, p < .001$).

We also conducted an exploratory analysis of the proportion of participants who suspected each character. For this analysis, we classified participants as either mostly suspecting the weaker performer (Coded 1) or mostly suspecting the stronger one (Coded 0) based on whether their ratings fell above or below the scale midpoint. The right panel of Figure 2 shows these results.

A binomial logistic regression revealed the same pattern observed with participants' ratings: There were main effects of who profited, $\chi^2(1) = 146.78, p < .001$, and area of leaderboard, $\chi^2(1) = 20.03, p < .001$, but no interaction, $\chi^2(1) = 2.71, p = .099$. A greater proportion of participants thought the culprit was the weaker performer when she profited from the misleading telegram, and also when the characters had middling rankings rather than high ones.

Discussion

Consistent with counterfactual reasoning, participants suspected the weaker-performing competitor more strongly when they profited from the sabotage (more prizes awarded) compared to when the stronger performing competitor profited (fewer prizes awarded). Participants also showed a tendency blame on the contestant originally in second place, as they suspected the stronger performer more strongly when the characters all had high ranks (and this character was originally expected to finish in second). This tendency, though, did not interfere with

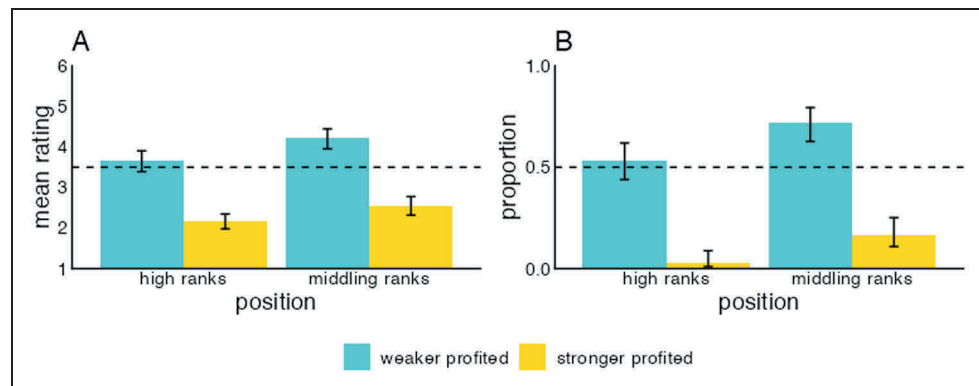


Figure 2. Mean results and 95% confidence intervals of Experiment 1.

Note. Participants read stories where a contestant with a high or middling rank was eliminated from a contest by sabotage (between-subjects). They then judged which of the two other contestants was responsible for the sabotage: a strong performer who ended up finishing in the victim's original position or a weaker performer who ended up finishing one spot below. The left panel (A) shows participants' ratings on a 6-point scale where lower values indicate greater suspicion of the stronger performer and higher values indicate greater suspicion of the weaker performer. The right panel (B) shows the proportion of participants who mostly suspected the weaker performer rather than the stronger one.

counterfactual reasoning as the effect of who profited did not diminish when ranks were high rather than middling.

Experiment 2

We next examined whether people would use counterfactual reasoning with graded judgments. Here, both suspects financially benefited from the sabotage, and so identifying the culprit required determining which character profited more. As before, we again examined whether counterfactual reasoning was disrupted by a tendency to pin the blame on the contestant originally in second place.

Methods

Participants. The experiment was successfully completed by 403 participants (mean age = 38 years; 199 women, 197 men, and 7 indicating other or preferring not to answer); an additional 13 participants were excluded for failing comprehension checks.

Procedure. Participants again read about a company that gives bonuses to its most productive employees. As before, Martin was set to win a bonus but was fired beforehand because of an email with false information about him. Participants saw a leaderboard showing the final rankings of six employees and judged which of the two employees, Pat or Sam, had sent the email. They responded using a 6-point scale ranging from "very likely Pat" (coded 1) to "very likely Sam" (coded 6).

The experiment used a 2×2 between-subjects design. One factor was whether the characters were ranked at the top or middle of the leaderboard. Either Pat took over from

Martin as the top performer with Sam ranked second, or else Pat took over fourth place with Sam ranked fifth.

The other factor was who profited most from Martin being fired. Here, this was manipulated by varying the value of the bonuses, which were either set so that Pat (who took over Martin's rank) gained \$4,000 from the sabotage while Sam (who was one rank lower) gained \$1,000, or so that these values were reversed.² For instance, when the characters had middling ranks and Sam profited the most, Pat's prize was \$6,000, Sam's was \$5,000, and the next employee won just \$1,000. In the corresponding condition where Pat profited most, Pat's prize was still \$6,000, Sam's was \$2,000, and the next employee again won just \$1,000. Hence, both Pat and Sam were guaranteed money even without the sabotage. But which of them profited more from depended on the prize structure.

Results

The left panel of Figure 3 shows mean ratings of who sent the misleading message; higher ratings indicate more confidence that this was the weaker performer. A 2 (who profited more: weaker performer, stronger performer) $\times$ 2 (area of leaderboard: high, middling) ANOVA revealed a main effect of who profited more, $F(1, 399) = 8.12$, $p = .005$, $\eta_p^2 = .020$. It resulted because participants gave somewhat higher ratings when it was the weaker performer who profited more. There was no main effect of area of leaderboard, $F(1, 399) = 3.75$, $p = .053$, $\eta_p^2 = .009$, and no interaction, $F(1, 399) = 0.12$, $p = .731$, $\eta_p^2 = .000$.

Single-sample t -tests against the scale midpoint (3.5) showed that participants predominantly suspected the stronger performer. They suspected this character more

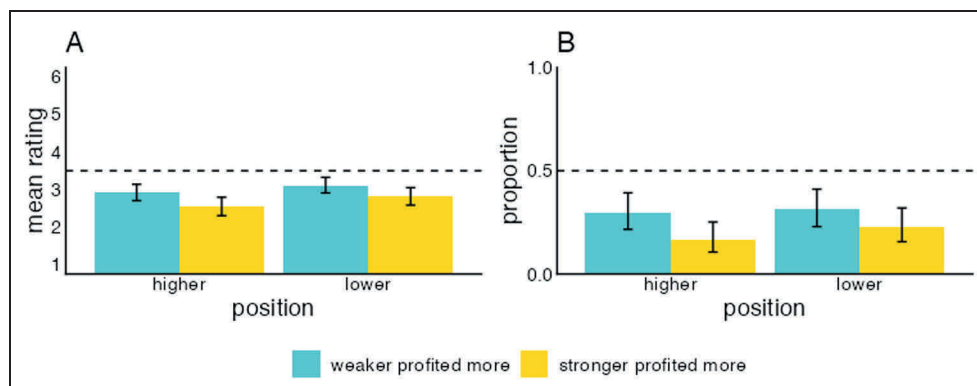


Figure 3. Mean results and 95% confidence intervals of Experiment 2.

Note. Participants read stories where a contestant with a high or middling rank was eliminated from a contest by sabotage (between subjects). They then judged who was responsible for the sabotage: a strong performer who ended up finishing in the victim's original position or a weaker performer who ended up finishing one spot below. The left panel (A) shows participants' ratings on a 6-point scale where lower values indicate greater suspicion of the stronger performer and higher values indicate greater suspicion of the weaker performer. The right panel (B) shows the proportion of participants who mostly suspected the weaker performer rather than the stronger one.

both when the weaker performer reaped the greater benefit ($M = 3.02$, 95% CI [2.86, 3.17], $t(199) = -6.01$, $p < .001$) and when the stronger performer did ($M = 2.67$, 95% CI [2.50, 2.85], $t(202) = -9.33$, $p < .001$).

We again conducted an exploratory analysis of the proportion of participants who suspected each character. Participants who mostly suspected the weaker performer were coded 1, while those who mostly suspected the stronger performer were coded 0. The right panel of Figure 3 shows these results. A binomial logistic regression revealed the same pattern observed with ratings: There was a main effect of who profited more, $\chi^2(1) = 6.46$, $p = .011$, but no main effect of area of leaderboard, $\chi^2(1) = 0.98$, $p = .323$, and no interaction, $\chi^2(1) = 0.44$, $p = .506$.

Discussion

Opposite to the first experiment, participants did not appear to reason counterfactually. Although they showed some sensitivity to who gained more from the sabotage, they mostly suspected the stronger performer even when the weaker performer profited more.

Experiment 3

One explanation for the contrast between Experiments 1 and 2 is that while some people use counterfactuals to make binary judgments (*did each character profit?*), they struggle with graded judgments (*how much did each character profit?*). We tested this possibility by directly comparing these two uses of counterfactuals. We also attempted to minimize task demands that could impede counterfactual reasoning (particularly with graded judgments). To this end, we used a story about three alphabetically named

characters (as in the *Case of the Misleading Telegram*) and only asked about top performers. In addition, after asking participants about which character was the saboteur, we asked them which character profited more from the initial leader's expulsion from the contest. This question probes counterfactual reasoning without requiring participants to uncover past events.

Methods

Participants. The experiment was successfully completed by 406 participants (mean age = 36 years; 223 women, 175 men, and 8 indicating other or preferring not to answer); an additional four participants were excluded for failing comprehension checks.

Procedure. Participants read about a company that gives employees bonuses based on their speed in assembling advanced cameras. Abe is the fastest worker in the company, Bill is second fastest, and Carl is third fastest. Abe misses out on getting a bonus, however, because he leaves work in response to a text message falsely informing him about a family emergency. Because this story was about just three characters, the final results were described in a single sentence rather than shown on a leaderboard.

This experiment again used a 2×2 between-subjects design. One factor was whether the reward structure (and corresponding counterfactual judgment) was binary or graded. With the binary structure, all bonuses were for the same amount; with graded structure, faster workers earned higher bonuses.

The other factor was who gained the most from the sabotage. This was manipulated through further variations in the reward structure. With binary rewards, the company

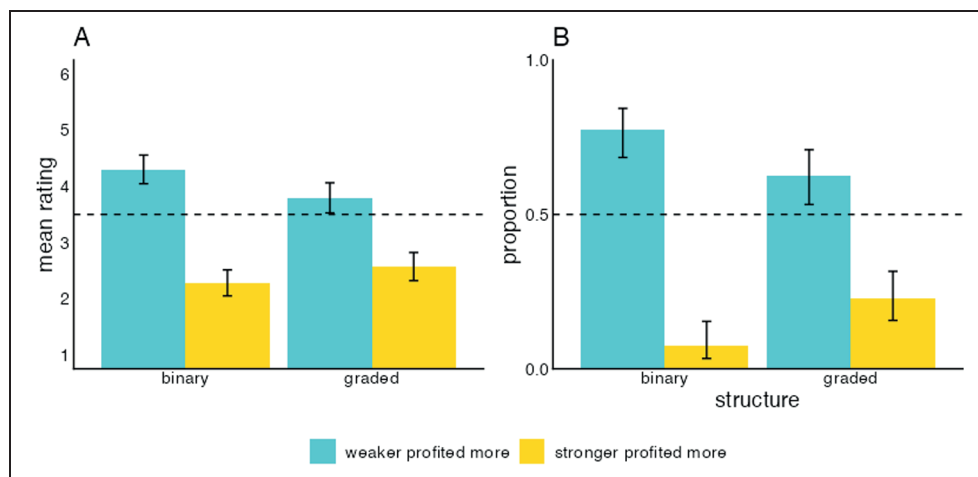


Figure 4. Mean results and 95% confidence intervals of Experiment 3.

Note. Participants read a story where an employee was eliminated from a contest by sabotage and then judged which of the two other contestants was responsible for the sabotage: a strong performer or a weaker performer who ended up finishing one spot below. The left panel (A) shows participants' ratings on a 6-point scale where lower ratings indicate greater suspicion of the stronger performer and higher ratings indicate greater suspicion of the weaker performer. The right panel (B) shows the proportion of participants who mostly suspected the weaker performer rather than the stronger one.

awarded either one or two \$1,000 bonuses. The faster suspect (Bill) benefited when there was just one bonus, whereas the slower suspect (Carl) benefited when there were two. With graded rewards, the first-place bonus was always \$1,000, and the second-place bonus was either \$100 or \$900. When the second-place bonus was \$100, the sabotage allowed the faster suspect to gain \$800 and the slower suspect to gain \$100. These gains reversed when the second-place bonus was \$900.

Participants were asked who sent the false message and responded using a 6-point scale ranging from "very likely Bill" (coded 1) to "very likely Carl" (coded 6). On a separate screen, they were also asked a forced-choice question about whether Bill or Carl gained the most money from Abe's departure.

Results

The left panel of Figure 4 shows mean ratings of who sent the misleading message; higher ratings indicate more confidence that this was the weaker performer. A 2 (who profited more: weaker performer, stronger performer) $\times$ 2 (structure: binary, graded) ANOVA revealed a main effect of who profited more, $F(1, 402) = 145.61$, $p < .001$, $\eta_p^2 = .266$, no effect of structure, $F(1, 402) = 0.68$, $p = .411$, $\eta_p^2 = .002$, and a significant interaction, $F(1, 402) = 8.66$, $p = .003$, $\eta_p^2 = .021$.

Participants more strongly chose the weaker performer when he had the most to gain, compared with when the stronger performer did, though this effect was greater when rewards were binary, $\Delta M = 2.02$, 95% CI [1.63, 2.41], $t(402) = 10.21$, $p < .001$, rather than graded, $\Delta M = 1.23$,

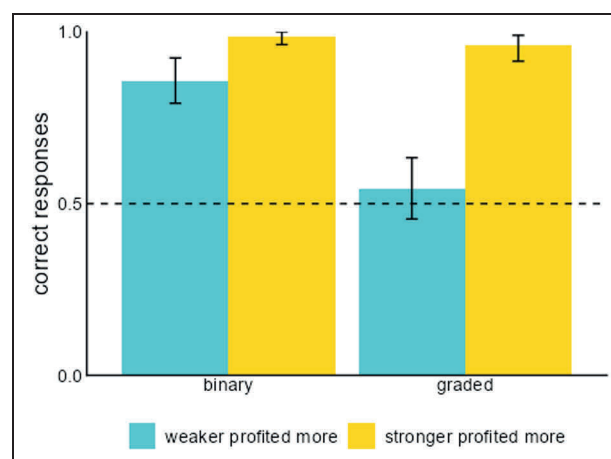


Figure 5. Mean judgments about who profited more and 95% confidence intervals of Experiment 3.

Note. The proportion of correct responses to the follow-up question in Experiment 3 about which character profited more from the sabotage.

95% CI [0.87, 1.59], $t(402) = 6.73$, $p < .001$. Single-sample t -tests against the scale midpoint (3.5) found that participants mostly concluded the message was sent by the weaker performer when this character profited more: binary, $M = 4.30$, 95% CI [4.04, 4.57], $t(105) = 6.04$, $p < .001$; graded, $M = 3.79$, 95% CI [3.52, 4.07], $t(111) = 2.09$, $p = .039$. They likewise mostly concluded it was sent by the stronger performer when this character profited more: binary, $M = 2.28$, 95% CI [2.06, 2.50], $t(81) = -10.81$, $p < .001$; graded, $M = 2.57$, 95% CI [2.30, 2.83], $t(105) = -6.93$, $p < .001$.

As in the previous experiment, we also conducted an exploratory analysis of the proportion of participants who suspected each character. Participants who suspected the weaker performer were coded 1, while those who suspected the stronger performer were coded 0; see the right panel of Figure 4. A binomial logistic regression revealed the same pattern observed with participants' ratings. There was a main effect of who profited more, $\chi^2(1) = 135.41$, $p < .001$, no main effect of structure, $\chi^2(1) = 1.12$, $p = .289$, and a significant interaction, $\chi^2(1) = 14.42$, $p < .001$. A higher proportion of participants suspected the weaker performer when he had the most to gain, compared with when the stronger performer did, though this effect was greater when rewards were binary, $\Delta M = 0.70$, 95% CI [0.60, 0.80], $z = 14.07$, $p < .001$, rather than graded, $\Delta M = 0.40$, 95% CI [0.28, 0.52], $z = 6.51$, $p < .001$.

We also examined responses to the question about which character profited more. Figure 5 shows the mean proportion of participants who correctly chose the character who profited more over the character who profited less. A 2 (who profited more: weaker performer, stronger performer) $\times$ 2 (structure: binary, graded) generalized linear model revealed a main effect of who profited more, $\chi^2(1) = 49.24$, $p < .001$, a main effect of structure, $\chi^2(1) = 7.28$, $p = .007$, but no interaction, $\chi^2(1) = 0.15$, $p = .702$.³ Participants were more likely to respond correctly when the stronger performer profited more than when the weaker one did, and with binary judgments compared with graded ones. We were surprised that the model did not yield a significant interaction since the plot in Figure 4 strongly suggests one. Failure to detect an interaction may have resulted because of ceiling effects in cells where the stronger performer profited more.

In an exploratory analysis, we revisited judgments of which competitor was the saboteur but focused on participants who correctly indicated which had profited more from the sabotage. With these participants, a 2 (who profited more) $\times$ 2 (structure) ANOVA revealed only a main effect of who profited more, $F(1, 331) = 263.30$, $p < .001$, $\eta_p^2 = .443$, with no effect of structure, $F(1, 331) = 0.00$, $p = .964$, $\eta_p^2 = .000$, and no interaction, $F(1, 331) = 2.87$, $p = .091$, $\eta_p^2 = .009$. Hence, participants who recognized which suspect profited more were similarly successful in identifying the likely saboteur regardless of judgment type. They predominantly suspected the weaker performer when this suspect profited more, $M = 4.54$, 95% CI [4.36, 4.72], $t(151) = 11.20$, $p < .001$, and predominantly suspected the stronger performer when they profited more, $M = 2.40$, 95% CI [2.22, 2.58], $t(182) = -12.28$, $p < .001$.

Discussion

Participants suspected the weaker-performing competitor more strongly when they profited more from the sabotage

than when the stronger performing competitor profited more, though the effect was greater with binary than graded judgments. The difference between binary and graded judgments disappeared, though, when we restricted analysis to participants who responded correctly to a follow-up question about which character profited most.

General Discussion

Using counterfactual reasoning to uncover past events is not just for legendary detectives like Sherlock Holmes. Many of our participants were able to use counterfactual reasoning in this way too. In determining who sabotaged a competitor, they assessed which suspect profited from the sabotage—a judgment which requires comparing actual outcome with the counterfactual outcome, which would have resulted if not for the sabotage. Although people are known to use counterfactual reasoning in myriad ways, previous work had not shown that people use counterfactual reasoning to infer motives and thereby uncover past events. So besides contributing to our understanding of ways that people use counterfactual reasoning, this work is also informative about how they infer history (Jara-Ettinger & Schachner, 2024) and others' motives (Liefgreen et al., 2021).

Our findings also reveal a factor that affects the difficulty of counterfactual reasoning. Participants more often showed signs of thinking about counterfactual outcomes when this supported binary judgments (*did each suspect profit?*) than graded ones (*how much did each profit?*). This difference is apparent when comparing the findings of the first and second experiments, and we directly tested it in the final experiment. Recent work on counterfactual reasoning has also distinguished between binary and graded uses of counterfactual reasoning (e.g., Gerstenberg et al., 2021; Lagnado et al., 2013). To our knowledge, though, the present work is the first to directly compare the difficulty they pose. One question for future research is whether this difference would hold if participants were incentivized to identify the true culprit.

Does determining who profited the most from some event, like sabotage, always require thinking about counterfactual outcomes? No. Participants probably did not need to think counterfactually when the contests offered a single prize. In this situation, the culprit can be identified by looking to the final outcome alone—who ended up with the prize? But even with other prize structures, the need for counterfactual reasoning depends on what information is available. In our tasks, participants were told the final results of the contest and had to work out what would have happened if not for the sabotage (i.e., the counterfactual outcome). The need for counterfactual reasoning would be removed, then, if we instead told participants the outcome

expected before the sabotage happened and had them work out the final results.

Do people use counterfactual reasoning to uncover past events in other ways, that is, in ways besides inferring who profited most? One potential place people might do this is in thinking about the implications of missing evidence. This idea is suggested by one of the original Sherlock Holmes stories: *The Adventure of Silver Blaze* (Conan Doyle, 1892). In the story, Holmes concludes that a racehorse was taken from its stable by an insider. The key clue is that a nearby dog did not bark during the incident. This solution appears to rest on counterfactual reasoning: If the culprit had been a stranger, the dog would have barked (Collier, 2011; Einhorn & Hogarth, 1982). However, this case does not require counterfactual reasoning. Instead, it can be solved using the generalizations *dogs bark at strangers* and *dogs remain silent with familiar people*. It remains to be seen, then, if there are other ways in which people use counterfactual reasoning to uncover past events. In our studies, this hinged on the connection between real-world events (the sabotage) and mental states (motives). Thinking about other ways these can intersect might be the key to recognizing further uses of counterfactual reasoning.

ORCID iDs

Jordan Bauman  <https://orcid.org/0009-0002-1318-9589>
 Stephanie Denison  <https://orcid.org/0000-0002-6658-4139>
 Ori Friedman  <https://orcid.org/0000-0003-2346-9787>

Ethical Considerations

This research was approved by the Office of Research Ethics at the University of Waterloo (REB31953).

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: It was funded by separate Discovery Grants from the Natural Sciences and Engineering Research Council of Canada awarded to OF and SD.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

Preregistrations, materials, data, and analysis code for all experiments are available on OSF at https://osf.io/hg6rp/overview?view_only=b37fd8fdb7e4103be9fcc281fe8d587. We disclose all measures, manipulations, and exclusions.

Supplemental Material

The supplemental material is available in the online version of the article.

Notes

1. Thinking about other ways an outcome could have resulted *can* involve counterfactual reasoning in other circumstances. It involves counterfactuals when a person already knows how an outcome resulted but then considers other ways it could have been produced. But it does not involve counterfactuals when a person does not yet know how the outcome resulted—from the person's perspective, each way of producing the outcome could be correct.
2. This way of manipulating potential gains across conditions was intended to equate the employees' motivations. For example, if participants felt that the stronger performer could be motivated to commit sabotage to gain \$4,000, they should likewise conclude this about the weaker performer. Admittedly, this does not take into account expectations about diminishing marginal utility which could increase suspicion of the weaker performer (e.g., gaining an extra \$4,000 should seem more valuable the *less* one would otherwise have had).
3. The pre-registration for this analysis specified running an ANOVA for this analysis, but that would be inappropriate since the data are binomial.

References

- Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. <https://doi.org/10.32614/CRAN.package.papaja>
- Byrne, R. M. J. (2016). Counterfactual thought. *Annual Review of Psychology*, 67(1), 135–157. <https://doi.org/10.1146/annurev-psych-122414-033249>
- Collier, D. (2011). Understanding process tracing. *PS: Political Science & Politics*, 44(4), 823–830. <https://doi.org/10.1017/s1049096511001429>
- Conan Doyle, A. (1892). The adventure of silver blaze. *Strand Magazine*, 4, 645–660. <https://archive.org/details/StrandMagazine24>
- De Brigard, F. (2025). Counterfactual thinking. In M. C. Frank & A. Majid (Eds.), *Open encyclopedia of cognitive science* (pp. 1–7). MIT Press.
- De Brigard, F., Hanna, E., St Jacques, P. L., & Schacter, D. L. (2018). How thinking about what could have been affects how we feel about what was. *Cognition and Emotion*, 33(4), 646–659. <https://doi.org/10.1080/02699931.2018.1478280>
- Einhorn, H. J., & Hogarth, R. M. (1982). *A theory of diagnostic inference. I. Imagination and the psychophysics of evidence* (Nos. TR-2). Chicago University Center for Decision Research. <https://apps.dtic.mil/sti/html/tr/ADA115940/>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/bf03193146>
- FitzGibbon, L., Komiya, A., & Murayama, K. (2021). The lure of counterfactual curiosity: People incur a cost to experience regret. *Psychological Science*, 32(2), 241–255. <https://doi.org/10.1177/0956797620963615>
- Fleming, R. W., & Schmidt, F. (2019). Getting “fumbered”: Classifying objects by what has been done to them. *Journal of Vision*, 19(4), 15. <https://doi.org/10.1167/19.4.15>
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of

- causal judgments for physical events. *Psychological Review*, 128(5), 936–975. <https://doi.org/10.1037/rev0000281>
- Goulding, B. W., Stonehouse, E. E., & Friedman, O. (2025). Time from structure: Children infer the temporal order of past events from visual arrays. *Psychonomic Bulletin & Review*, 32(4), 1731–1739. <https://doi.org/10.3758/s13423-025-02659-9>
- Jara-Ettinger, J., & Schachner, A. (2024). Traces of our past: The social representation of the physical world. *Current Directions in Psychological Science*, 33(5), 334–340. <https://doi.org/10.1177/09637214241268145>
- Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, 37(6), 1036–1073. <https://doi.org/10.1111/cogs.12054>
- Lenth, R. V. (2023). *Emmeans: Estimated marginal means, aka least-squares means*. <https://CRAN.R-project.org/package=emmeans>
- Liefgreen, A., Yousif, S. R., Keil, F. C., & Lagnado, D. A. (2021). Motive on the mind: Explanatory preferences at multiple stages of the legal-investigative process. *Cognition*, 217, 104892. <https://doi.org/10.1016/j.cognition.2021.104892>
- Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2022). Social inferences from physical evidence via bayesian event reconstruction. *Journal of Experimental Psychology: General*, 151(9), 2029–2042. <https://doi.org/10.1037/xge0001182>
- Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, 71(3), 450–463. <https://doi.org/10.1037/0022-3514.71.3.450>
- Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology*, 29(1), 87–109. <https://doi.org/10.1006/jesp.1993.1005>
- Medvec, V. H., Madey, S. F., & Gilovich, T. (1995). When less is more: Counterfactual thinking and satisfaction among Olympic medalists. *Journal of Personality and Social Psychology*, 69(4), 603–610. <https://doi.org/10.1037/0022-3514.69.4.603>
- Roese, N. J. (1994). The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology*, 66(5), 805–818. <https://doi.org/10.1037/0022-3514.66.5.805>
- Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In J. M. Olson, & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 56, pp. 1–79). Academic Press. <https://doi.org/10.1016/bs.aesp.2017.02.002>
- Schmidt, F., Phillips, F., & Fleming, R. W. (2019). Visual perception of shape-transforming processes: ‘Shape scission’. *Cognition*, 189, 167–180. <https://doi.org/10.1016/j.cognition.2019.04.006>
- Schneider, W. J. (2026). *apaquarto* [Computer software]. <https://github.com/wjschne/apaquarto>
- Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2024). *Afex: Analysis of factorial experiments*. <https://CRAN.R-project.org/package=afex>
- Smallman, R., & Summerville, A. (2018). Counterfactual thought in reasoning and performance. *Social and Personality Psychology Compass*, 12(4). <https://doi.org/10.1111/spc3.12376>
- Summerville, A. (2011). Counterfactual seeking. *Personality and Social Psychology Bulletin*, 37(11), 1522–1533. <https://doi.org/10.1177/0146167211413295>
- Teigen, K. H. (1996). Luck: The art of a near miss. *Scandinavian Journal of Psychology*, 37(2), 156171. <https://doi.org/10.1111/j.1467-9450.1996.tb00648.x>
- Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, 56(2), 161–169. <https://doi.org/10.1037/0022-3514.56.2.161>
- Wong, K. W., Shah, A. D., & Scholl, B. (2025). Seeing from the ground up: Spontaneous perception of ‘causal history’ due to intuitive physics. *Journal of Vision*, 25(9), 2006. <https://doi.org/10.1167/jov.25.9.2006>

Author Biographies

Jordan Bauman is an undergraduate student at the University of Waterloo. He is interested in cognitive and social responses to uncertainty and the intersection of psychology and religion.

Hailey Pawsey is now an elementary school teacher. She contributed to this work while completing a MASc at the University of Waterloo.

Stephanie Denison is a professor of psychology at the University of Waterloo. Much of her work investigates the development in infancy and childhood of probabilistic reasoning, inferences about mental states, and heuristics and biases.

Ori Friedman is a professor of psychology at the University of Waterloo. His interests include how children and adults recognize others’ thoughts and emotions, how they think about ownership rights and responsibilities, and how they determine whether events are possible.