The Desirability Bias in Predictions under Aleatory and Epistemic Uncertainty

Paul D. Windschitl[a], Jane E. Miller[a], Inkyung Park[a], Shanon Rule[a],
Ashley Jennings[a], Andrew R. Smith[b]
[a]University of Iowa
[b]Appalachian State University

Author Note

Paul D. Windschitl (https://orcid.org/0000-0002-4058-3779), Jane E. Miller (https://orcid.org/0000-0002-3487-9602), Inkyung Park (https://orcid.org/0000-0002-1681-6288), Shanon Rule, Ashley Jennings, Andrew R. Smith (https://orcid.org/0000-0001-5302-3343).

This work was supported by Grant SES-1851738 to Paul Windschitl and Andrew Smith from the National Science Foundation.

Correspondence concerning this article should be addressed to Paul D. Windschitl, Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA,52242. E-mail: paul-windschitl@uiowa.edu; Phone: 319-335-2435

Declarations of interest: none.

Abstract

The desirability bias (or wishful thinking effect) refers to when a person's desire regarding an event's occurrence has an unwarranted, optimistic influence on expectations about that event. Past experimental tests of this effect have been dominated by paradigms in which uncertainty about the target event is purely stochastic—i.e., involving only aleatory uncertainty. In six studies, we detected desirability biases using two new paradigms in which people made predictions about events for which their uncertainty was both aleatory and epistemic. We tested and meta-analyzed the impact of two potential moderators: the strength of evidence and the level of stochasticity. In support of the first moderator hypothesis, desirability biases were larger when people were making predictions about events for which the evidence for the possible outcomes was of similar strength (vs. not of similar strength). Regarding the second moderator hypothesis, the overall results did not support the notion that the desirability bias would be larger when the target event was higher vs. lower in stochasticity, although there was some significant evidence for moderation in one of the two paradigms. The findings broaden the generalizability of the desirability bias in predictions, yet they also reveal boundaries to an account of how stochasticity might provide affordances for optimistically biased predictions.

Key Words: desirability bias, wishful thinking, optimism, uncertainty, prediction

The Desirability Bias in Predictions under Aleatory and Epistemic Uncertainty

There are numerous strategies people can use to formulate predictions about uncertain events (Lagnado & Sloman, 2004). Unless a person has some control over an event, one factor they should ignore when making predictions is their own desire for how the event will turn out. For example, in predicting the weather for an outdoor ceremony, a desire for sunshine should not influence expectations. However, people often violate the imperative to base one's predictions on evidence, not desires. The term *desirability bias* (or *wishful thinking*) refers to when desiring an event inflates expectations that it will occur (Bar-Hillel & Budescu, 1995; Budescu & Bruderman, 1995; Krizan & Windschitl, 2007). The present work broadens the examination of the desirability bias from paradigms that involve just one form of uncertainty, to new paradigms in which people were faced with two forms of uncertainty. It also targets the role of two theoretically relevant moderators.

The desirability bias or wishful thinking has been examined in various ways. Correlational studies have revealed significant links between outcome preferences and expectations within domains like sports and politics (Babad, 1987; Granberg & Brent, 1983; Krizan et al., 2010; Markman & Hurt, 2002; Massey et al., 2011). However, because of potential confounds in correlational designs (e.g., people who strongly prefer a team also know more about that team than others), experimental designs have often been used to manipulate people's preferences and measure the causal impact on expectations. Some of these experimental studies have measured expectations as discrete predictions (asking people to indicate whether a target event will or will not happen; e.g., Irwin, 1953; Marks, 1951) and other studies have also measured expectations as likelihood judgments (e.g., Price & Marquez, 2005). The evidence that preferences impact likelihood judgments has been mixed (e.g., Bilgin, 2012; Harris, Corner, & Hahn, 2009; Krizan & Windschitl, 2007; Windschitl et al., 2010), with recent work suggesting

that predictions are generally more vulnerable to bias than are likelihood judgments (Park et al., in press).

Critically, studies that have manipulated preferences and measured predictions have relied heavily on paradigms in which the target events are purely stochastic events akin to those in games of chance. Most studies have used a variant of the marked-card paradigm (e.g., Budescu & Bruderman, 1995; Irwin, 1953; Marks, 1951; Price & Marquez, 2005). In the paradigm, participants predict whether a marked card, made desirable or not through a monetary manipulation, would be drawn from a deck containing a known portion of marked and unmarked cards. This proportion is usually also manipulated—e.g., from 10/90% to 90/10%. A meta-analysis revealed that the characteristic finding is that among 50-50 decks, predictions are heavily influenced by marked-card desirability (Krizan & Windschitl, 2007; see also Lench & Ditto, 2008; Lench et al., 2014; Windschitl et al., 2010).

Although the marked-card paradigm has produced remarkably consistent results, the paradigm suffers from an important generalizability problem. Uncertainty is not a unitary construct (Howell & Burnett, 1978; Kahneman & Tversky, 1982; Løhre & Teigen, 2015; Tannenbaum et al., 2017). Building on ideas discussed by Hacking (1975), Fox and Ulkümen (2011) discussed an important distinction between two forms of uncertainty: aleatory and epistemic. Aleatory uncertainty is the type that arises for stochastic events (or others) that can turn out differently with repeated runs under similar conditions. Examples include the uncertainty in games of chance and the uncertainty that is left after learning a forecaster's estimate for the chance of rain. Epistemic uncertainty arises when people realize they have incomplete information relevant to predicting an event that is, in principle, knowable. For example, when faced with the question about whether Paris or Madrid has a larger population, a person might have epistemic uncertainty. A physician's uncertainty about the cause of a patient's

symptom is also epistemic; the physician may not know the cause of the symptoms, but the cause is potentially knowable. Given that both aleatory and epistemic forms of uncertainty are relevant across a variety of life's domains, the heavy reliance on the marked-card paradigm—which only involves aleatory uncertainty—is limiting for fully understanding the desirability bias.

To address this limitation, Windschitl et al. (2010) developed an experimental paradigm that paralleled some of the structure of the marked-card paradigm, but tested for a desirability bias with items involving only epistemic uncertainty. Participants were presented with trivia items with two possible answers (e.g., "What animal makes a louder noise?—whale/lion"). A monetary manipulation was used to make participants prefer (or not prefer) that one of the answers was actually the factually valid answer. Participants' predictions were then solicited (i.e., predictions of which answer was the factually valid answer). The overall desirability bias was not significant. The desirability manipulation only significantly affected predictions on a limited group of the trivia items that were expressly designed to be almost impossibly difficult.

To explain this overall pattern, Windschitl et al. (2010) offered a *biased-guessing account*. The account assumes that people are often largely unbiased in assessing evidence, but people are inclined to give optimistic predictions and will take specific opportunities to do so. Even when evidence is stacked against predicting a preferred outcome, people might take an opportunity to make a preferred prediction if the prediction itself feels largely arbitrary—like a guess. With purely chance events, like in a marked-card paradigm, even when a preferred outcome only has a 40% chance, people may still predict the preferred outcome because the stochasticity of the event essentially frames the prediction as a guess (after all, a person who gives that prediction would be right 40% of the time). However, when there is no stochasticity—as in a trivia paradigm where all the uncertainty is epistemic—it is harder for a person to justify predicting a preferred answer when the evidence for that answer is viewed as weaker than the

evidence for the nonpreferred answer (Kunda, 1990). This is purportedly why, in the trivia paradigm from Windschitl et al. (2010), a desirability bias was only detected when people could not see any difference in the evidence supporting the two options. To summarize, the biased-guessing account suggests that stochasticity, which was part of the marked-card paradigm but not the trivia paradigm, is a key reason for why the desirability bias is much more robust in the former than the latter.

The present research addressed this potential role of stochasticity more directly and thoroughly. It involved two paradigms for which uncertainty was neither purely stochastic/aleatory (as in the marked card paradigm), nor purely epistemic (as in the trivia paradigm). Instead, the uncertainty felt by participants always had some degree of epistemicness but also varied in stochasticity. Testing the desirability bias in paradigms that involve both epistemic and aleatory uncertainty has obvious ecological relevance, since many events in everyday life involve a combination of both forms of uncertainty. For example, expectations about a baseball game depend on perceptions of the quality of the two teams as well as beliefs about unforeseeable random aspects within the game. Similarly, expectations about near-term profits for an agricultural company would depend on knowledge of relevant company characteristics but also considerations about possible stochasticity in weather patterns.

The first paradigm that we developed asked participants to predict the winners of miniature car races that were staged in the lab. In addition to a desirability manipulation (causing participants to prefer one outcome over another), there were two other key manipulations. First, we manipulated the apparent differential in speediness of the two cars in a given race; sometimes the two cars were matched in apparent speediness and sometimes they were mismatched. Second, we manipulated the extent to which stochastic/aleatory uncertainty was relevant to a given race. In high-stochasticity races, the cars raced on courses dotted with sporadic bumps and

obstacles. In lower-stochasticity races, the courses were less obstructed or completely smooth. The other paradigm that we developed involved the same conceptual components—manipulations of desirability, evidence strength, and stochasticity—but the events did not involve car races and were depicted virtually. The paradigm addressed generalizability concerns and afforded improved quantification of the moderator manipulations. It will be detailed later.

Our initial hypotheses were consistent with the biased-guessing account (Windschitl et al., 2010). We expected that the desirability manipulation would impact predictions, but this impact would be moderated by both of the other manipulated factors. First, we expected desirability to have a greater impact on predictions for races in which the two cars in a pair seemed evenly matched in apparent speediness than for races in which the two cars seemed mismatched. Second, and somewhat more importantly, we expected desirability to have a greater impact on predictions for races on high-stochasticity courses than for races on low-stochasticity courses. Again, the biased-guessing account suggests that, even when people have realistic evaluations of available evidence, people tend to provide optimistic predictions when the prediction itself feels largely arbitrary. Hence, when two cars appear virtually the same in their speediness, a participant's prediction would feel like a guess and thus be vulnerable to a desirability bias. Critically, we thought a high-stochasticity situation provided another circumstance in which a prediction could seem arbitrary, leading people to predict that their preferred car will win, even when the other car was slightly better. After all, in high-stochasticity races, a lesser car will sometimes win over a better car.

These two predictions about the moderation of the desirability bias can be related to other perspectives as well. Our prediction about the role of evenly matched vs. mismatched pairs fits broadly with other theoretical perspectives that assume that various biases play a greater role in responses under conditions of vagueness or ambiguity (Bar-Hillel & Budescu, 1995; Dunning,

Meyerowitz, & Holzberg, 1989; Kunda, 1990; Quattrone & Tversky, 1986). Our prediction about the role of stochasticity is also generally compatible with two additional accounts, although directly testing these accounts was not an a priori goal for these studies. The first is Tannenbaum et al.'s (2017) account of how epistemic and aleatory uncertainty relate to judgment extremity. They predicted and found evidence that even when a person's knowledge of evidence is held constant, an emphasis on the epistemic (vs. aleatory) basis of their uncertainty should produce probability judgments that are more extreme. Extending this logic to the car race paradigm, when races are on a low-stochasticity course, people might develop strong views of how likely it is that a better car would win, and they may be reluctant to predict the lesser (but preferred) of the two cars. For a high-stochasticity course, aleatory uncertainty might reduce the strength of views about how likely it is that a better car will win. This, consequently, might leave room for people to feel as though the lesser car (if preferred) has enough of a chance to warrant being the person's discrete prediction as the winner.

The second account that would also seem to anticipate an effect of stochasticity is a motivated probability perception account (Lench et al., 2014). According to this account, when a person has a desire for an outcome, this will enhance their perception of variability in the likelihood of that outcome; they will view the likelihood information—even a specific probability value—as less definitive or more open to interpretation. This enhanced perceived variability is proposed to play a causal role in allowing people to ultimately make a preferred prediction. In one study, participants were given the probabilities of drawing a winning or losing card in a simple card game with either a wide range (10-50%) or a narrow range (20-40%) that had the same midpoint (Lench et al., 2014). Participants' predictions were more apt to show a desirability bias when the probabilities were described with the wider range. Given all this, one might also expect that high stochasticity in the car-racing paradigm would yield a greater

desirability bias. High stochasticity effectively creates another layer of uncertainty beyond the uncertainty resulting from attempting to gauge the speediness of the cars. It essentially requires a person to widen their confidence interval for any implicit probability assessment they make based on their inspection of the two cars before they race.

Overview of the Studies

We conducted a series of studies to address the predictions outlined above. Our initial study, which used the new car-race paradigm, yielded robust evidence of a desirability bias, and the bias was larger when cars were evenly matched. However, contrary to our expectations, the desirability bias was not significantly moderated by the stochasticity manipulation. This null effect for moderation by stochasticity shaped the goals and methods for the subsequent studies. After finding similar results in a follow-up (Study 1.1), we began to wonder if some participants viewed the stochasticity-inducing obstacles in an egocentric way—focusing primarily on how they could disrupt their own car's performance (but not thinking of the competitor car's performance). Therefore, in Study 2 we tested how adding even more stochasticity would influence the levels of desirability bias observed. In Study 3, we used a framing manipulation to test whether we could trigger positive vs. negative construals of stochasticity and thereby alter whether stochasticity would augment or diminish desirability biases. Given no success at detecting a moderating role of stochasticity, we tested the initial hypotheses again in Studies 4a and 4b, using an entirely new paradigm.

A preregistration link can be found in the description of each study. Data sets for these studies are available at https://osf.io/58jpe/?view_only=b11762eababa471599aca2a40e2f5fff. For each study, we report all data exclusions, manipulations, conditions, and measures.

Study 1

Study 1 was an initial test of our hypotheses using the car paradigm. Desirability was manipulated, as were the potential moderators mentioned above (speediness differential and level of stochasticity). In the full design, half the participants provided predictions, and half provided likelihood judgments. However, because this paper focuses on predictions as the dependent variable, our analyses for Study 1 will focus only on the participants asked to give predictions; results for likelihood judgments are in the Supplemental Materials. The preregistration can be viewed at https://osf.io/34se8/?view_only=c8c6bda52a0f4c08884cabf593f2774c.

Study 1 Method

Participants and Design

Fifty-three University of Iowa students (21 males, 32 females, M_{age} = 19.13, SD = 1.37) participated in partial fulfillment of a research component for a course. The design was a 2 (Preferred Team: blue or yellow) x 3 (Pair Type: blue-team faster, equally matched, yellow-team faster) x 2 (Stochasticity: lower or higher) mixed design; the latter two factors were within-subject. There were also two counterbalancing factors, described later. The sample size provides 95% power to detect a medium-sized interaction (f = .25) between preferred team and stochasticity (all power analyses were computed using G-power 3.1.9.2; Faul et al., 2007).[1]

Stimuli (car) Selection

Twenty-four Matchbox-style cars were used. They were selected from a larger set of over 100 cars that varied widely in many attributes—including style, shape, weight, and age. We used informal pilot testing to learn how our participants used cues to judge the speediness of cars.

[1] Given the nature of our design, this interaction test could be recoded as a test of a within-subject main effect of whether participants more often predicted the car from their own team winning on trials within the high-stochasticity condition vs. the low-stochasticity condition.

Cars that looked old or damaged (scratched, dented, worn), or were made of thin lightweight plastic were expected by participants to be slow. Cars that looked new, undamaged, and of better materials were expected to be fast. With this knowledge in mind, we picked 12 pairs of cars that instantiated a key manipulation, as described in the next section.

Car Pairs and Teams

Each session involved the same 12 pairs. One car in each pair was on the "blue team" and one was on the "yellow team" (each car was marked with a colored dot and unique letter). We selected specific pairings of cars such that, for four of the pairs, the blue-team car had better speed cues than the yellow-team car. For another four pairs that we call the "evenly matched" pairs, the two cars looked and felt similar. They differed only in ways that did not have an implication for speed.[2] For the remaining four pairs, the yellow-team car had better speed cues than the blue-team car.

Informal pilot testing and data from the study bear out that the cues we used to construct pairs for this manipulation were predictive of participants' perceptions and of actual race outcomes. Consequently, participants' intuitions about what car would be faster had some validity but could leave room for epistemic uncertainty about which car was actually faster.

Racecourses and Stochasticity

A large, inclined board (4ft x 11ft) was used to stage the races. A given race involved two cars released from a starting line at the top of the incline. Figure 1 shows a picture of a board that was very similar to the one used in Study 1. On the far-right side of the board were two adjoining tracks that were used only for races in the low-stochasticity condition (otherwise removed).

[2] We could have used identical cars for the "evenly matched" category, but it seemed more interesting and less contrived to let the cars vary at least somewhat.

Figure 1. This shows the board used for staging races, similar to how it appeared in Studies 1 and 1.1. The two orange tracks on the far right of the image were used only for races in the low-stochasticity condition; they were removed for high-stochasticity races. The starting line is near the top. The start bar, which is resting at the very top/center, gets moved from that location for use. The finish line was below the bottom edge of the picture. By design, features are sporadically located across the 6 lanes (tape, rough patches of caulk, fuzzy black strips, white round posts), creating the stochasticity needed for the high-stochasticity condition.

These two tracks were smooth and narrowly sized to ensure that the cars run straight. The rest of the board was used for races in the high-stochasticity condition. It was divided into six broad lanes, each containing a haphazard array of bumps and obstacles (e.g., posts, rough strips, and dried caulk affixed to the race surface). The haphazardness was by design—creating the stochasticity by slowing and diverting cars unpredictably.

Procedure

Participants were tested individually. An audio-visual presentation informed the participant that they would be making predictions about a series of races between cars on their team and another team. Race rules were explained. Participants also learned they would receive 10 points every time a car from their team won, and that if enough points were gained, they would receive a choice of snack (e.g., candy bar) from the basket visible in the room.

The experimenter reiterated parts of that information before bringing the participant to a table on which the two teams of cars rested. The two teams could not be seen because each was

under a cover—one labeled with "A" and the other with "B." Participants picked either A or B, and the team under that cover was designated as "their team." The assignment of team to the A cover or B cover was randomized, so although participants picked a team—an action designed to encourage a feeling of team affiliation—their actual assignment to a team was fully random. After the team was selected, both covers were removed and the cars were visible, with pairs arranged in a random order. Participants were handed a clipboard with a questionnaire; the first question asked participants to indicate their team color.

Next were 24 individual trials—each consisting of an examination of cars, a prediction, and a race. The 12 pairs of cars raced twice, once in each of the stochasticity conditions. We counterbalanced whether the first block of 12 races was a low- or high-stochasticity block. At the start of a block of 12, the experimenter previewed the process. For the high-stochasticity block, the two cars in a race would be released down the same lane, determined by a coin flip. This means that participants could not anticipate which bumps or lanes on the board would be relevant for the race. A start bar, shown at the top of Figure 1, was used to space the cars evenly within the lane and ensure a fair release. For the low-stochasticity block, the same two tracks were always used. They were stored on the floor except for the point at which that block of races started, when the experimenter would place those tracks on a smooth portion of the larger board.

To begin each trial, the experimenter put the pair of cars in front of the participant and instructed them to examine the cars before making a prediction. The participant could touch and hold the cars but not push them. The participant then recorded their prediction on a questionnaire sheet that read "For this race, which car do you predict will win? *Car X will win*/*Car Y will win*" – with the option letters corresponding to car letters. The pair was then raced (preceded by a dice roll in the high-stochasticity condition, to determine lane placement). The experimenter would announce the winner and its implication for points. For ties, the race was held again.

After the 24 trials, participants completed miscellaneous measures and a key manipulation check described below. Finally, if their point total exceeded a median point threshold, participants were given their choice of a full-sized candy bar or snack. [3] [4]

Manipulation Check and Other Items/Measures

The most important of the additional measures was a 4-item version of the Epistemic-Aleatory Rating Scale (EARS; Fox et al., 2016—cited in Ülkümen et al., 2016). This scale was developed to assess the extent to which epistemic and aleatory uncertainty is associated with a particular event, as perceived by a respondent. Here, we used it as a manipulation check for how people perceived races on the low- and high-stochasticity courses. First, participants were asked to think about races on the two narrow tracks (the low-stochasticity course) and give 7-point ratings for each of the four items below:

> *As you were making a prediction about races on these tracks, to what extent did the outcome seem like…*
>
> *... it was something that had an element of randomness.*
> *... it would be determined by chance factors.*
> *... it was knowable in advance, given enough information.*
> *... it was something that well-informed people would agree on.*

Fox et al. (2016) designed the first and second items to assess perceptions of aleatory uncertainty. The third and fourth items assess perceptions of epistemic uncertainty. Participants then replied to the four items again, but with regards to the high-stochasticity races.

Exploratory measures included a scale assessing tendencies to check for good vs. bad news. We also asked about gender, age, ACT/SAT scores, interest in candy, care about winning

[3] No extra incentives were promised to motivate people to be accurate in their predictions. Participants seemed to be inherently motivated to be accurate in predictions of this type, and our results were not moderated by self-reported motivation to be accurate. Meta-analysis of past research on the desirability bias suggests that extra incentives (e.g., monetary) do not change the magnitude of the desirability bias (Krizan & Windschitl, 2007; see also Lench & Ditto, 2008; Simmons & Massey, 2012)

[4] Experimenters were not blind to manipulations. However, they were trained to follow standardized steps and to use the same instructional prompts throughout the races, irrespective of conditions and races outcomes.

and being accurate, superstitiousness, and dispositional optimism. Remaining items included open-ended questions (e.g., guesses about study purpose, factors used to make predictions). An overview of these measures and their findings are in the Supplemental Materials.

Study 1 Results

Manipulation Check Findings

Full results for the 3 key manipulation checks are in the Supplemental Materials. They were all successful, as summarized here. First, the EARS data confirmed that participants perceived races on high (vs. low) stochasticity courses to be more a matter of aleatory uncertainty ($p < .001$, $d_{av} = 2.43$) and less a matter of epistemic uncertainty ($p < .001$, $d_{av} = 1.62$). Second, actual race outcomes revealed that car pairs in the three categories performed largely as expected: The yellow team won a minority of the races in blue-faster category (7.6%), about half in the equally matched category (50.9%), and a majority in the yellow-faster category (85.3%). Third, actual race outcomes revealed that races on the high-stochasticity course were, in fact, more stochastic: There were more surprise outcomes on the high-stochasticity tracks (18.0%) than on the low-stochasticity tracks (4.2%) (p-value for the difference < .001).

Main Analyses

We calculated the percentage of times within each pair type that a participant predicted that the yellow-team car would win. We then submitted those percentages to a 2 (Preferred Team: blue or yellow) x 3 (Pair Type: blue-team faster, equally matched, yellow-team faster) x 2 (Stochasticity: lower or higher) repeated-measures ANOVA (see Figure 2 for patterns and Appendix A for a table of means).[5]

[5] Preliminary analyses that included the counterbalancing factor for the order of low- and high-stochasticity races did not reveal findings that substantially changed the main conclusions here or in subsequent studies, so we omitted counterbalancing as a factor in the analyses we describe.

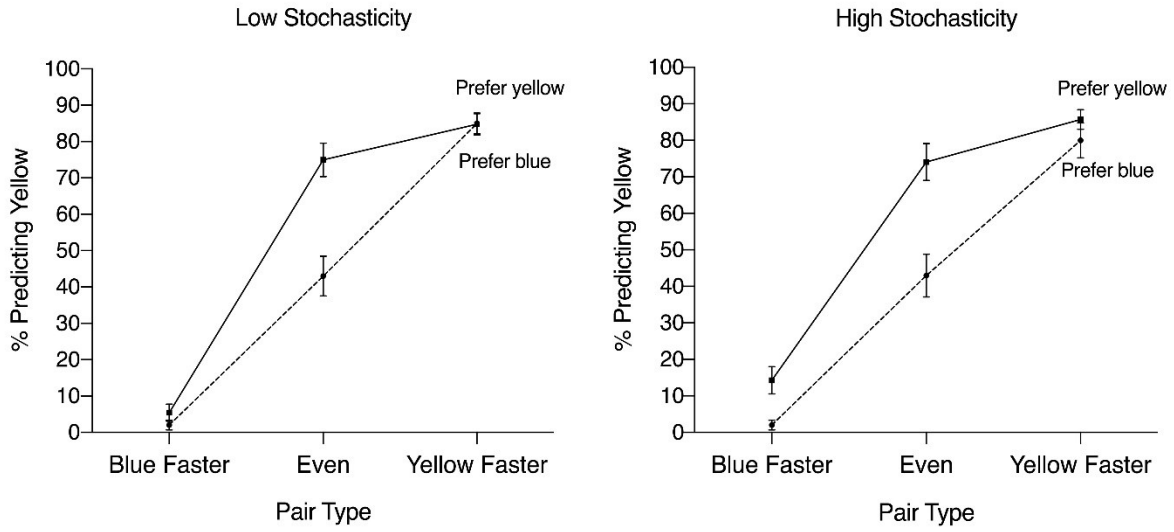

Figure 2. From Study 1, the proportion of races in which participants predicted that the car from the yellow team would win, as a function of stochasticity (i.e., whether the race was on low or high stochasticity track), pair-type (i.e., whether the apparent speediness of the two cars favored the blue team, neither, or the yellow team) and team preference (i.e., whether the participant's team was the yellow or blue team).

Our predictions involved interactions and a simple effect, but we will first discuss the main effects. A significant main effect for team preference constituted an overall desirability bias, $F(1, 51) = 28.27$, $p < .001$, adj $\eta_p^2 = .344$.[6] That is, people were more likely to predict that the yellow team would win when yellow, rather than blue, was their preferred team. A main effect of pair-type reveals that participants did indeed use speed cues for making predictions; they were most likely to predict a yellow win in the races of yellow-faster pairs and least likely in the blue-faster pairs, $F(1.68, 85.78) = 327.11$, $p < .001$, adj $\eta_p^2 = .862$.[7] Unremarkably, the main effect of stochasticity was not significant, $F(1, 51) = 0.13$, $p = .724$, $\eta_p^2 = .002$.

One of our preregistered predictions was that the desirability bias would be larger for races in which the cars in a pair were evenly matched versus when they were not evenly

[6] Throughout the paper, we report the adjusted version of partial eta squared, as established in Mordkoff (2019), except in places where the adjustment would drop the value below zero.
[7] Here and elsewhere, we report Greenhouse-Geisser adjustment for *F* tests when there are sphericity-assumption violations (Greenhouse & Geisser, 1959).

matched. The results were consistent with this prediction. A significant Team Preference x Pair Type interaction revealed the effect of desirability varied across pair types, $F(2, 102) = 12.20$, $p < .001$, adj $\eta_p^2 = .177$. For races with evenly matched pairs, there was a strong desirability bias; people were more likely to predict that the yellow team would win when yellow was their preferred team ($M = 74.55\%$, $SD = 19.69$) than when blue was their preferred team ($M = 43.00\%$, $SD = 24.23$), $t(51) = 5.23$, $p < .001$, $d = 1.43$, 95% CI [19.43, 43.67]. For blue-faster pairs, the tendency to predict the yellow team was slightly larger when yellow was preferred ($M = 9.82\%$, $SD = 14.17$) than when blue was preferred ($M = 2.00\%$, $SD = 4.68$), $t(33.49) = 2.76$, $p = .009$, $d = 0.72$, 95% CI [2.05, 13.59]. For yellow-faster pairs, the tendency to predict the yellow team was not significantly larger when yellow was preferred ($M = 85.27\%$, $SD = 11.31$) than when blue was preferred ($M = 82.50\%$, $SD = 16.14$), $t(51) = 0.72$, $p = .469$, 95% CI [-4.85, 10.39].

We also made the preregistered prediction that level of stochasticity would significantly moderate the desirability bias. However, it did not; the Team Preference x Stochasticity interaction was not significant, $F(1, 51) = 1.58$, $p = .215$, adj $\eta_p^2 = .011$. Figure 2 shows that the desirability bias was about as strong in the low-stochasticity condition as in the high-stochasticity condition. As expected, the simple effect of preference in the high-stochasticity condition was significant, $t(51) = 4.98$, $p < .001$, $d = 1.40$, 95%CI [9.77, 22.97]. But, it was also significant in the low-stochasticity condition, $t(51) = 3.71$, $p = .001$, $d = 1.04$, 95%CI [5.39, 18.06].

No other interactions were significant. This includes the Pair Type x Stochasticity interaction, $F(2, 102) = 1.10$, $p = .336$, adj $\eta_p^2 = .001$, and the 3-way interaction, $F(2,102) = 0.60$, $p = .548$, adj $\eta_p^2 = .004$.

Study 1 Discussion

The detection of a desirability bias in the car-race paradigm is both a conceptual replication and unique extension of findings from the marked-card paradigm (e.g., Budescu & Bruderman, 1995; Irwin, 1953; Marks, 1951; Price & Marquez, 2005). Whereas that paradigm involved events that were purely stochastic and in which evidence was summarized for participants as a numeric proportion (the proportion of marked to unmarked cards), the car-race paradigm involves both aleatory and epistemic uncertainty. Participants had to use various cues to draw inferences about what car was more likely to win, which they clearly did. Also, in support of one of the two predicted moderator effects, when the cues clearly favored one team over another, the impact of desirability was reduced.

However, a key surprise in the results concerned whether people would show more of a desirability bias for racing conditions that were higher, rather than lower, in stochasticity. We had hypothesized that, with the presence of stochasticity, people would feel like their prediction was somewhat arbitrary, allowing them to guess in a way that maintained a preference for optimism. Yet, the magnitude of the desirability bias did not differ between the high and low stochasticity conditions. This cannot be attributed to a general insensitivity to the stochasticity factor because the manipulation-check data from the EARS measure showed a large effect: Participants perceived the high-stochasticity course as involving much more aleatory uncertainty than the low-stochasticity course. The reverse was true for epistemic uncertainty.

Given the potential importance of the null finding for whether stochasticity moderates the desirability bias, we conducted follow-up studies to further test this effect. The first of the studies, Study 1.1, is detailed in the Supplemental Materials, but we summarize the key components here. One difference between Study 1 and Study 1.1 is that we removed the evenly-matched-pairs category. In its place, we included pairs of cars for which the two cars were not

particularly similar in appearance, but differed only in ways that did not obviously indicate which car was faster. As in Study 1, the key manipulation checks were successful, and there was a significant desirability bias—albeit smaller than in Study 1 because of the lack of an evenly-matched-pairs category. Unlike in Study 1, the team Preference x Pair Type interaction was not significant; the desirability bias remained generally consistent in size across the pair types. Most importantly, however, was the that the null findings regarding the stochasticity manipulation replicated. That is, stochasticity did not significantly moderate the desirability bias; the bias was significant in both the low- and high-stochasticity conditions.

Study 2

After seeing the null effects of stochasticity in Studies 1 and 1.1, we began to wonder whether the bumps and obstacles on the high-stochasticity track were being viewed—at least by some participants—as a chaotic circumstance that was likely to hurt their car's chance of winning. This idea is related to the shared-circumstance effect (Moore & Kim, 2003; Windschitl et al., 2003). Work on the shared-circumstance effect and related phenomena suggests that people in a competition are often egocentric in how they consider salient, shared circumstances when estimating their likelihood of winning (Camerer & Lovallo, 1999; Davidai & Gilovich, 2016; Larrick et al., 2007; Moore, 2005; Moore & Kim, 2003; Windschitl et al. 2003; Windschitl et al., 2008). Depending on whether the circumstance is a shared benefit (i.e., generally helps performances) or a shared adversity (i.e., generally hurts performances), thinking about the circumstance egocentrically can bias competitors' optimism either upward or downward. For example, when shared adversities are salient (e.g., rain in a soccer match), players dwell more on how the adversity will affect them than how it will affect their competitor, which thereby reduces everyone's optimism about winning. Regarding the present studies, we questioned whether the racecourse characteristics that created high stochasticity might have essentially been interpreted

by some participants in the way that shared adversities often are. If so, this might be a countervailing influence on optimistic predictions. This reasoning led to Study 2.

For Study 2, we tested how adding even more stochasticity would influence the levels of desirability bias observed in this paradigm. We reasoned that, to the extent that the random bumps and obstacles on a course were being perceived as adversities, and to the extent that people might dwell more on how those adversities might ruin their car's run rather than how it might affect the other car's run, then adding more obstacles would lead to less optimistic predictions. We had two stochasticity conditions: 1) a *moderately-high-stochasticity condition* that involved almost as much stochasticity as the high-stochasticity conditions of the previous studies, and 2) a *very-high-stochasticity condition* that involved even more stochasticity. Unlike our previous preregistered predictions about stochasticity, for this study our prediction was that piling in even more stochasticity (into the very-high-stochasticity condition) would trigger more pessimism, thereby offsetting optimistic predictions. Consequently, we predicted a smaller desirability bias in the very-high-stochasticity condition than in the moderately-high-stochasticity condition. The preregistration for this study can be viewed at https://osf.io/p9xns/?view_only=7d39db10e8af46989eec82dba5689f09.

Study 2 Method

The participants (N = 72; 35 males, 37 females, M_{age} = 19.25, SD = 2.78) were from the same university pool as in Study 1. The design, materials, and procedures were also the same, except as noted. The same car sets were used, and all participants provided predictions (not likelihood judgments). The key difference between the studies was in how the stochasticity manipulation was implemented. We slightly reconfigured the large inclined race board (see Figure S2 in Supplemental Materials). Half of the board held three wide lanes with bumps and obstacles similar to (yet not quite as numerous as) those found in the high-stochasticity condition

of Studies 1 and 1.1. The other half of the board held three wide lanes that had many more bumps and obstacles than both the other side and what was used in the earlier studies. At the start of the study, half of the board was covered; the covered side depended on the counterbalancing of which stochasticity condition was experienced first. Only after the first 12 trials did the experimenter reveal the other half of the board—and hence the other condition. The narrow smooth tracks used in the previous studies were not used in this study. For each race, a die roll was used to determine which of the three lanes on a given side were used for the given race.

A minor change in Study 2 was the inclusion of an exploratory measure at the end of each session. Participants indicated how often each car in a pair would win if raced 100 times.

The sample size of 72 participants provides 99% power to detect a medium-sized interaction (f=.25) between preferred team and stochasticity.

Study 2 Results

Manipulation Check Findings

The manipulation checks returned the expected results. EARS data confirmed that participants perceived races on high (vs. low) stochasticity courses to be more a matter of aleatory uncertainty ($p < .001$, $d_{av} = 1.00$) and less a matter of epistemic uncertainty ($p < .001$, $d_{av} = 0.86$). Actual race outcomes confirmed that car pairs in the three categories performed largely as expected, and the high-stochasticity tracks did create more stochasticity. See Supplemental Materials for full reporting on these checks.

Main Analyses

Figure 3 illustrates the means for the main analysis—an ANOVA like in Study 1 (see Appendix B for means). Starting with the main effects, we found, as predicted, a significant effect of team preference (i.e., the desirability bias), $F(1,70) = 35.72$, $p < .001$. adj $\eta_p^2 = .329$.

The main effect of pair-type was again significant, $F(2,140) = 298.74$, $p < .001$, adj $\eta_p^2 = .807$, and the main effect of stochasticity was again not significant, $F(1,70) = 0.06$, $p = .804$, $\eta_p^2 = .001$.

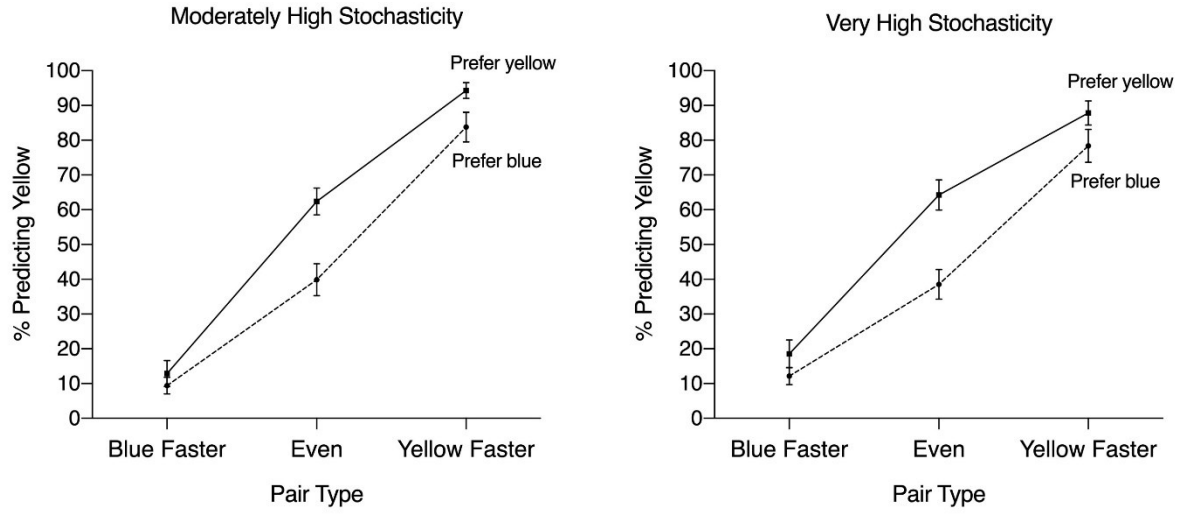

Figure 3. From Study 2, the proportion of races in which participants predicted that the car from the yellow team would win, as a function of stochasticity (i.e., whether the race was on the moderately high or very high stochasticity track), pair-type (i.e., whether the apparent speediness of the two cars favored the blue team, neither, or the yellow team) and team preference (i.e., whether the participant's team was the yellow or blue team).

Consistent with one moderation prediction, the desirability bias was larger for predictions about matched pairs than mismatched ones, as revealed by a significant Team Preference x Pair Type interaction, $F(2, 140) = 5.60$, $p = .005$, adj $\eta_p^2 = .061$.

Contrary to the preregistered prediction about stochasticity, but consistent with the previous studies, the desirability bias did not vary as a function of stochasticity; the Team Preference x Stochasticity interaction was not significant, $F(1, 70) = 0.21$, $p = .650$, $\eta_p^2 = .003$. The simple effects of team preference were significant in both the moderately-high-stochasticity condition, $t(70) = 4.26$, $p < .001$, $d = 1.02$, 95% CI [6.46, 17.81], and very high stochasticity-condition, $t(70) = 4.70$, $p < .001$, $d = 1.13$, 95% CI [8.00, 19.77]. No other interactions were significant. This includes the Pair Type x Stochasticity interaction, $F(2,140) = 1.96$, $p = .145$, adj $\eta_p^2 = .027$, and the 3-way interaction, $F(2,140) = .011$, $p = .897$, $\eta_p^2 = .002$.

Study 2 Discussion

Study 2 provides replication of the finding that level of stochasticity does not moderate the desirability bias. We pushed stochasticity to a very high level in the very-high-stochasticity condition, such that in races involving mismatched pairs, the slower car happened to win 32.99% of the time (rather than 0% of the time, as would happen in the absence of stochasticity). The findings from Study 2 provide no support for the idea that very high stochasticity might be viewed akin to how shared adversities sometimes are—i.e., egocentrically and with the consequence of causing pessimism overall (e.g., Windschitl et al., 2003). However, the findings also do not support the idea that elevated levels of stochasticity allow people to be wishfully optimistic—i.e., show a desirability bias. In short, although there are intuitively plausible reasons why stochasticity could either fuel a desirability bias or its opposite, the results are more in line with the conclusion that stochasticity does not substantially interact with desirability in affecting people's predictions. Study 3 provided an additional test relevant to this issue.

Study 3

In Study 3, we tested the idea that, although people do not seem to generally interpret stochasticity as an optimistic or pessimistic influence, contextual factors might moderate this. Pure randomness in a competition is neutral in the sense that it does not systematically favor or disfavor one's own team over another. However, we reasoned that this neutral quality also means it is pliable; randomness can be described as an opportunity for good things to happen or as chaotic and dangerous. And if people are led to think of the randomness as an opportunity for good things (vs. the opposite), they might focus primarily on what that means for their own team, even though the good things could happen for either team (e.g., Moore & Kim, 2003; Windschitl et al., 2003).

Study 3 was similar to Study 1, except participants saw one of two large and salient signs hanging above the racecourse. One sign read “Opportunity Parkway” and contained images of a rainbow, pot of gold, two checkered flags, and a series of four-leaf clovers. The other sign read “Hazard Parkway” and contained images of a red and white warning triangle around an exclamation point, a road sign with skull and crossbones, and a barbed wire fence. See Figure 4.

Figure 4. From Study 3, the two signs that acted as the framing manipulation. Participants were randomly assigned to see the Opportunity Parkway sign in the positive framing condition, or the Hazard Parkway sign in the negative framing condition.

This is essentially an attribute framing manipulation, and we tested whether this framing manipulation would trigger different reactions to stochasticity (Schneider, Burke, Solomonson, & Laurion, 2005). We hypothesized that under an “Opportunity Parkway” frame as opposed to a “Hazard Parkway” frame, there would be a greater desirability bias. We also hypothesized that frame would moderate the impact of stochasticity on the desirability bias. The preregistration can be viewed at https://osf.io/wk97f/?view_only=ecda92c7ba654bca819b6d7bd0e8e1fe.

Study 3 Method

Study 3 (N = 128, 62 females, 62 males, 4 unreported, M_{age} = 19.45, SD = 1.98) was the same as Study 1, except for the following: 1) The race board for higher-stochasticity races was split into five lanes, and the tracks for the lower stochasticity races always rested on the inclined board rather than being put into place only when needed. 2) Participants were either in an

Opportunity-Parkway or Hazard-Parkway Condition. The relevant sign (described earlier) hung above the racecourse and was seen in the introductory audio-visual presentation. 3) None of the races took place until all predictions were made. We didn't want observed race outcomes to clash with the connotation of the sign. 4) We staged three practice races to ensure participants understood how races were conducted before they made predictions. 5) We used the same car sets as in Studies 1 and 2, except we first inserted replacements for four cars because of wear and performance issues. 6) We used the same exploratory measures as in Study 2 except we added a luck-belief scale and a question asking participants about how bumps influenced their expectations. The sample size of 128 participants provides >99% power to detect a medium-sized interaction (f=.25) between preferred team and stochasticity. It provides 80% power to detect medium-sized interaction between preferred team and frame.

Study 3 Results

The manipulation-check results from the EARS data and outcome data were as expected—paralleling those from Studies 1 and 2. See Supplemental Materials for a full report.

Figure 5 illustrates the means for the main analysis, which was an ANOVA like that from the previous studies but included the framing factor (see also Appendix C for detailed reporting of *M*s and *SD*s and Supplemental Materials for comprehensive table of ANOVA results). Although the overall results of the study were similar to those from the previous ones (see details in next paragraph), the predictions about the new frame factor were not supported, even directionally speaking. The desirability bias was not larger in the "Opportunity Parkway" condition vs. the "Hazard Parkway" condition $F(1, 124) = 0.63$, $p = .427$, adj $\eta_p^2 = -.003$. Even when analyses were restricted to only the high-stochasticity races, there was no greater desirability bias in the "Opportunity Parkway" condition, $F(1, 124) = 2.07$, $p = .153$, adj $\eta_p^2 = .008$.

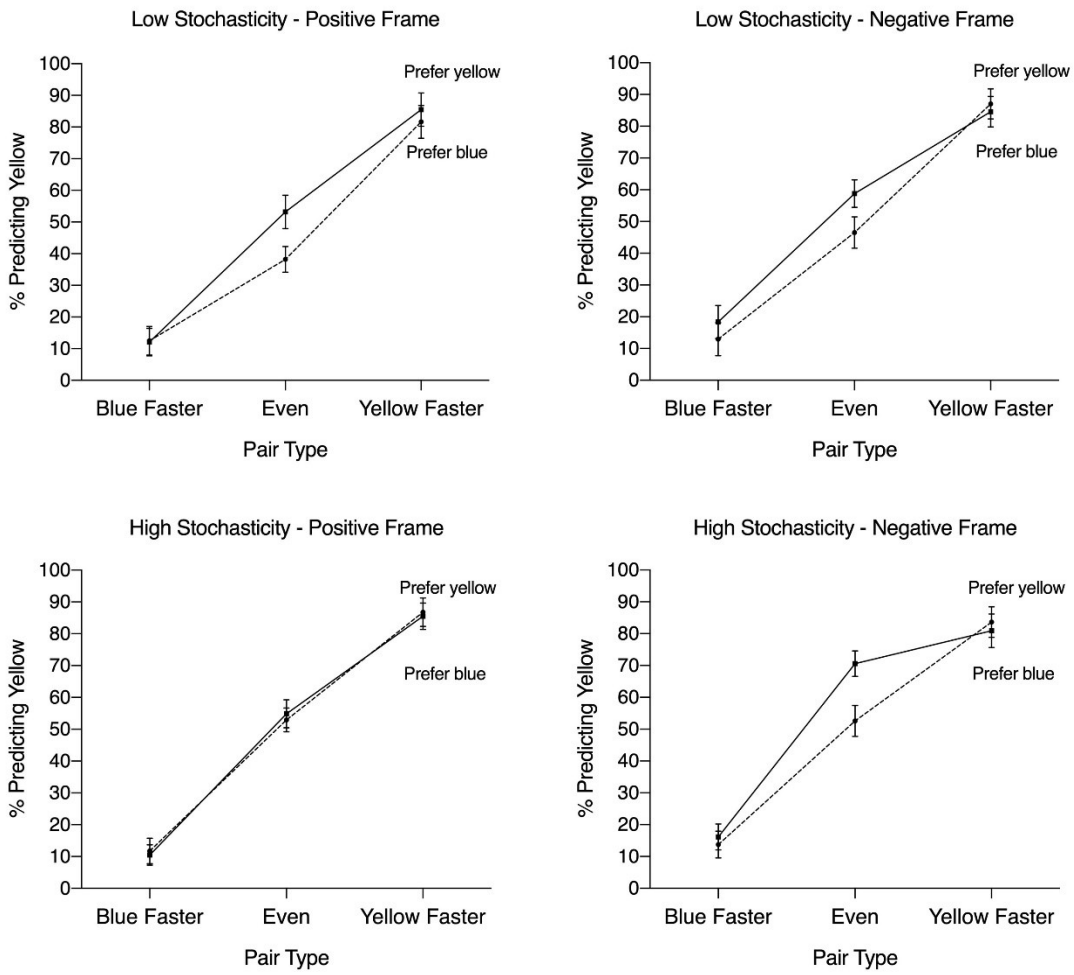

Figure 5. From Study 3, the proportion of races in which participants predicted that the car from the yellow team would win, as a function of framing (i.e., "Opportunity Parkway" or "Hazard Parkway" sign), stochasticity, pair-type, and team preference.

We had also preregistered to look at the effects of desirability and interactions with sign on predictions about only the evenly-matched pairs of cars. For these pairs, the desirability bias was significant, $F(1, 124) = 10.83$, $p = .001$, adj $\eta_p^2 = .073$, but the sign factor did not significantly interact with the desirability bias, $F(1, 124) = 0.87$, $p = .352$, adj $\eta_p^2 = -.001$, nor was it part of a significant three-way interaction, $F(1, 124) = 3.15$, $p = .079$, adj $\eta_p^2 = .017$.

Setting aside the framing factor, the other results in the study were similar to those from the previous ones. The overall desirability bias was significant, albeit smaller than in Studies 1

and 2, $F(1, 124) = 7.16$, $p = .008$, adj $\eta_p^2 = .047$. The reduced size might be due to the fact that participants made all their predictions before seeing any cars race, except in practice trials. The main effect of pair-type was again significant ($p < .001$). The Pair-Type x Desirability interaction was not significant ($p = .07$), but the means took the same distinct pattern as in the previous studies. There was a significant desirability bias for races with evenly matched cars ($p = .001$); the bias was not significant for the non-evenly-matched pairs ($p = .69$). In a more targeted analysis in which pair-type was coded at only 2 levels (evenly matched or non-evenly matched), the Pair-Type x Desirability interaction was significant ($p = .002$). For unknown reasons, the stochasticity main effect neared significance, ($p = .087$). More importantly, the Stochasticity x Desirability interaction was again not significant, $F(1, 124) = 1.02$, $p = .314$, adj $\eta_p^2 = .000$.

Study 3 Discussion

Inspired by findings on shared-circumstance effects (e.g., Moore & Kim, 2003; Windschitl et al., 2003), our new hypothesis for Study 3 had been that the framing of the different signs would set up different effects. With an "Opportunity Parkway" frame, participants would view the uncertainty tied to the bumps on the high stochasticity track as providing an affordance for being optimistically biased in their predictions. With a "Hazard Parkway" frame, participants would view the bumps as an adversity that, although shared by both cars in any pair, would nevertheless result in pessimistic predictions. Despite the prior plausibility of these predictions, the results suggest that reactions to stochasticity were not readily pliable—i.e., not something that was easily nudged toward having optimistic or pessimistic implication for predictions. The framing of stochasticity did little to change the impact of desirability. We verified in a follow-up study that this latter finding could not be attributed to participants being somehow unaware of the signs and their messages (see Study 3.1 in the Supplemental Materials).

Aside from framing results, this study provides another replication of the finding that level of stochasticity does not substantially moderate the desirability bias. With this finding well established across several studies using a car race paradigm, Studies 4a and 4b explored the same issue but using an entirely new paradigm.

Studies 4a and 4b – A New Paradigm

To test the generalizability of the main findings presented thus far, we developed a new paradigm called the *grid-dashing paradigm*. We ran two studies that differed in only one part of the method, which did not meaningfully affect key results, so we describe the two studies together and report results from the combined data set. The lone methodological difference between the two studies involved whether participants received (Study 4b) or did not receive (Study 4a) an explicit warning about the potential for a stochastic element to influence competition outcomes (see ***Procedures and Primary Measures*** below for a description).

Studies 4a and 4b involved the same conceptual variables as Studies 1 & 1.1, with manipulations of evidence strength, stochastic uncertainty, and outcome desirability. The uncertain events about which participants made predictions were part of a game-like task that was presented virtually to online participants. The preregistration's for 4a & 4b can be viewed at https://osf.io/fg6tc/?view_only=69399128c95f4c668b1beb6ab1f06201 and https://osf.io/fxqwn/?view_only=1e4f2a2288f442c1ac1c254c08d5340a.

Study 4a and 4b Method

Overview of the Grid-dashing Paradigm and Procedures

Participants were introduced to a grid-dashing game with multiple rounds involving two robots named "Zuli" and "Remi." There was a desirability manipulation that made a participant prefer that one robot—"their robot"—wins (because winning avoids slowdowns). In each round, a key color was announced, and the robots were said to search the warehouse—each looking for

a grid that had as many squares of the key color as possible. After each robot was said to have secured their own grid, participants briefly saw the two grids and were asked to predict which robot would win. In a no-stochasticity condition, the winner was based on which robot had more squares of the key color over their entire grid. The participant's uncertainty about which robot would win was epistemic; the uncertainty would be completely resolvable if the participant were shown the grid for long enough to count the relevant squares. In a medium-stochasticity condition, the winner was based on which robot had more squares of the key color within a to-be-randomly-determined subsection of the grid. This element of randomness made the outcomes semi-stochastic. Therefore, the participant's uncertainty about which robot would win was both epistemic and stochastic. The strength of evidence was manipulated across rounds by changing the proportions of squares of the key color that appeared in one grid or the other—in some rounds the proportions were almost 50-50 and in others they were more extreme. See below and the Supplemental Materials for more paradigm details and program access.

Participants and Design for 4a and 4b

The participants (N=256 with half in 4a, 144 males, 111 females, 1 not reporting, M_{age} = 38.98, SD = 11.68) were Amazon Mechanical Turk workers secured through CloudResearch (Litman, Robinson, & Abberbock, 2017) and paid $1.20. The design was a 2 (Preferred Robot: Zuli or Remi) x 4 (Pair Proportion Bins: Clearly Favors Zuli, Slightly Favor Zuli, Slightly Favors Remi, Clearly Favors Remi) x 2 (Stochasticity: None vs. Medium) mixed design; the first and third manipulations were between-subject. Study number (4a or 4b) was also a factor in analyses to test for study-based differences in instructions given to participants (see below for more information). There was also a counterbalancing factor, described later. For each sub-study, we preregistered to collect data until reaching the sample size of 128 after exclusions. This sample

size of 256 provided 99% power (or 80% power per study) to detect a medium-sized interaction ($f = .25$) between preferred robot and stochasticity.

Procedures and Primary Measure

Participants were informed that they would be observing a game with a set number of rounds. They were told they needed a robot and to select from two boxes labeled "Robot." After selecting, they were told which of the two robots they had selected—Zuli or Remi. This gave participants the illusion of blindly selecting a robot, although the actual assignment was random.

Participants learned that the robots live in a warehouse full of multi-color grids. They play a game in which, for each round, a game host calls out a *key color* (e.g., "green"), and the robots buzz around the warehouse to each find a grid that has as much of that color as possible. Then they load their respective grids into a scanner to see who wins that round (see Figure 6).

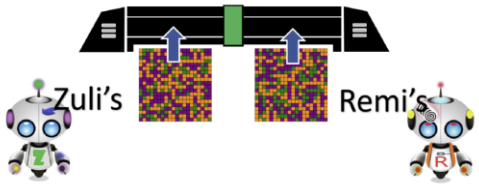

Figure 6. Screenshot from the instruction screens showing grids being loaded in a scanner.

At this point, the instructions in the two stochasticity conditions deviated. In the *no-stochasticity condition*, participants learned that the scanner simply counts the number of squares with the key color on each grid, and the robot with more wins. Figure 7 shows an example.

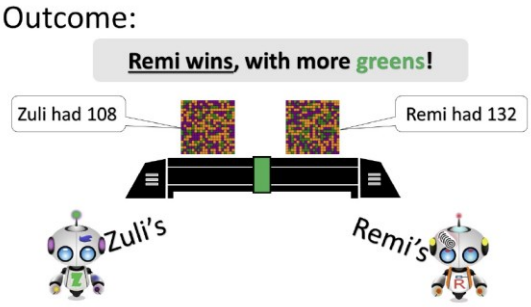

Figure 7. Screenshot showing how a winner is determined in the no-stochasticity condition. The scanner simply counts the number of squares with the key color on each grid (which is green in this example), and the robot with more of that color wins.

In the *medium-stochasticity condition*, participants learned that the scanner first randomly selects 1 of 8 possible subsections to count from each grid. The robot with more squares of the key color in that section wins (see Figure 8).

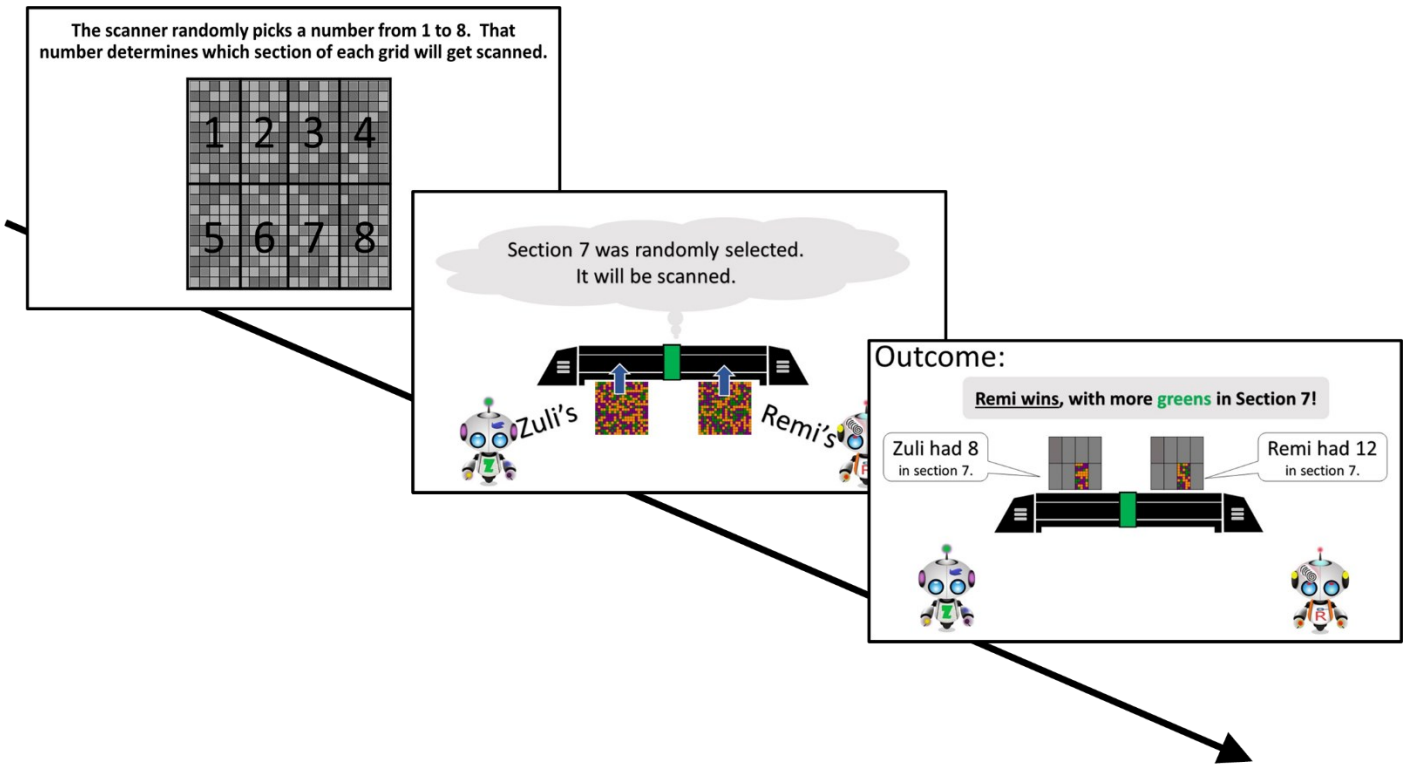

Figure 8. Screenshots showing how a winner is determined in the medium-stochasticity condition of Studies 4a and 4b. The scanner first randomly selects a section number between 1 and 8. Then, within that section on each grid, the scanner counts the number of squares with the key color (which is green in this example). Finally, the robot with more of that color within the section wins.

The only methodological difference between Study 4a and 4b occurred at this point in the instruction sequence. Participants in Study 4b saw a set of screens not included in Study 4a. In the medium-stochasticity condition, participants in Study 4b saw a screen that reviewed the instructions and that explicitly drew two conclusions (counterbalanced in order) about the potential consequences of stochasticity:

> *So, even if your robot's FULL grid has more of the key-color squares than the other robot's grid, your robot might still lose.*
> *And, even if your robot's FULL grid has fewer of the key-color squares than the other robot's grid, your robot might still win.*

Two subsequent questions then reinforced the information. These reinforcement questions required participants to select answers that matched the information that they had just learned about—i.e. how a winner would be determined on each round and what the consequences of stochasticity were. In the no-stochasticity condition, participants in Study 4b saw an analogous review and pair of reinforcement questions. Any participant who answered a reinforcement question incorrectly was informed that their answer was wrong and presented with corrective information.

Next, for all participants, the instructions described information relevant to the desirability manipulation. Participants were informed that every time their robot won, they would briefly see a celebratory screen and immediately move on to the next round. Every time their robot lost, they would have to perform a "loser's task," which involved identifying which of four sentences included a spelling error (and this was accompanied by a sad-faced robot picture). Finally, the instructions notified people that on each round, they would see the two grids for a limited time and be asked to predict which robot would win.

After seeing these instructions and being offered a chance to revisit them, participants began the first of 18 rounds. The first and second of these rounds were practice rounds for which the proportion differences were extreme (one favoring Zuli and the other Remi), making it easy for participants to see who had more of the key-colored squares in their grid. The next 16 rounds were randomly ordered and involved the grid pairs that are described more in the following section. The rounds proceeded as the instructions suggested. For each round, the key color was announced and participants saw the pair of grids for four seconds before being asked "Who do

you predict will win this round?" (Zuli/Remi). As the instructions promised, each time a participant's robot lost, there was a "loser's task" to complete before proceeding. After the rounds, participants completed a modified EARS and other measures (see more information about these near the end of this Method section).

The Grid Pairs, Their Bins, and Outcome Determinations

There were 16 pairs of grids used in the study, which were always presented in a random order. Each grid contained 400 squares. Every pair involved three different colors of squares, one of which was the key color. We used three rather than two colors per pair simply to make it more challenging for participants to determine which grid had more of the key color. There were always 240 key-colored squares in a pair, the distribution of which is described in the next paragraph.

Recall that in the car-race paradigm, we created different pair types that varied in how clearly the evidence favored one car over another in each pair. Analogous to that, we created different pairs of grids that varied in how much the evidence favored one robot over another. The 16 pairs of grids were all unique in this respect, but on an *a priori* basis, we organized them into four bins (see Table S3 in the Supplemental Materials). For pairs in the first bin, the key-colors were distributed in a way that clearly favored Zuli. Specifically, for pairs in that bin, 54% of the key-colored squares were in Zuli's grid and 46% were in Remi's (on average). For pairs in the fourth bin, the distribution favored Remi by the same margin. Based on informal pilot testing, we knew that these proportion ratios allowed people to be generally accurate at detecting which grid from a pair had more of the key color. For the second and third bins, the proportions were less distinct. For pairs in the second bin, the distribution of key colors slightly favored Zuli (with a share of 51%). For the third bid the distribution slight favored Remi (with a share of 51%). Said

differently, the proportions of key colors held by Remi's grids (vs. Zuli's) for pairs in Bin 1-4 were 46%, 49%, 51%, and 54%, respectively.

In the no-stochasticity condition, the announced winner of a round was, of course, directly determined by the proportions. In medium-stochasticity condition, we used programmed schemes to determine outcomes per round, such the frequencies of surprise outcomes would approximate the objectively expected rate (see Supplemental Materials for details).

Modified EARS

We used a modified version of the EARS (Fox et al., 2016) to fit the context of this study. Our wording started with:

> *After you saw the pair of grids on a given round (at the point you were asked to make a prediction), you might still have felt uncertain about the final outcome of which robot would win. To what extent did the following contribute to uncertainty, even after you had seen the grids?*

The first two items were relevant to aleatory uncertainty: "There was still an element of randomness to the final outcome," and "The final outcome would be influenced by chance factors." The other two items were relevant to epistemic uncertainty: "You didn't have enough time to see the grids," and "Your evaluation of the grids felt incomplete."

Other Measures and Exclusion-Check Items

Aside from routine demographic items, there were items that checked on participants understanding of game rules, disappointment about losing, care about being accurate, and any potential confusions. There were also two items that were used for preregistered exclusions. One asked: "Which robot's team were you on?" and the other was an open-ended question asking them to explain why they hoped a given robot would win on a typical round. An incorrect response on the first question or a nonsensical response on the open-ended question led us to

exclude a participant's data from further analysis. The total number of exclusions (not part of our final sample size) was 37.

Study 4a & 4b Results

Manipulation Check Findings

Analyses confirm our preregistered expectation about the EARS data, which focused on perceptions of aleatory uncertainty. The composite scores from the first two items asking about the aleatory nature of the game were significantly and substantially higher in the medium-stochasticity condition (M = 5.55, SD = 1.46) than in the no-stochasticity condition (M = 2.77, SD = 1.92), $t(227.24) = 12.92$, $p < .001$, $d = 1.64$, 95% CI[2.35, 3.19]. We also confirmed that desire was successfully manipulated, with participants rating their hope for a Remi (vs. Zuli) win as higher when assigned to Remi (M = 4.50, SD = 0.66) rather than Zuli (M = 1.56, SD = 0.71), $t(254) = 34.25$, $p < .001$, $d = 4.28$, 95% CI[2.77, 3.11].

Main Analyses

The analysis strategy was analogous to that used for the car-race paradigm. We first calculated the percentage of the times within each bin that a participant predicted a Remi win. These percentages were submitted to an ANOVA. Figure 9 illustrates the observed pattern of means relevant to this analysis; see Appendix D for detailed reporting of Ms and SDs.

First, in a 2 (Study) x 2 (Preferred Robot) x 2 (Stochasticity) x 4 (Proportion Bin) ANOVA with Proportion Bin as repeated, we tested whether there were any consequential differences in the results between Studies 4a and 4b. In short, there were none. The only significant effect that involved the Study factor was a three-way interaction that did not include stochasticity (p = .03 for the Proportion x Robot x Study term). Therefore, we collapsed across

studies for all remaining analyses. For analyses per study, see the Supplemental Materials.

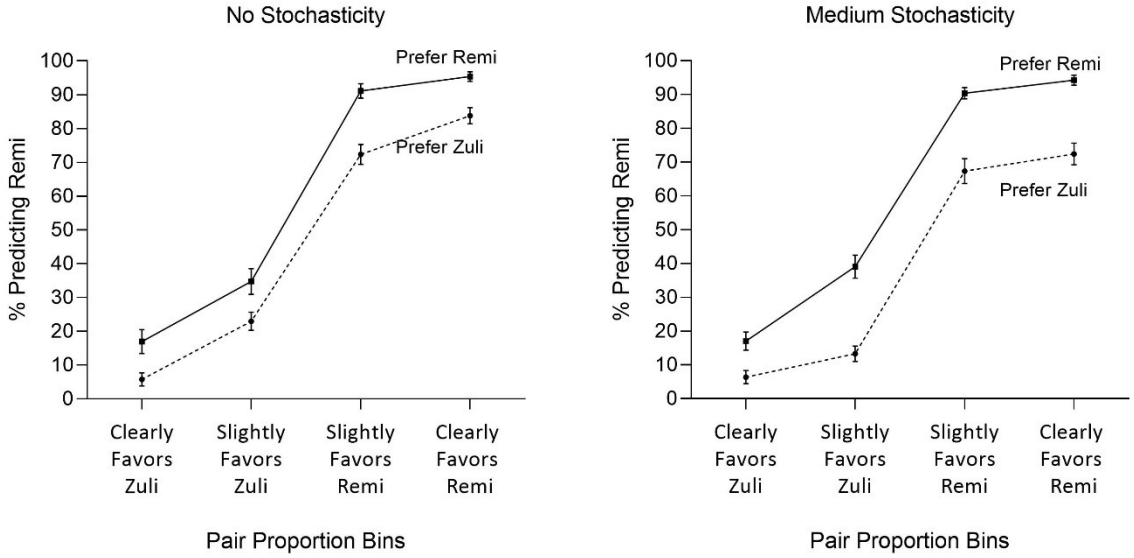

Figure 9. From Studies 4a and 4b (combined), the proportion of rounds in which participants predicted that Remi would win, as a function of stochasticity (i.e., whether the round had no or medium stochasticity), team preference (i.e., whether the participant's team was Remi or Zuli's team), and pair proportion bins (i.e., whether the apparent proportion of squares with the key-color clearly favored Zuli, slightly favored Zuli, slightly favored Remi, or clearly favored Remi). For the pair proportion bins, the average proportion of key colors held by Remi's grids was 46, 49, 51, and 54%, respectively.

Although two of our key predictions involved interactions, we discuss the main effects first. Overall, the main effects replicated the findings from previous studies using the car-race paradigm. Consistent with our preregistered prediction, the results indicated a robust desirability bias – i.e., a main effect in which people were more likely to predict that Remi would win when Remi was the preferred robot rather than Zuli, $F(1, 252) = 127.48$, $p < .001$, adj $\eta_p^2 = .357$. Also, there was a main effect of proportion bin, $F(2.66, 671.44) = 1058.22$, $p < .001$, adj $\eta_p^2 = .807$. This reveals that participants used the proportions of the key-colored squares for making predictions; they were most likely to predict a Remi win for grid pairs with the strongest evidence for Remi (pairs from the 54% bin). Unsurprisingly, the main effect of stochasticity was not significant, $F(1, 252) = 2.45$, $p = .119$, $\eta_p^2 = .010$.

Analogous to findings in the car-race paradigm, we expected that the desirability bias would be more robust when the evidence for the two robots winning (i.e., number of key-colored squares in Zuli vs. Remi's grids) was nearly equal rather than more unequal. In a preregistered analysis, we found that the Robot Preference x Proportion Bin interaction was not quite significant, $F(2.66, 671.44) = 2.50$, $p = .066$, adj $\eta_p^2 = .062$. The effect was small, and the pattern of the effect is not as obvious as in previous studies. For a more targeted analysis, we collapsed the two bins in which the evidence was nearly equal (the 49 and 51% bins) and separately collapsed the two bins in which the evidence was more unequal (the 46 and 54% bins). In an ANOVA that included those two new bins as a repeated measure, the Robot Preference x Bin interaction was significant but still small, $F(1, 252) = 8.54$, $p = .004$, adj $\eta_p^2 = .029$.

The most important preregistered question being tested in this study was whether the level of stochasticity would significantly moderate the desirability bias. The simple effect of preference was significant in both the no-stochasticity, $t(122) = 6.71$, $p < .001$, $d = 1.21$, 95% CI [8.87, 16.30] and medium-stochasticity conditions, $t(130) = 9.21$, $p < .001$, $d = 1.61$, 95% CI [15.72, 24.32]. But a significant Robot Preference x Stochasticity interaction revealed that the preference effect (i.e., the desirability bias) was slightly larger in the medium stochasticity condition, $F(1, 252) = 6.63$, $p = .011$, adj $\eta_p^2 = .022$.[8] Unlike in the cars-paradigm, this result suggests that stochasticity moderates the desirability bias.

Finally, the Robot Preference x Stochasticity x Proportion Bin interaction was not significant, $F(2.66, 671.44) = 1.89$, $p = .137$, adj $\eta_p^2 = .003$, but the Stochasticity x Proportion Bin interaction was significant, $F(2.66, 671.44) = 3.44$, $p = .021$, adj $\eta_p^2 = .009$. The pattern of

[8] This moderation effect was not significantly larger in Study 4b, were there was an explicit reminder about the possible influence of stochasticity, than in Study 4a, $F(1, 248) = 0.08$, $p = .773$, adj $\eta_p^2 < .0001$ for the Study x Robot Preference x Stochasticity interaction.

this interaction makes sense, with participants' predictions being slightly less affected by bin proportions in the medium-stochasticity condition than in the no-stochastic condition.

Study 4a & 4b Discussion

The creation of the grid-dashing paradigm was a success—with all the manipulation checks and main effects working as anticipated. Critically, a robust desirability bias was detected.

But even more important were the results for interaction tests for two types of moderation of the desirability bias. There was evidence for both types. In an analysis that specifically compared the size of the desirability bias across two proportion-bin types, we found that the bias was significantly larger when evidence presented in the grids was more extreme (i.e., clearly favoring one robot) vs. less extreme. And we also found that the desirability bias was significantly larger in the medium-stochasticity trials—where participants could not know for sure which of 8 grid sections would be counted to determine the winner—than in the no-stochasticity condition. This latter result draws a distinction with results from the car paradigm, where stochasticity was not a significant moderator of the desirability bias. Although the two moderation effects were statistically significant, it is also instructive to examine their relative magnitudes. For the moderation of the desirability bias by evidence extremity, the effect size was relatively small ($\eta_p^2 = .033$). The effect size of the moderation by stochasticity was even smaller ($\eta_p^2 = .026$). By comparison, the magnitude of the main effect of desirability was much larger ($\eta_p^2 = .336$).

General Discussion

Understanding how directional motives like outcome desirability might influence expectations is a fundamental issue in the field of judgment and decision making (Hastie, 2001). Among past studies that have manipulated outcome desirability to assess its impact on

predictions, most have relied on the same essential paradigm—the marked-card paradigm—in which the events being predicted are purely stochastic (for review, see Krizan & Windschitl, 2007). Overreliance on paradigms that involve purely stochastic uncertainty has limited our understanding of the desirability bias, given that lay people and professionals often make predictions about events for which their uncertainty is partially or fully epistemic, not just aleatory. Although desirability biases in the fully stochastic, marked-card paradigm tend to be robust, detecting a similar bias in a trivia paradigm that involved only epistemic uncertainty seemed to require a narrow circumstance in which a trivia item was impossibly difficult, such that people could not see any notable differences in the evidence supporting the two options (Windschitl et al., 2010).

In the present studies, we addressed the generalizability gap. We developed two paradigms for studying the desirability bias in predictions about events for which both aleatory and epistemic uncertainty are simultaneously relevant. A desirability bias was detected in both paradigms and in all studies. The meta-analytically combined effect size for the desirability bias across the studies was relatively large and significant ($d = 1.11$, $Z = 5.88$, $p < .001$, 95% CI[0.74, 1.48], $BF_{10} = 178$; see Supplemental Materials for more information about these and other meta-analytic results mentioned below). These findings demonstrate that the desirability bias is not restricted to purely stochastic situations or impossibly difficult trivia questions.

Critically, we also tested for two possible moderation effects. First, in each study we tested whether the desirability bias varied as a function of how balanced vs. extreme the evidence was in supporting one outcome over another. To meta-analytically evaluate the results for this type of moderation, we converted the effect sizes for the relevant interactions to Cohen's *d*. The combined effect size for the moderation effect was statistically significant ($d = 0.35$, $Z = 3.36$, p

<.001, 95% CI[0.15,0.55]), and the Bayes factor indicated that there was strong evidence for the presence (vs. absence) of moderation (BF_{10} = 10.05).

The magnitude of the moderation effect also varied across studies, $Q(6) = 18.33$, $p <$ .005), and closer inspection of this variability across studies reveals a sensible pattern. Recall that in most studies using the car paradigm, the methods included both evenly and unevenly matched categories of car pair (Studies 1, 2, 3, and 3.1). However, for one of the car-paradigm studies and for both grid-dashing studies, there were no evenly-matched pairs (in Study 1.1 the cars within a pair always looked different from each other, and in Studies 4a and 4b there were no 50-50 grids). The effect sizes for the moderation effect were smallest in these latter studies. This makes sense because it essentially means that, in studies where the methodology included a less strong manipulation of evidence extremity, the effect sizes for moderation were smaller. Overall, the moderation of the desirability bias by evidence extremity is consistent with a more general category of findings in which motivated biases are more prominent when evidence is vague or there are not strong differences among options in a choice set (Dunning et al., 1989; Kunda, 1990).

The second key moderator that we tested was the level of stochastic/aleatory uncertainty. Using items from the EARS as a manipulation check, we verified that our manipulations successfully altered participants' perceptions of stochasticity/aleatoryness. Nevertheless, the manipulations did not have the expected effects as a moderator of the desirability bias. We initially expected (and pre-registered) that the desirability bias would be larger for car races occurring on the higher-stochasticity courses versus the lower-stochasticity courses. The rationale for our prediction was based in the biased-guessing account (Windschitl et al., 2010); we thought stochasticity would allow a person to give an optimistic prediction as a guess, even when their preferred car was viewed as slower than the alternative car. After failing to detect the

expected Stochasticity x Desirability interactions in Studies 1 and 1.1, we speculated about a countervailing force related to shared-circumstance effects (Davidai & Gilovich, 2016; Moore & Kim, 2003; Windschitl et al, 2003). Namely, we posited that a relatively high level of stochasticity (i.e., many bumps and obstacles on the racecourse) was being interpreted as an adversity that some people would view egocentrically, thereby reducing optimism about winning (and mitigating any optimistic boost from stochasticity). We ramped up the stochasticity in the very-high-stochasticity condition of Study 2, expecting that this would cause increased pessimism in predictions. However, the stochasticity manipulation again had no significant interaction with desirability. For Study 3, we tested whether high stochasticity was pliable in how it was interpreted and how it would affect the desirability bias. However, framing the stochasticity in a positive light ("Opportunity Parkway") produced no more of a desirability bias than framing it in a negative light ("Hazard Parkway"). Finally, to examine if the absence of a substantial effect of stochasticity was paradigm specific or more generalizable, we designed a new paradigm for use in Studies 4a and 4b. Unlike in the car-racing paradigm, stochasticity was a significant moderator in the grid-dashing paradigm, although the magnitude of the effect was small relative to the size of the desirability bias.

The same findings can also be viewed through a meta-analytic lens. Across the studies from both paradigms, the combined effect size of the moderation by stochasticity was not significant (d = .12, Z = 1.54, p = .123, 95% CI[0.03, 0.28]). The Bayes factor indicated that there was moderate evidence for the absence of the moderation effect (i.e., a null effect) over its the presence (BF_{01} = 3.83). Although a general test for heterogeneity was not significant, $Q(6)$ = 9.32, p = .157), a test of whether the magnitude of the effect was significantly larger in the grid paradigm (d = 0.35) than in the car paradigm (d = 0.12) was significant (Z = 2.28, p = .023). The

Bayes factor was weakly supportive of a moderation effect by stochasticity in the grid paradigm (BF_{10} = 2.42) but strongly supportive of a null effect in the car paradigm (BF_{01} = 11.16).

The Tale of Two Stochasticities?

There are too many differences between the car-racing and grid-dashing paradigms to draw firm conclusions about discrepancies in results between them. Nevertheless, we will speculate about why there was a small moderating effect of stochasticity in the grid-dashing context but not the car-racing context. In the car-racing context, stochasticity was operationalized as unpredictable elements that are presumably construed of as impediments. This is an important way of operationalizing stochasticity because there are many everyday contexts in which unpredictable impediments create added uncertainty about how an event or competition will turn out. In the grid-dashing context, the stochasticity was framed as inherent to the game and presumably was not construed of by participants as an impediment. This too represents many everyday contexts. Our point is that stochasticity that affects performance or events via impediments (versus affecting them via something more neutral) might have different impacts on bias in people's predictions. One way of testing this would be to operationalize the stochasticity manipulation in the car paradigm differently—perhaps by using a random process to determine which of several racecourses of varying lengths (with no obstacles) would be the course used for the race.

Another speculative idea is that participants' underweighting of the role of stochasticity might be related to a general tendency people have to adopt an *inside view* when making a prediction—i.e., attending primarily to case-specific information rather than also attending to or adjusting based on more distributional data (Kahneman, 2011; Kahneman & Tversky, 1979). In the car paradigm, participants had a wealth of case-specific cues about each pair of cars, and those strongly influenced their predictions (alongside an influence of outcome desirability).

Stochasticity, which impacted the distribution of actual race outcomes, might have seemed more like a background feature that just did not get much attention in the prediction process.[9] In the grid-dashing paradigm, where stochasticity did have some influence on people's predictions, perhaps the operationalization of stochasticity made it more salient at the same time that people were attending to case-based information. Specifically, although each pair of grids displayed the cased-based information vividly (i.e., the relevant grid colors), participants in the stochasticity condition were perhaps well aware that only one-eighth of what they were looking at would be relevant (because only one of the eight sections would be the one that is scanned).

Biased-Guessing and Related Accounts

On the basis of our findings, it would be difficult to argue that stochasticity plays a central role in enabling or moderating the desirability bias. In all the studies, the desirability bias was significant even when assessed only within the conditions in which virtually no stochasticity was present. In other words, stochasticity was not at all necessary for finding wishful thinking effects. And, across the studies in the car paradigm, there was no significant evidence of the moderating role of stochasticity. The one bright spot for accounts that propose a moderating effect of stochasticity comes from the grid-dashing paradigm, were the moderation effect was significant, albeit small.

Therefore, these results constrain the purview of, but do not rule out, the three accounts we cited earlier as justification for our moderation-by-stochasticity hypothesis. First is the biased-guessing account, which proposed that people tend to use high stochasticity as an implicit or explicit justification for making an optimistic prediction (Windschitl et al., 2010). The second

[9] There is a normative caveat to point out here. Because we solicited dichotomous predictions, stochasticity should not influence any single prediction if participants are using an appropriate maximization strategy—i.e., always predicting the better of the two cars to win. However, we know from pattern of predictions that matching or some other non-maximization strategies must have been relevant to participants predictions, yet we still did not see an influence of stochasticity.

was related to work by Tannenbaum at al. (2017). We noted that a high-stochasticity condition might emphasize aleatory uncertainty, an emphasis that Tannenbaum et al. had shown can lead to less extreme likelihood judgments. In applying this account here, we speculated that less extreme likelihood evaluations for two cars (or grids) in a pair might offer more opportunity for people to feel as though the lesser car—if a preferred one—has enough of a chance to warrant a prediction as the winner. According to the third account (Lench et al., 2014), when a person has a motivated interest in an outcome, this will enhance their perception of variability in the likelihood of that outcome, which is crucial in allowing people to ultimately make a preferred prediction. Given that this account assumes that wider perceived variability fuels predictions of desired outcomes, one would also expect that high stochasticity in the car-racing or grid-dashing paradigms would have also yielded a greater desirability bias. These current studies were not specifically designed to test these latter two accounts, but we nonetheless find it important to note how the present findings relate to those accounts.

Regarding the biased-guessing account, although we are surprised at how limited the evidence was for stochasticity triggering biased guessing in the present studies, we note that this does not mean that stochasticity and biased guessing fail to play a key role in other findings, like those from the marked-card paradigm. It still might be the case that for events that are *purely* stochastic, the reason why people predict a desirable but improbable outcome will occur is that the randomness of the event allows for arbitrary guesses in trying to specify what will happen. Research on people's tendency to *match* (i.e., exhibit prediction rates that match evidence proportions) rather than to *maximize* (i.e., always predict the more probable outcome) demonstrates people's willingness to occasionally make arbitrary predictions in favor of a lower probability outcome (James & Koehler, 2011; Schulze, James, Koehler, & Newell, 2019).

Notably, the findings on matching vs. maximizing tend to come from paradigms that involve purely stochastic events.

Limitations and Future Research

One limitation of this work is that we did not solicit likelihood judgments or confidence estimates—with an exception in Study 1 (see Supplemental Materials). In past work, the desirability bias was larger when the outcome measures were predictions rather than likelihood judgments (Krizan & Windschitl, 2007; Price & Marquez, 2005; Windschitl et al., 2010), so focusing on predictions seemed sensible. Nevertheless, we cannot assume that people's predictions simply reflect their comparisons of judged probabilities. It would be useful to conduct more tests with the current paradigms, but to also incorporate likelihood judgments or confidence estimates as dependent variables (for discussions, see Bar-Hillel & Budescu, 1995; Bar-Hillel et al., 2008; Bilgin, 2012; de Molière & Harris, 2016; Harris, 2017; Krizan & Windschitl, 2009; Lench et al., 2016; Park et al, in press; Vosgerau, 2010; Windschitl, & Stuart, 2015).

Another potential limitation concerns our desirability manipulation. We manipulated desirability by associating the participant with a team and then linking something desirable with the outcomes of that team. Although the desirability that these methods triggered may pale in comparison to the desirability people might feel about some everyday outcomes (e.g., desirability about one's preferred presidential candidate winning), this does not mean that our methods of manipulating desirability were ineffective or unimportant. The desirability bias was significant in all studies. And, although increasing the desirability of outcomes might (or might not) change the overall level of bias observed, we have no reason to expect that it would interact with our other manipulated factors and thereby change our main conclusions. With all this said, exploring other types of desirability manipulations would be worthwhile in future studies.

Conclusion

These were the first studies we know of that tested how predictions are impacted by outcome desirability, under systematically varied levels of both epistemic and aleatory uncertainty. Two paradigms were established and used to extend the generalizability of the desirability bias. As expected, the desirability bias tended to be larger when evidence for the possible outcomes was relatively balanced. A key hypothesis about the desirability bias being moderated by stochasticity was supported (with a small effect size) in one paradigm but not the other.

References

Babad, E. (1987). Wishful thinking and objectivity among sports fans. *Social Behaviour, 2*(4), 231–240.

Bar-Hillel, M., & Budescu, D. (1995). The elusive wishful thinking effect. *Thinking & Reasoning, 1*, 71-103. https://doi.org/10.1080/13546789508256906

Bar-Hillel, M., Budescu, D. V., & Amar, M. (2008). Predicting World Cup results: Do goals seem more likely when they pay off? *Psychonomic Bulletin & Review*, *15*, 278-283. DOI: 10.3758/pbr.15.2.278

Bilgin, B. (2012). Losses loom more likely than gains: Propensity to imagine losses increases their subjective probability. *Organizational Behavior and Human Decision Processes*, *118*, 203-215. https://doi.org/10.1016/j.obhdp.2012.03.008

Budescu, D. V., & Bruderman, M. (1995). The relationship between the illusion of control and the desirability bias. *Journal of Behavioral Decision Making, 8*, 109-125. https://doi.org/10.1002/bdm.3960080204

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *The American Economic Review, 89*(1), 306-318.

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*, 835–851. https://doi.org/10.1037/pspa0000066

de Molière, L., & Harris, A. J. L. (2016). Conceptual and direct replications fail to support the stake-likelihood hypothesis as an explanation for the interdependence of utility and likelihood judgments. *Journal of Experimental Psychology: General, 145*(4), e13–e26. https://doi.org/10.1037/xge0000124

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082-1090.

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.

Fox, C. R., Tannenbaum, D., Ülkümen, G., Walters, D., & Erner, C. (2016). Skill versus luck and attributions of uncertainty: Validating an epistemic–aleatory rating scale (EARS). Unpublished Manuscript.

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkeboen, & H. Montgomery (Eds.), *Perspectives on Thinking, Judging and Decision Making* (pp. 21-35). United Kingdom: Universitetsforlaget.

Granberg, D., & Brent, E. (1983). When prophecy bends: The preference–expectation link in US presidential elections, 1952–1980. *Journal of Personality and Social Psychology*, *45*, 477–491. https://doi.org/10.1037/0022-3514.45.3.477

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112. https://doi.org/10.1007/BF02289823

Hacking, I. (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference.* London; New York: Cambridge University Press.

Harris, A. J. L. (2017). Understanding the coherence of the severity effect and optimism phenomena: Lessons from attention. *Consciousness and Cognition, 50,* 30-44. https://doi.org/10.1016/j.concog.2016.10.014

Harris, A. J., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, *110*, 51-64. DOI: 10.1016/j.cognition.2008.10.006

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology*, *52*, 653-683. https://doi.org/10.1146/annurev.psych.52.1.653

Howell, W. C., & Burnett, S. A. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance*, *22*, 45-68. https://doi.org/10.1016/0030-5073(78)90004-1

Irwin, F. W. (1953). Stated expectations as a function of probability and desirability of outcomes. *Journal of Personality, 21,* 329–335. https://doi.org/10.1111/j.1467-6494.1953.tb01775.x

James, G., & Koehler, D. J. (2011). Banking on a bad bet: Probability matching in risky choice is linked to expectation generation. *Psychological Science*, *22*, 707-711. https://doi.org/10.1177/0956797611407933

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1979). Intuitive Prediction: Biases and Corrective Procedures. *TIMS Studies in Management Science*, 12, 313-327.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*, 143-157. https://doi.org/10.1016/0010-0277(82)90023-3

Krizan, Z., Miller, J. C., & Johar, O. (2010). Wishful thinking in the 2008 U.S. presidential election. *Psychological Science, 21*, 140-146 .https://doi.org/10.1177/0956797609356421

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin, 133*, 95-121. https://doi.org/10.1037/0033-2909.133.1.95

Krizan, Z., & Windschitl, P. D. (2009). Wishful thinking about the future: Does desire impact optimism?. *Social and Personality Psychology Compass*, *3*, 227-243. https://doi.org/10.1111/j.1751-9004.2009.00169.x

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lagnado, D. A., & Sloman, S. A. (2004). Inside and outside probability judgment. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 157-176). Blackwell Publishing. https://doi.org/10.1002/9780470752937.ch8

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*, 76-94. https://doi.org/10.1016/j.obhdp.2006.10.002

Lench, H. C., & Ditto, P. H. (2008). Automatic optimism: Biased use of base rate information for positive and negative events. *Journal of Experimental Social Psychology*, *44*, 631-639. https://doi.org/10.1016/j.jesp.2007.02.011

Lench, H. C., Smallman, R., & Berg, L. A. (2016). Moving toward a brighter future: The effects of desire on judgments about the likelihood of future events. *Motivation Science, 2*, 33–48. https://doi.org/10.1037/mot0000029

Lench, H. C., Smallman, R., Darbor, K. E., & Bench, S. W. (2014). Motivated perception of probabilistic information. *Cognition*, *133*, 429-442. https://doi.org/10.1016/j.cognition.2014.08.001

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433-442.

Løhre, E., & Teigen, K. H. (2016). There is a 60% probability, but I am 70% certain: Communicative consequences of external and internal expressions of uncertainty. *Thinking & Reasoning*, *22*, 369-396. https://doi.org/10.1080/13546783.2015.1069758

Markman, K. D., & Hirt, E. R. (2002). Social prediction and the "allegiance bias". *Social Cognition*, *20*, 58-86. https://doi.org/10.1521/soco.20.1.58.20943

Marks, R. W. (1951). The effect of probability, desirability, and "privilege" on the stated expectations of children. *Journal of Personality, 19,* 332–351. https://doi.org/10.1111/j.1467-6494.1951.tb01107.x

Massey, C., Simmons, J. P., & Armor, D. A. (2011). Hope over experience: Desirability and the persistence of optimism. *Psychological Science, 22*, 274–281. https://doi.org/10.1177/0956797610396223

Moore, D. A. (2005). Myopic biases in strategic social prediction: Why deadlines put everyone under more pressure than everyone else. *Personality and Social Psychology Bulletin*, *31*, 668-679. https://doi.org/10.1177/0146167204271569

Moore, D. A., & Kim, T. G. (2003). Myopic social prediction and the solo comparison effect. *Journal of Personality and Social Psychology, 85,* 1121–1135. https://doi.org/10.1037/0022-3514.85.6.1121

Mordkoff, J.T. (2019). A simple method for removing bias from a popular measure of standardized effect size: Adjusted partial eta squared. *Advances in Methods and Practices in Psychological Science*, 2, 228-232. https://doi.org/10.1177/2515245919855053

Okada, K. (2017). Negative estimate of variance-accounted-for effect size: How often it is obtained, and what happens if it is treated as zero. *Behavior Research Methods, 49*, 979-987. doi: 10.3758/s13428-016-0760-y

Price, P. C., & Marquez, C. A. (2005). *Wishful thinking in the predictions of a simple repeatable event: Effects on deterministic versus probabilistic predictions.* Unpublished manuscript.

Quattrone, G. A., & Tversky, A. (1986). Self-deception and the voter's illusion. *The multiple self*, 35-38.

Schneider, S. L., Burke, M. D., Solomonson, A. L., & Laurion, S. K. (2005). Incidental framing effects and associative processes: A study of attribute frames in broadcast news stories. *Journal of Behavioral Decision Making*, *18*, 261-280. https://doi.org/10.1002/bdm.500

Schulze, C., James, G., Koehler, D. J., & Newell, B. R. (2019). Probability matching does not decrease under cognitive load: A preregistered failure to replicate. *Memory & Cognition*, *47*, 511-518. doi: 10.3758/s13421-018-0888-3.

Simmons, J. P., & Massey, C. (2012). Is optimism real? *Journal of Experimental Psychology: General*, *141*, 630–634. https://doi.org/10.1037/a0027405

Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2017). Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Science*, *63*(2), 497-518. https://doi.org/10.1287/mnsc.2015.2344

Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology: General, 145*, 1280-1297. https://doi.org/10.1037/xge0000202

Vosgerau, J. (2010). How prevalent is wishful thinking? Misattribution of arousal causes optimism and pessimism in subjective probabilities. *Journal of Experimental Psychology: General*, *139*, 32–48. https://doi.org/10.1037/a0018144

Windschitl, P. D., Kruger, J., & Simms, E. N. (2003). The influence of egocentrism and focalism on people's optimism in competitions: When what affects us equally affects me more.

Journal of Personality and Social Psychology, 85, 389–408. doi:10.1037/0022-3514.85.3.389

Windschitl, P. D., Rose, J. P., Stalkfleet, M. T., & Smith, A. R. (2008). Are people excessive or judicious in their egocentrism? A modeling approach to understanding bias and accuracy in people's optimism. *Journal of Personality and Social Psychology, 95*, 253-273. https://doi.org/10.1037/0022-3514.95.2.253

Windschitl, P. D., Smith, A. R., Rose, J. P., Krizan, Z. (2010). The desirability bias in predictions: Going optimistic without leaving realism. *Organizational Behavior and Human Decision Processes, 111*, 33-47. https://doi.org/10.1016/j.obhdp.2009.08.003

Windschitl, P. D., & Stuart, J. O. (2015). Optimism biases: Types and causes. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making, Vol. II* (pp. 431-456). https://doi.org/10.1002/9781118468333.ch15

Appendix A

Study 1 Means (and SD) Reflecting the Proportion of Races in which Participants Predicted that the Car from the Yellow Team Would Win.

	Low Stochasticity				High Stochasticity				Overall
	Blue Faster	Even	Yellow Faster	Overall	Blue Faster	Even	Yellow Faster	Overall	
	M (SD)				*M (SD)*				*M(SD)*
Prefer Yellow	5.36 (12.47)	75.00 (24.53)	84.82 (15.72)	55.06 (10.72)	14.29 (19.75)	74.11 (26.77)	85.71 (14.32)	58.04 (9.75)	56.55 (7.56)
Prefer Blue	2.00 (6.92)	43.00 (27.50)	85.00 (14.43)	43.33 (12.27)	2.00 (6.92)	43.00 (29.33)	80.00 (23.94)	41.67 (14.02)	42.50 (11.47)
Overall	3.77 (10.28)	59.91 (30.36)	84.91 (14.98)	49.53 (12.81)	8.49 (16.22)	59.43 (31.86)	83.02 (19.47)	50.31 (14.43)	49.92 (11.86)

Appendix B

Study 2 Means (and SD) Reflecting the Proportion of Races in which Participants Predicted that the Car from the Yellow Team Would Win.

	Low Stochasticity				High Stochasticity				Overall
	Blue Faster	Even	Yellow Faster	Overall	Blue Faster	Even	Yellow Faster	Overall	
	M (SD)				*M (SD)*				*M(SD)*
Prefer Yellow	12.86 (22.17)	62.38 (22.81)	94.29 (13.67)	56.51 (10.59)	18.57 (23.75)	64.29 (25.93)	87.86 (20.45)	56.90 (13.02)	56.71 (8.83)
Prefer Blue	9.46 (14.85)	39.86 (27.93)	83.78 (25.83)	44.37 (13.33)	12.16 (15.16)	38.51 (26.08)	78.38 (28.97)	43.02 (20.30)	43.69 (9.60)
Overall	11.11 (18.71)	50.81 (27.80)	88.89 (21.35)	50.27 (13.46)	15.28 (19.93)	51.04 (28.90)	82.99 (25.46)	49.77 (42.61)	50.02 (11.27)

Appendix C

Study 3 Means (and SD) Reflecting the Proportion of Races in which Participants Predicted that the Car from the Yellow Team Would Win.

		Low Stochasticity				High Stochasticity				Overall
		Blue Faster	Even	Yellow Faster	Overall	Blue Faster	Even	Yellow Faster	Overall	
		M (SD)				*M (SD)*				*M (SD)*
Positive Frame	Prefer Yellow	12.10 (24.04)	53.23 (29.40)	85.48 (29.42)	50.27 (13.52)	10.48 (18.00)	54.84 (24.51)	85.48 (23.07)	50.27 (13.52)	50.27 (10.76)
	Prefer Blue	12.50 (26.29)	38.24 (24.02)	81.62 (30.35)	44.12 (11.52)	11.76 (23.22)	52.94 (21.99)	86.76 (26.28)	50.49 (29.65)	47.30 (8.45)
	Overall	12.31 (25.05)	45.38 (27.56)	83.46 (29.74)	47.05 (12.79)	11.15 (20.74)	53.85 (23.06)	86.15 (24.62)	50.38 (31.29)	48.72 (9.66)
Negative Frame	Prefer Yellow	18.38 (30.35)	58.82 (25.29)	84.56 (28.21)	53.92 (10.10)	16.18 (23.74)	70.59 (23.41)	80.88 (30.81)	55.88 (11.70)	54.90 (7.98)
	Prefer Blue	12.93 (28.05)	46.55 (26.50)	87.07 (25.55)	48.85 (11.30)	13.79 (22.74)	52.59 (26.17)	83.62 (26.11)	50.00 (8.90)	49.43 (8.24)
	Overall	15.87 (29.20)	53.17 (26.37)	85.71 (26.83)	51.59 (10.88)	15.08 (23.13)	62.30 (26.13)	82.14 (28.55)	53.17 (10.84)	52.38 (8.49)

Note: Positive frame refers to the condition where participants saw the "Opportunity Parkway" sign, while the negative frame refers to the condition where participants saw the "Hazard Parkway" sign.

Appendix D

Study 4a/4b Means (and SD) Reflecting the Proportion of Gird-dashing Rounds in which Participants Predicted that Remi Would Win.

	No Stochasticity					Medium Stochasticity					Overall
	Clearly Favors Zuli	Slightly Favors Zuli	Slightly Favors Remi	Clearly Favors Remi	Overall	Clearly Favors Zuli	Slightly Favors Zuli	Slightly Favors Remi	Clearly Favors Remi	Overall	
	M (SD)					*M (SD)*					*M (SD)*
Prefer Zuli	5.74 (16.32)	22.97 (22.19)	72.30 (24.33)	83.78 (20.03)	46.20 (9.97)	6.38 (14.98)	13.28 (17.23)	67.32 (28.3)	72.40 (24.62)	39.84 (13.11)	43.25 (11.92)
Prefer Remi	16.95 (26.44)	34.75 (28.24)	91.10 (15.92)	95.34 (10.86)	59.53 (11.97)	16.99 (22.97)	39.1 (29.5)	90.38 (14.09)	94.23 (12.69)	60.18 (11.84)	59.9 (11.85)
Overall	10.71 (22.04)	28.2 (25.64)	80.64 (22.95)	88.91 (17.52)	52.11 (12.73)	12.21 (20.4)	27.46 (27.82)	79.99 (24.47)	84.39 (21.85)	51.01 (16.01)	51.55 (14.5)