Organizational Behavior and Human Decision Processes 111 (2010) 33–47

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

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

The desirability bias in predictions: Going optimistic without leaving realism

Paul D. Windschitl [a,*], Andrew R. Smith [a], Jason P. Rose [b], Zlatan Krizan [c]

[a] Department of Psychology, E11 SSH, University of Iowa, Iowa City, IA 52242, United States
[b] Department of Psychology, University Hall 6516, University of Toledo, Toledo, OH, 43606, United States
[c] Department of Psychology, W112 Lagomarcino Hall, Iowa State University, Ames, IA, 50011, United States

ARTICLE INFO

Article history:
Received 20 August 2008
Accepted 16 August 2009
Available online 12 September 2009

Accepted by Maurice Schweitzer

Keywords:
Desirability bias
Wishful thinking
Optimism
Motivated reasoning
Likelihood judgment
Subjective probability

ABSTRACT

Does desire for an outcome inflate optimism? Previous experiments have produced mixed results regarding the *desirability bias*, with the bulk of supportive findings coming from one paradigm—the classic *marked-card paradigm* in which people make discrete predictions about desirable or undesirable cards being drawn from decks. We introduce a biased-guessing account for the effects from this paradigm, which posits that people are often realistic in their likelihood assessments, but when making a subjectively arbitrary prediction (a guess), they will tend to guess in a desired direction. In order to establish the validity of the biased-guessing account and to distinguish it from other accounts, we conducted five experiments that tested the desirability bias within the paradigm and novel extensions of it. In addition to supporting the biased-guessing account, the findings illustrate the critical role of moderators (e.g., type of outcome, type of forecast) for fully understanding and predicting desirability biases.

© 2009 Elsevier Inc. All rights reserved.

Introduction

Julie, who works at the west branch of a company, gets a stunner from her morning newspaper: The corporate office is closing either the east or west branch, to be announced later. Julie scours the rest of the story looking for clues about which branch will close.

While vacationing in Seattle, Bob is tickled to hear that if the weather conditions are right, the Blue Angels Squadron will perform a flight demonstration near his hotel. He promptly checks several weather forecasts.

Does the fact that Julie wants to keep her job and Bob wants to see the flight demonstration cause them to be biased in an optimistic direction, with Julie expecting that her branch will be safe and Bob expecting the weather to cooperate? In more general terms, the question being raised is whether people tend to show a *desirability bias*—an effect in which the desire for an outcome inflates optimism about that outcome.

Research on the desirability bias (also known as the *wishful thinking effect*) has not produced a consistent set of findings. Perhaps the most widely known studies that have directly tested the desirability bias used a paradigm developed by Marks (1951) in which people are asked to make dichotomous predictions about whether a marked card will be drawn from a deck (e.g., Crandall, Solomon, & Kellaway, 1955; Irwin 1953; Irwin & Metzger, 1966). These studies tend to produce robust desirability biases—that is, participants predict a marked card more often when the drawing of a marked card would result in a monetary gain. However, outside this marked-card paradigm, detection of a consistent desirability bias seems to be more elusive (see Bar-Hillel & Budescu, 1995; Bar-Hillel, Budescu, & Amar, 2008a, 2008b; for review see Krizan & Windschitl, 2007a). To date, relatively little is known about the underlying causal mechanisms that yield desirability biases in the marked-card paradigm, and why these mechanisms have not produced consistent effects outside the paradigm.

Therefore, the overall goal of the present research was to identify the key mechanisms accounting for the desirability biases in the marked-card paradigm, and to investigate the applicability of these mechanisms when key aspects of the paradigm are altered. Addressing these issues is critical for achieving a better understanding of how desires impact people's expectations. In the next sections, we first briefly summarize findings from a recent meta-analysis on desirability effects, before then discussing possible mechanisms that will be tested in our experiments.

Evidence regarding the desirability bias

Krizan and Windschitl (2007a) recently conducted meta-analysis of studies in which the desirability of outcomes was experimentally manipulated and in which the dependent variable was some

[*] Corresponding author. Address: Department of Psychology, University of Iowa, Iowa City, IA 52242, United States. Fax: +1 319 335 0191.
E-mail address: paul-windschitl@uiowa.edu (P.D. Windschitl).

0749-5978/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2009.08.003

form of a forecast. The analysis was also restricted to cases in which respondents did not have an ability to control the outcome; as illustrated in the opening vignettes, such cases are common and important in everyday life. Each study in the analysis was classified into one of four categories, defined by whether the study concerned outcomes that were purely stochastic in nature (e.g., card-draw outcomes) or had some nonstochastic determinants (e.g., competition outcomes), and whether participants were asked to provide a discrete outcome prediction or some form of a likelihood or confidence judgment about an outcome. For each of these four categories, Fig. 1 displays the number of studies that were located for the review and the relevant meta-analyzed effect sizes for the desirability bias. The figure reveals some critical complexities. One cell is entirely empty because no studies in that category were located despite a concerted search. More importantly, studies in the stochastic-predictions cell (upper left) appear to produce large desirability effects, whereas the overall effect in the stochastic-likelihood cell is essentially nil, and the overall effect in the nonstochastic-likelihood cell is small yet significant. In short, one cell stands out—studies in the stochastic-predictions cell have produced desirability biases at a level and consistency that is not matched by other cells. Naturally, there is good reason peer deeper into the studies and effects within that cell.

Of the 14 studies in that cell, 12 involved the classic marked-card paradigm or a close variant (e.g., Crandall et al., 1955; Irwin 1953; Marks, 1951). In the prototypical study, participants are first told the proportion of cards that are marked (which might be manipulated from 10% to 90%) and then are told whether drawing a marked card will mean that they gain or lose some specified amount of money (or points). Participants make predictions about numerous decks before learning anything about the outcomes of the card draws. Of the 12 studies using this marked-card paradigm and soliciting dichotomous outcome predictions, all 12 produced significant desirability biases (see Krizan & Windschitl, 2007a). That is, participants predicted a marked card more often when a marked card would result in a gain rather than a loss. The bias tended to be largest for decks that contained 50% marked cards. Monetary and instructional incentives to be accurate in one's predictions did not tend to reduce the size of the desirability bias in this paradigm. Because findings from the marked-card paradigm have tended to be robust and replicable, they have become the hallmark example of scientific evidence that people are prone to suffer from a desirability bias in their forecasts.

Possible mechanisms

Although numerous studies have produced a desirability bias in the marked-card paradigm, explanations as to how such a bias operates or why it might be greater in some paradigms than in others has tended to be discussed in only a cursory fashion (notable exceptions include Budescu & Bruderman, 1995; Price & Marquez, 2005). In this paper, we explicitly consider four types of accounts for the desirability bias in the marked-card paradigm.

The first account refers to an artifactual explanation that has not been adequately tested. In previous studies using the marked-card paradigm, participants were told by the experimenter what the value of drawing a marked card would be. The same experimenter would also orally solicit a prediction about whether the drawn card would be marked. This procedure is clearly vulnerable to experimenter bias and demand characteristics (e.g., Rosenthal & Fode, 1963). It is easy to imagine that the way in which an experimenter asks the "Will it be a marked card?" question could be different if drawing a marked card would have good rather than bad consequences for the participant, and it is easy to imagine that the respondent might feel some pressure to respond in a certain way when the experimenter is directly posing the questions.

The second type of account, which we will call the *biased-evaluation account*, posits that desire for an outcome biases the way in which the evidence for that outcome is perceived or evaluated. In the broader literature on motivated reasoning, there are several empirical demonstrations that suggest that evidence for a desired conclusion is viewed as stronger or with less skepticism than would the same evidence for an undesired conclusion (for reviews see Balcetis, 2008; Kunda, 1990; Pyszczynski & Greenberg, 1987; Trope & Liberman, 1996; see also Krizan & Windschitl, 2007a). As applied to the marked-card paradigm, the biased-evaluation account (or any variant thereof) would suggest that the stated proportion of marked cards somehow seems larger or more favorable when marked cards are desirable rather than undesirable. Although some readers might question whether a precise and fully relevant statement about the proportion of marked cards (e.g., "4 of the 10 cards are marked") could be differentially evaluated, we note that there have been numerous studies showing that even the most precise numeric information can be viewed as bigger or smaller as a function of context or presentational features (see e.g., Hsee, 1996; Kirkpatrick & Epstein, 1992; Klein, 1997; Peters et al., 2006; Windschitl, Martin, & Flugstad; 2002; Windschitl & Weber, 1999). Therefore, it is theoretically tenable that desire for a marked card could make "4 out of 10" seem larger than it otherwise would.

The third type of account, which we will call the *biased-threshold account*, assumes that the evaluation of the evidence for a marked card is unbiased, but the decision threshold for predicting that a marked card will be drawn is lower when the marked cards are desired rather than undesired. Therefore, when the subjective probability of a marked card is 40%, this might trigger a prediction

	Discrete Outcome Prediction	Likelihood Judgment
Stochastic	14 Studies (12 from Marked-Card Paradigm) 13 Had Significant Effects Overall Odds Ratio: OR = 2.26*	9 Studies 2 Had Significant Effects Overall Effect Size: g = 0.01 (ns)
Non-Stochastic	0 Studies	7 Studies 4 Had Significant Effects Overall Effect Size: g = 0.20*

Fig. 1. A summarized representation of the experimental studies on the desirability bias that met the inclusion criteria for Krizan & Windschitl (2007a) review and meta-analysis. Note: * Indicates that the 95% confidence interval around the population estimate of the standardized mean difference or odds-ratio excluded 0 or 1, respectively.

of a marked card when the card is desirable, but not when it is undesirable. Price and Marquez (2005) described this account and its relation to a signal detection framework. The account is also related to the "Can I/Must I" distinction, which assumes that people require lower evidence standards for drawing palatable conclusions rather than unpalatable conclusions (see Dawson, Gilovich, & Regan, 2002; Gilovich, 1991).

Although the artifactual account, the biased-evaluation account, and the biased-threshold account are all tenable, we are—in this paper—introducing a fourth account. We call it the *biased-guessing account*. The account posits that the desirability bias found in a typical marked-card study comes from an asymmetric approach to guessing an outcome—i.e., guessing more often in an optimistic rather than pessimistic direction. By the term *guess*, we refer to the act of making a prediction that is, in a substantial way, subjectively arbitrary. For example, when a respondent in the marked-card paradigm encounters a deck with five marked and five unmarked cards, he or she is essentially forced to guess. Even when there is an imbalanced deck—say four marked and six unmarked cards—a respondent might still make a guess when generating a prediction, because the outcome seems unknowable from his or her position. The respondent can guess or predict the marked card if he or she sees no contradiction between knowing that there are fewer marked than unmarked cards and anticipating a marked card. After all, a marked card is possible and will indeed occur 40% of the time.[1]

In sum, we have described four accounts of desirability bias in the marked-card paradigm, the first of which refers to a potential artifact. The biased-evaluation account refers to a bias in the way evidence is assessed or evaluated, whereas the biased-threshold and biased-guessing accounts do not. Rather, the latter two can be applied to the decision–prediction processes. The main distinction between the biased-threshold and biased-guessing account is that the biased-guessing account assumes that the key process responsible for the bulk of the desirability bias in outcome predictions is guessing. That is, when people believe that part of their prediction is arbitrary (a guess), they will tend to guess optimistically. When there is no subjectively arbitrary element to their prediction, the biased-guessing account does not predict a desirability bias, but the biased-threshold account would still predict a bias due to a lowered threshold for desirable outcomes.

The present experiments

We believe that the biased-guessing account describes most of what drives the desirability biases that have been detected within the marked-card paradigm. Testing this notion was a key goal for the present research. An interrelated goal was to test the predictions of the biased-guessing account versus other accounts for desirability biases outside the typical marked-card paradigm—namely in cases when the target events are nonstochastic rather than stochastic (corresponding to the nonstochastic-predictions cell in Fig. 1) or in cases when people are asked to provide likelihood judgments rather than discrete outcome predictions (corresponding to the stochastic-likelihood cell). Investigating both of these cases is critical for achieving a more complete understanding of the desirability bias.

The first step in our empirical work was to test for a desirability bias in an improved version of the classic marked-card paradigm, one that allowed us to rule out artifactual accounts of the classic effects described earlier. Having produced a reliable effect in this paradigm (Experiment 1), we then used it as a general platform that we systematically modified for the remaining experiments. In Experiments 2 and 3, we retained many critical features of the paradigm, but we introduced modifications that allowed us to test for a desirability bias when the target outcomes were nonstochastic rather than purely stochastic. As we will discuss in more detail later, guessing is typically less relevant to nonstochastic outcomes than to stochastic ones, so the biased-threshold and biased-guessing accounts differ in their predictions for these two types of outcomes. Then in Experiment 4, we again slightly modified the paradigm from Experiment 1 in order to test for a desirability bias when likelihood judgments rather than dichotomous predictions were solicited. Whereas guessing and decision thresholds can play a role in trying to anticipate the specific outcome of an event, they do not play the same role in how people typically estimate the likelihood of an outcome. Therefore, the biased-guessing and biased-threshold accounts make different predictions from the biased-evaluation accounts for the results of Experiment 4. Finally, in the most direct test of the guessing account (Experiment 5), we used a novel scale-juxtaposition method and special instructions to test whether people would exhibit a desirability bias when specifically encouraged to express their guesses on a likelihood scale.

Experiment 1

Our main goal for Experiment 1 was to test for a desirability bias in a new and improved version of the classic marked-card paradigm—one that would preclude artifactual explanations that are potentially applicable to the effects previously found in the classic paradigm. Like the classic marked-card studies, we presented people with a series of decks, we manipulated the desirability of specific cards (through monetary means), we manipulated the stated frequencies of these cards, and we had participants make dichotomous outcome predictions. Also, although manipulations of accuracy incentives have not had systematic effects on predictions in marked-card studies (see Krizan & Windschitl, 2007a), we wanted to provide some external incentive for accuracy, so participants were told that they would receive a monetary bonus for each accurate prediction.

The most critical change from the classic paradigm was that we made our experimenters blind to the value of drawing a marked card. As mentioned earlier, the experimenters in previous studies were not only aware of the value of a marked card, but they were also responsible for soliciting predictions from participants, which opened a clear potential for demand characteristics. In our Experiment 1, we used a computer for specifying the value of drawing a marked card and recording the participant's prediction—with both the value specification and prediction unknown to the experimenter.

The second notable change in our paradigm concerned the markings on the cards. In the classic paradigm, each card in each deck is either marked (with the same marking, such as an X) or unmarked. This fact might pressure participants to avoid providing the same response in runs of three or more, which could thereby increase the number of nonoptimal predictions and inflate the observed biases. In our paradigm, each card in a deck contained one of two markings, and the markings for one deck were entirely different from the markings for other decks. For example, in one deck, each card was either marked with blue or orange, whereas in another deck, each card was either marked with a triangle or a square. Therefore, rather than being asked whether a marked card will or will not be drawn, the participants were asked whether the drawn card will be one marked with blue or orange, for example. From our perspective as researchers, some cards in a deck were

[1] While researchers might readily identify this as a nonoptimal strategy, studies on probability matching show that some people will occasionally predict the less likely of two outcomes rather than use a maximization strategy for their predictions, in which they would always predict the more likely outcome (e.g., Gal & Baron, 1996; Peterson & Ulehla, 1965).

designated as critical (i.e., contained the mark we designated as critical) and the others were noncritical.

With these two changes, Experiment 1 constituted the most stringent test of the desirability bias in a marked-card paradigm to date.

Method

Participants and design

Fifteen undergraduate students participated in Experiment 1. Participants in the experiments in this paper received credit toward a research exposure component of their Elementary Psychology Course. The main design was a 3 (value of the critical card: +$1, 0, −$1) × 5 (frequency of critical card: 3, 4, 5, 6, or 7 out of 10) within-subject design. There was also one counterbalancing factor, described later. Each participant actually provided 2 data points per cell of this 3 × 5 design, but we collapse all results across this replication factor.

Procedures

The experimenter and participant were seated at opposite sides of a table on which sat a computer screen that faced only the participant. There were 30 decks of cards on a table behind the experimenter. The participant was informed that he or she would start with $3 and that this amount would change depending on the outcomes of card draws in 30 rounds and the accuracy of his or her predictions about those draws. Detailed instructions about how the 30 rounds would proceed included the following information: (1) each deck contained exactly 10 cards, (2) there were two possible markings for cards within a deck, (3) the drawing of a given mark could be worth +$1, $0, or −$1 as specified, (4) for each accurate prediction, the participant would receive $0.25, (5) the experimenter would be the person who drew from the deck, and (6) no outcomes would be revealed until the end of the 30 rounds. After these instructions, there were two practice rounds without feedback, followed by the 30 real rounds, which were randomized separately for each participant.

Each round proceeded as follows. A recorded voice announced the round/deck number. On the screen, the participant viewed value information about the two types of markings for the current deck. For example, some participants read that if a card marked with Z was drawn, they would gain $1, but if a card marked with Y was drawn, they would get $0. (Critical marks always had values of +$1, $0, or −$1; noncritical marks always had a value of $0.) After a short delay, the computer prompted the experimenter to announce the frequencies of the two types of marks and also provide the participant with a sheet of paper stating this information. Returning to our example, the participant would hear and read that four cards were marked with Z and six cards with Y. Finally, the dependent measure would appear on screen: “What is your prediction about which card will be drawn?” After the participant responded (by clicking one of two buttons), the experimenter shuffled the deck, selected an arbitrary card, placed the card face down on the top of the deck, and returned the deck to the back table. At this point, the next round would begin.

At the end of the 30 rounds, participants completed individual-difference measures. (Details about the individual-difference measures and relevant findings are reported in Appendix A for this and all the remaining experiments.) Then the outcomes for the 30 rounds were revealed, and participants were paid, debriefed, and dismissed.

Decks, outcome values, and counterbalancing

All the cards in the experiment were 6.3 × 8.8 cm standard playing cards with the same design on the backside of each card. We assembled 30 decks of 10 cards. A given deck had 3, 4, 5, 6, or 7 cards with a critical mark (on the face side), and the remaining cards had a noncritical mark. For each participant, the critical marks for some decks were imbued with a +1 value, others with a $0 value, and others with a −$1 value. The noncritical marks always had a value of $0. The full crossing of the frequency factor and the value factor required 15 decks, but we also added an internal replication, so 30 decks were used. A between-subject counterbalancing ensured that for a given deck, the critical card was imbued with each of the possible values equally often across participants. Also, the left–right order of on-screen information and response options regarding the critical and noncritical markings was equally balanced across the 30 rounds and within any value condition. Finally, the critical and noncritical markings were always unique to a particular deck; we used various pairs of colors, letters, and shapes for the markings.

Results

Fig. 2 shows the percentage of times that respondents predicted the critical mark as a function of its frequency and value (see Appendix B for the exact means and standard deviations relevant to Fig. 2). The pattern in Fig. 2 is fully consistent with patterns from previous marked-card studies. Of course, the most important element of this pattern is how the desirability of a critical mark influenced participants’ tendencies to predict it. Overall, when a critical mark was desirable (i.e., it would yield +$1 whereas the noncritical mark would yield $0), participants predicted the critical mark 68.7% of the time. When a critical mark was neutral (i.e., both it and the noncritical mark would yield $0), participants predicted it 50.0% of the time. When a critical mark was undesirable (i.e., it would yield −$1 whereas the noncritical mark would yield $0), participants predicted it only 38.0% of the time.

For inferential tests, we scored the prediction of a critical and noncritical card as a 1 and 0, respectively. These scores were then averaged, within subjects and cells, to create composite scores, which were then submitted to ANOVAs and *t*-tests. A repeated measures ANOVA on these composites revealed a significant desirability or value effect, $F(2, 13) = 9.69$, $p < .01$. *t*-Tests also revealed that the rate of selecting the critical card was greater in the +$1 condition than in the $0 condition, $t(14) = 4.52$, $p < .001$, and greater in the $0 condition than in the −$1 condition, $t(14) = 2.74$, $p < .05$.

The ANOVA also revealed a significant effect of frequency, $F(4, 11) = 57.51$, $p < .001$. That is, people were sensitive—albeit

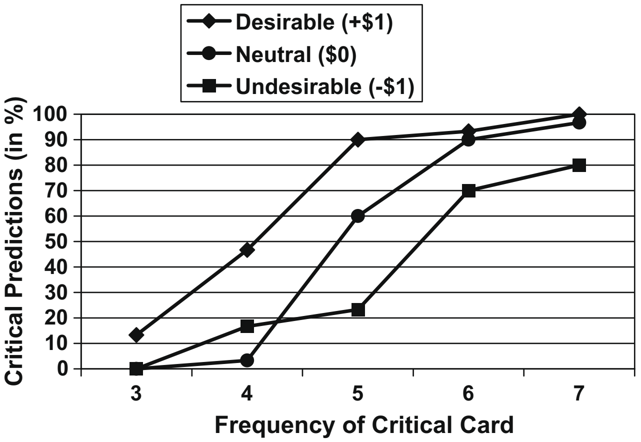

Fig. 2. From Experiment 1, the percent of trials on which the critical card was predicted as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1).

normatively undersensitive—to the frequency of the marked card (see sloping lines in Fig. 2).

The Desirability × Frequency interaction was also significant, $F(8, 7) = 14.54$, $p < .01$. In Fig. 2, the desirability bias appears to be larger when the frequency of critical marks is 5 rather than 3, 4, 6, or 7. For each participant, we calculated a composite of the desirability bias within each frequency condition by subtracting the rate of selecting the marked card in the −$1 condition from the same rate in the +$1 condition. A series of paired t-tests confirm that the magnitude of the desirability bias was indeed larger in the five-card condition than in any other condition (all $p < .05$).

Finally, we should also note that the main effects in Fig. 2 were not driven only by a small subset of participants. In fact, of the 15 participants, 12 exhibited results that were directionally consistent with the desirability bias (i.e., they predicted more critical cards in the +$1 condition than the −$1 condition), and the remaining three exhibited neutral results. All 15 participants exhibited results consistent with a sensitivity to frequency information.

Discussion

Experiment 1 detected a robust desirability effect in a new and improved paradigm. Because the experimenter was unaware of the value of a marked card in a given deck and because participants' responses were not immediately visible to the experimenter, this paradigm rules out the possibility that experimenter bias accounts for previous results and it minimizes the potential role of demand characteristics. The paradigm also removed some pressure on participants to avoid providing the same response on consecutive decks—a pressure that might have augmented non-normative responding. Finally, because it provided a successful demonstration of the desirability bias, Experiment 1 and its paradigm can serve as a platform for examining whether the desirability bias observed in the marked-card paradigm extends beyond its usual confines. This is important for reasons of external validity, but it is also important for determining which of the other three accounts we mentioned earlier best explains the observed bias.

Experiment 2

In Experiment 2, we tested whether the desirability bias would extend to events that were nonstochastic rather than purely stochastic. Such events are common in everyday life, yet the Krizan and Windschitl (2007a) review found no experiments that tested for wishful thinking in outcome predictions regarding such events (the now empty cell of Fig. 1). The experiment was a mixed design, with some participants in a card (stochastic) condition and some in a trivia (nonstochastic) condition. The card condition was identical to Experiment 1. The trivia condition was constructed to be as similar as possible to the card condition, except for the nonstochastic nature of the questions that participants encountered. For this trivia condition, we constructed a list of 30 trivia questions that each had two possible responses. For example, "What animal makes a louder noise?—blue whale or lion." Participants were asked to predict the factually correct option, and they were promised $0.25 for every accurate prediction.[2] Recall that for the card paradigm, we arbitrarily deemed one of the two markings from a deck as the critical one, and if that card happened to be the drawn card, the participant would receive +1, $0, or −$1 (regardless of their prediction). Similarly, for the trivia condition, we arbitrarily deemed one of the two options from a trivia question as the critical one, and if that option happened to be the factual option, the participant would receive +1, $0, or −$1 (regardless of their prediction). For example, some participants were told that if the blue whale was louder (i.e., if it was the factual option), they would win $1, but if the lion was louder, they would get $0. Hence, these participants would desire that blue whale was the factual option because this would yield a dollar (irrespective of their prediction). Of course, they would be wise to ignore this desire when formulating their prediction; in order to maximize their chances of gaining the $0.25 accuracy reward, they should base their prediction—as best they can—on their relevant knowledge of blue whales and lions. In short, the card and trivia conditions had important parallels and both tested whether people would tend to predict the more desirable of the two outcomes.

Not only does the trivia condition in Experiment 2 explore the generalizability of the findings from the marked-card paradigm, but it also helps distinguish among the biased-evaluation, biased-threshold, and biased-guessing accounts. If biased evidence evaluation was the key mediator of the effect in the card paradigm, we should see similarly robust effects in the trivia paradigm of Experiment 2. In fact, the evidence that a person might consider seems much more malleable in the trivia paradigm than in the card paradigm, so one might even expect a larger desirability bias in the trivia paradigm if biased evidence evaluation is a key driver of the desirability effect in the marked-card paradigm. Similarly, if biased decision thresholds were key in producing the effects in the card paradigm, the same robust effects should be observed in the trivia paradigm. That is, if biased predictions occurred because less evidence (or lower evidential support) was required to trigger predictions of desired outcomes than undesired outcomes, then this same differential-threshold process has full potential to occur in the trivia paradigm.

However, if guessing was a key process in producing the effect in the card paradigm, we would expect substantially smaller effects in the trivia paradigm. Recall that, by the term guessing, we are referring to the act of making a prediction that is, in a substantial way, subjectively arbitrary. In the card paradigm, this would clearly occur for decks in which the proportions of the critical and noncritical marks are exactly equal. It could also occur when they are unequal; a respondent can still guess or predict that a minority mark will be drawn if he or she sees no clear contradiction between such a guess and knowing that there are fewer of those marks than the other marks. In short, some people might feel quite comfortable with the following logic: "I know there are only four cards with X, but I think it will be X on this draw." Now consider guessing in the trivia paradigm. If a person evaluates the evidence for the two possible outcomes (e.g., his or her relevant knowledge of lion and blue whale) and sees absolutely no imbalance in the evidence, he or she would guess and might then be vulnerable to a desirability bias (i.e., guessing blue whale because it is more desired as a factually true outcome). However, if there is any imbalance—if the person's knowledge leans slightly in one direction—the person would be compelled to make the guess or prediction that goes in the same direction as their knowledge. If not, this would present an internal inconsistency in reasoning—e.g., "My knowledge points to lion, but I'm going to say blue whale." Thus, we suggest that a tendency to keep one's predictions consistent in direction with one's knowledge will preclude a desirability bias whenever one's knowledge supports one trivia outcome more than another.

For the trivia questions used in Experiment 2, we selected questions that people would be unlikely to have previously learned the correct answer but would have at least cursory background knowledge that they could use as a foundation for making a prediction. We presumed that people's knowledge for the two possible outcomes of a question would rarely support both in a perfectly equal fashion, so we expected the desirability bias to be generally small in the trivia condition.

[2] We use the term *prediction* to refer to participants' task of indicating the factually correct answer, even though the event on which the answer is based is already determined. The question of whether it is important that these "predictions" are really postdictions is addressed later in the paper.

In summary, the biased-evaluation and biased-threshold accounts predict that any desirability bias in the card condition should readily extend to the trivia condition. However, our biased-guessing account predicts that although there should be a replication of the desirability bias in the card condition, it should not extend with much robustness to the trivia condition.

Method

Participants and design

Our plan was to randomly assign our participants to either the card or trivia condition, which we did for the first 30 participants. After analyzing these data and discovering a healthy desirability bias in the card condition but null effects in the trivia condition, we decided it was important to rule out a Type II error in the trivia condition by substantially increasing the sample size. Hence, there were a total of 15 participants in the card condition and 39 in the trivia condition.

The card condition involved the same counterbalancing and within-subjects factors as Experiment 1. The trivia condition included the value factor (+$1, 0, −$1) and a counterbalancing factor (described below). Although there was no frequency factor for the trivia condition, we did construct the question set such that the critical options would range from somewhat weak (analogous to a case in which there were few critical marks in a deck) to somewhat strong.

Procedures

The procedures in the card condition were identical to those used in Experiment 1. In the trivia condition, the procedures were designed to be as similar or parallel as possible. The participant was informed that he or she would start with $3 and that this amount would change depending on the dollar values associated with factual answers to 30 trivia questions and on his or her prediction accuracy. Detailed instructions about how the 30 rounds would proceed included the following information: (1) for each question, the participant would see two options and be asked to predict which was the factual option, (2) the computer would randomly assign a +$1, $0, or −$1 dollar value to each option, (3) for the factual option, the participant would win or lose the assigned amount regardless of his or her prediction, (4) for each accurate prediction, the participant would receive $0.25, and (5) no outcomes would be revealed until the end of the 30 rounds. After these instructions, there were two practice rounds without feedback, followed by 30 real rounds, randomized separately for each participant.

During each round, the participant viewed, on screen, value information about the two options for the trivia question—even before the question was revealed. For example, participants were told that if blue whale was the factual option, they would win $1, but if lion was the factual option, they get $0, regardless of how they responded. We had arbitrarily and surreptitiously deemed one option as the critical one, which was assigned a value of +$1, $0, or −$1; noncritical options always had a value of $0. After a short delay, the experimenter provided the participant with a sheet of paper stating the question (e.g., “What animal makes a louder noise?”). Finally, the dependent measure (“What is your prediction about the true answer?”) would appear on screen with the relevant response buttons below it.

At the end of the 30 rounds in both the card and trivia conditions, participants completed individual-difference measures (see Appendix A). Participants in the trivia condition also read each trivia question again and provided subjective probability estimates for both items (critical and noncritical) of a given question (adding to 100%). The questionnaire that solicited these subjective probability estimates did not list the outcome values (+$1, $0, −$1). Then participants received accuracy feedback, were paid, debriefed, and dismissed.

Trivia questions, outcome values, and counterbalancing

Five example trivia questions can be found in Appendix C. The critical options for the 30 questions were always assigned a value of +$1, $0, or −$1 (for half the questions, the option that was deemed to be critical was also the factually correct option). Counterbalancing ensured that for a given question, the critical option was imbued with each of the possible values equally often across participants. Also, the left–right order of on-screen information and response options regarding the critical and noncritical options was equally balanced across the 30 rounds and within any value condition.

Results

For the card condition, the results are remarkably similar to those of Experiment 1—see Fig. 3 or see Appendix B for exact means. Overall, the rate at which the critical mark was predicted was 66.7% in the +$1 condition, 54.0% in the $0 condition, and 40.7% in the −$1 condition. For inferential tests, we used the same coding and analyses as described for Experiment 1. A repeated measures ANOVA on the composite scores revealed a significant desirability effect, $F(2, 13) = 7.66$, $p < .01$. *t*-Tests also revealed that the rate of predicting the critical card was greater in the +$1 condition than in the $0 condition, $t(14) = 2.74$, $p < .05$, and greater in the $0 condition than in the −$1 condition, $t(14) = 1.96$, $p = .07$. As in Experiment 1, there was also a robust frequency effect, $F(4, 11) = 93.38$, $p < .001$. Finally, the overall Desirability × Frequency interaction was not significant, $F(8, 7) = 1.53$, $p > .20$. However, in a set of paired *t*-tests that more directly examined the desirability bias across frequencies, the desirability bias was larger when there were five critical cards in a deck than when there were 3 or 7 critical cards (both *ps* < .05; with the remaining tests nonsignificant).

Did the robust desirability bias that was detected for the card condition also extend to the trivia condition? As expected, the answer was no. Overall, the rate at which the critical option was predicted was 57.9%, 54.0% and 52.3% in the +$1, $0, and −$1 conditions, respectively. A repeated measures ANOVA on the composites scores for these conditions revealed that desirability did not have a significant effect on predictions, $F(2, 37) = 1.17$, $p > .20$. An overall ANOVA produced a significant Desirability (+$1 or −$1) × Event Type interaction, which confirmed that the desirability bias

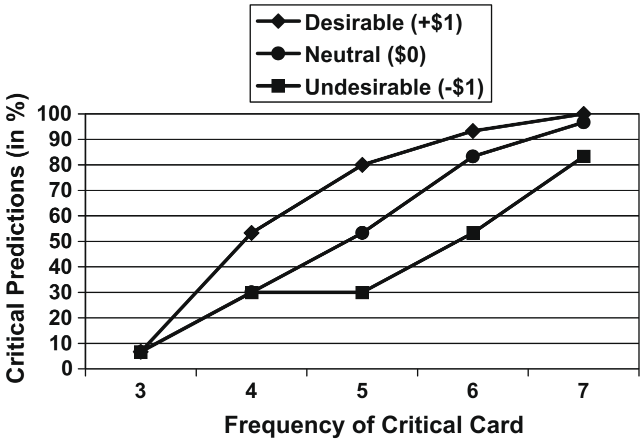

Fig. 3. From the card condition of Experiment 2, the percent of trials on which the critical card was predicted as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1).

was larger in the card condition than in the trivia condition, $F(1, 52) = 7.36$, $p < .01$.

Although we did not systematically manipulate the strength of the critical trivia options, they did vary from somewhat weak to somewhat strong. We therefore organized the data on this basis and produced a graph of the trivia results that is analogous to the graph of the results from the card condition (see Fig. 4). To create Fig. 4, we first ordered the 30 trivia questions according to the strength of the critical items (based on the sample's mean probability estimates for these critical items—collected at the end of the experimental session). Then we split the questions into five groups and plotted the prediction rates as a function of value condition. As seen from Fig. 4, evidence for a desirability bias is minimal, at best, regardless of whether the critical options were generally weak (far left), moderate (middle), or strong (far right).

Finally, to check whether the subjective probability estimates that were collected at the very end of the session were affected by the earlier manipulations of desirability, we conducted a repeated measures ANOVA on the estimates for the critical items. There was no evidence of a desirability bias on the subjective probability estimates, $F(2, 36) = 1.01$, $p > .20$. The average estimate was 51.8%, 49.6%, and 51.8% in the +$1, $0, and −$1, conditions, respectively.

Discussion

Although there was a robust replication of the desirability bias in the card condition, this did not extend to the trivia condition. For the sake of comparison, the difference in rate of predicting the critical mark/option when in the +$1 condition versus the −$1 condition, which is one metric of the desirability bias, was 30.7% in Experiment 1, 26.0% in the card condition of Experiment 2, and only 5.6% in the trivia condition of Experiment 2. This pattern is consistent with our proposal that the desirability bias in the marked-card paradigm is primarily driven by biased guessing, rather than by biased evidence evaluation or biased decision thresholds. Had biased evaluations or thresholds driven the results in the card paradigm used in Experiments 1 and 2, we would have seen at least similar levels of bias in the trivia conditions.

Experiment 3

The biased-guessing account does not assume that predictions about trivia questions (or other questions involving epistemic uncertainty) are always invulnerable to a desirability bias. We believe that if a person considers the evidence for the two possible outcomes and sees absolutely no imbalance in the evidence, his or her guess for a prediction is open to a desirability bias. For the questions we created in Experiment 2, we assumed that most participants would typically have some background knowledge that would at least point them in a tentative prediction direction, thereby precluding a role for biased guessing. However, in Experiment 3, we reran a trivia condition and included a subset of questions that were specifically designed to leave participants with the sense that the two options were equally plausible. These questions, which can be colloquially described as ridiculously difficult, are precisely the type of questions that we believe are vulnerable to biased guessing.

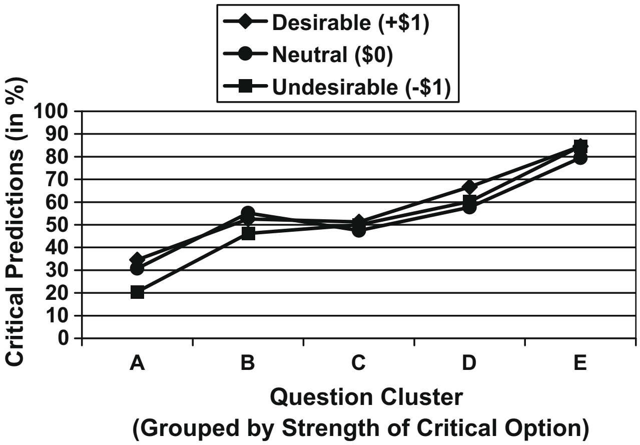

Fig. 4. From the trivia condition of Experiment 2, the percent of trials on which the critical item was predicted as a function of the overall strength of the critical item (weakest in Question Cluster A, strongest in Question Cluster E) and whether it was desirable (+$1), neutral ($0), or undesirable (−$1) for the critical items to be factually correct.

Method

Thirty undergraduates participated. The design and procedures were identical to those used in the trivia condition of Experiment 2. The only change was that we replaced 12 of the 30 questions with new questions that were designed to have options that participants would view as essentially indistinguishable. For example, one question was: Who invented the lava lamp?—Andrew Jenkins or Edward Walker. Another was: The first police force was established in Paris in what year?—1676 or 1667. Additional examples can be found in Appendix C.

Results

Consistent with our main prediction, there was a robust desirability bias detected for the new questions—as revealed by an ANOVA on the composite scores for the +$1, $0, or −$1 conditions, $F(2, 28) = 6.18$, $p < .01$. Among these new questions, the overall rate at which the critical mark was predicted was 48.3%, 42.5%, and 30.8% in the +$1, $0, and −$1 conditions, respectively.[3] The old questions again yielded a nonsignificant effect, $F(2, 28) = 1.56$, $p > .20$. Among the old questions, the overall rate at which the critical mark was predicted was 61.7%, 53.3%, and 52.8% in the +$1, $0, and $−1 conditions, respectively. We did not eliminate the most difficult question from our old set, so the modest (yet nonsignificant) effect in the old set is not surprising. Therefore, it is also not necessarily surprising nor problematic that the interaction between question type and desirability (+$1 or −$1) was not significant, $F(1, 29) = 1.34$, $p > .20$.

Fig. 5 depicts the results across both the new and old items, using the same grouping scheme as in Fig. 4. The 12 new questions tended to fall in the second and third groups of questions on the figure, which is precisely where there is a clear separation in the prediction rates for the +$1, $0, and −$1 conditions.

Finally and not surprisingly, a repeated measures ANOVA revealed no evidence of a desirability bias on the subjective probability responses that were collected late in the experimental sessions, $F(2, 28) = 0.13$ $p > .20$. This was true even when the analysis was restricted to the new questions, $F(2, 28) = 1.65$ $p > .20$. Also, consistent with our intent of using difficult questions, participants tended to respond with "50%" for critical options on the new questions (specifically, 71% of the time).

[3] Readers might wonder why prediction rates for the critical items on the new questions were below 50%. Although we counterbalanced whether a critical item served in the +$1, $0, or −$1 conditions, the determination of which answer/item for a question would be the critical rather than noncritical item was done randomly when designing the experiment and was the same for each participant. Therefore, the sub-50% prediction rates simply reflect the fact that the items randomly deemed to be critical were slightly less attractive as guesses than were the items deemed as noncritical. This is not a concern for any of our conclusions.

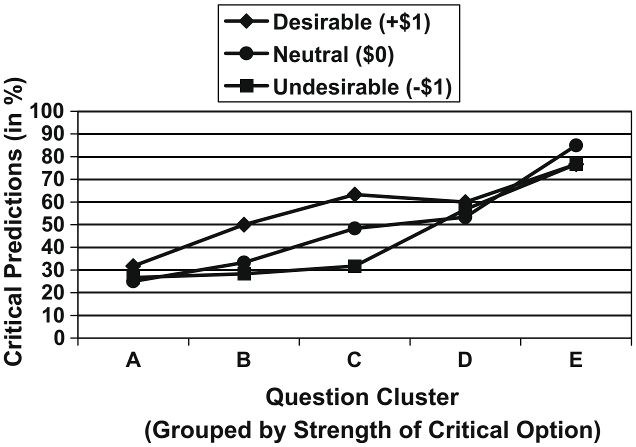

Fig. 5. From the new and old questions in Experiment 3, the percent of trials on which the critical item was predicted as a function of the overall strength of the critical item (weakest in Question Cluster A, strongest in Question Cluster E) and whether it was desirable (+$1), neutral ($0), or undesirable (−$1) for the critical items to be factually correct.

Discussion

These results show that predictions for trivia questions are not immune to the desirability bias. Consistent with our biased-guessing account, when people see no imbalance in the evidence for two options, their predictions are guesses that are vulnerable to a desirability bias.

Experiment 4

Thus far, we have focused on people's discrete outcome predictions. Yet, there are many everyday contexts in which people must estimate the likelihood of an event, not merely make a prediction. Although bias observed in outcome predictions is sometimes assumed to serve as evidence of bias in subjective likelihood, this has been identified as a questionable assumption (see discussion by Bar-Hillel & Budescu, 1995; see also Kahneman & Tversky, 1982). Therefore, testing the desirability bias with likelihood judgments is just as critical as testing the bias with discrete predictions. In Experiment 4, we returned to the card paradigm and directly compared the degree of the desirability bias in a likelihood-judgment condition and an outcome-prediction condition.

This direct comparison is particularly useful given the mixed findings from studies that have examined the desirability bias in likelihood judgments about stochastic events. For example, Biner, Huffman, Curran, and Long (1998) used a food reward to make a specific outcome of a card drawing desirable, and they found a significant desirability effect on a confidence measure. However, Bar-Hillel and Budescu (1995) conducted four studies in which the subjective probability of a chance outcome (e.g., a random selection from a visually presented matrix) was not significantly impacted by the desirability of the outcome. Most relevant is Price and Marquez (2005), who found that neither confidence estimates nor subjective probabilities were influenced by outcome desirability in a paradigm that was essentially the classic marked-card paradigm.

Our direct comparison between likelihood judgments and outcome predictions also provided another test of whether biased-evaluation, rather than biased-guessing or biased-thresholds, can account for the desirability bias in outcome predictions in the marked-card paradigm. If biased-evaluation processes are critical, then we should expect that the desirability bias would be comparable in magnitude when likelihood judgments or predictions are solicited. However, if biased guessing is critical, then we would not expect to see a robust desirability bias when the dependent measure solicits likelihood judgments. When likelihood judgments are solicited (at least under typical conditions; see Experiment 5 for an alternative), people would use the available evidence to generate their likelihood estimates, and there is no point at which it would seem suitable to insert an arbitrary component or guess. Therefore, there is no point at which to insert an arbitrary sense of optimism about an outcome. We should note that the biased-threshold account also does not predict a desirability bias for likelihood judgments, because there is not a role for a decision/prediction threshold in the judgment process.

In Experiment 4, we used a slightly modified version of our marked-card paradigm (discussed in the next paragraph) and we randomly assigned participants to either provide predictions or likelihood judgments—corresponding to two conditions we will call *dichotomous* and *continuous*. Critically, we made these two conditions as similar as possible. In fact, the only difference was the wording and formatting of the response anchors and scale that appeared below the question: "What is your prediction about which card was drawn?" In the dichotomous condition, two labeled response buttons (e.g., "Z" and "Y") appeared. In the continuous condition, a slider scale was used. More specifically, participants placed or slid a red marker along a line that had three anchors (e.g., "was definitely Z" on the left, "equal chances of Z and Y" in the middle, and "was definitely Y" on the right). Our key question of interest was whether the rather precise differences between our prediction and likelihood measures would result in different degrees of desirability bias.

We also used Experiment 4 to check on a counter-explanation for why the desirability bias was robust in the card condition but not the trivia condition of Experiment 2. Within the trivia condition of Experiment 2, but not in the card condition, the predictions were technically postdictions, because the factual outcomes or answers to the trivia questions were already determined yet unknown to the participant. Previous research has detected betting and confidence differences between prediction and postdiction (Rothbart & Snyder, 1970). To test whether *pre*-diction is a prerequisite for observing a desirability bias (and thereby assess whether the postdiction–prediction difference is a valid counter-explanation for the results of Experiment 2), we solicited only postdictions. That is, participants in both the dichotomous and continuous conditions provided their responses after the experimenter had already drawn the card on a given round. We expected the usual desirability bias in the dichotomous condition, because the role of guessing should not depend on whether the outcome has yet to occur or has already occurred but is still unknown.

Method

Forty-six undergraduates were randomly assigned to either the dichotomous or continuous condition. The other factors, procedures, and materials were identical to those of Experiment 1 with two exceptions. First, the experimenter always selected a card from the deck immediately before the participant was prompted to respond. Second, the scales in the two conditions differed as described previously (see two paragraphs above). For the accuracy incentive, all participants heard the same instructions ($0.25 per correct response). If asked for more information by a participant in the continuous condition, the experimenter explained that accuracy was based on whether their response was on the correct side of the scale.

Results

For the dichotomous condition, the results were similar to those from Experiment 1—see Fig. 6. Most importantly, a repeated

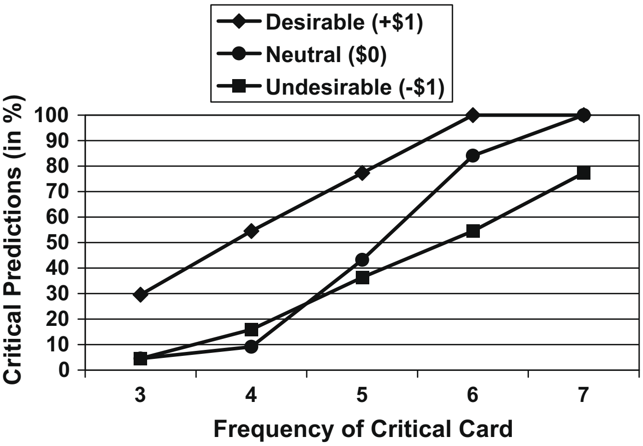

Fig. 6. From the dichotomous condition of Experiment 4, the percent of trials on which the critical card was postdicted as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1).

measures ANOVA revealed a significant desirability effect, $F(2, 20) = 12.85$, $p < .001$. The overall rate at which the critical mark was predicted was 72.3%, 48.2%, and 37.7% in the +$1, $0, and −$1 conditions, respectively. In short, even though a postdiction paradigm was used, the desirability bias was strong as usual.

For the continuous condition, we coded responses from 0% (for a response located at the endpoint anchored by the noncritical item) to 100% (for a response located at the endpoint anchored by the critical item). Fig. 7 displays the data pattern. In terms of inferential analyses, the most important finding was that a repeated measures ANOVA revealed a nonsignificant yet borderline desirability effect, $F(2, 22) = 2.82$, $p = .08$. The mean likelihood judgments regarding the critical options were 52.8%, 50.2%, and 47.8% in the +$1, $0, and −$1 conditions, respectively.

We also dichotomized the continuous data based on whether a participant's response was or was not on the side of the critical option (see Fig. 8). This allows us to directly compare results from the dichotomous and continuous conditions. There was a nonsignificant desirability effect, $F(2, 22) = 2.16$, $p = .14$. The overall rate at which the critical mark was predicted was 54.2%, 51.3%, and 46.3% in the +$1, $0, and −$1 conditions, respectively. More important, in a larger mixed-design ANOVA, the interaction between dependent-measure format (either dichotomous or continuous format—with the data dichotomized) and value (+$1 or −$1) was significant, $F(2, 43) = 6.53$, $p < .01$. This finding indicates that the desirability bias was significantly larger in the dichotomous condition than in the continuous condition.

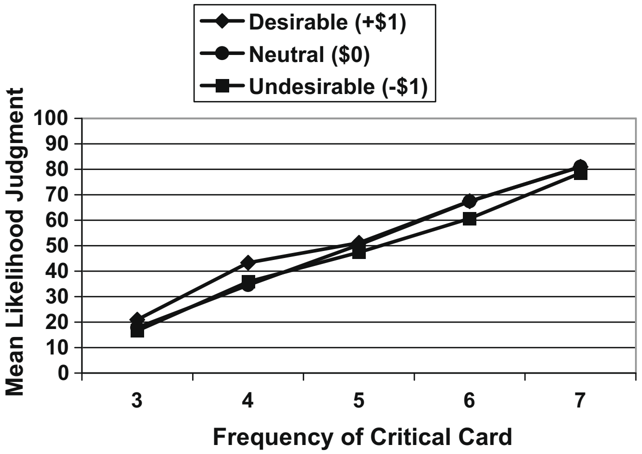

Fig. 7. From the continuous condition of Experiment 4, average likelihood judgment for the critical card as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1).

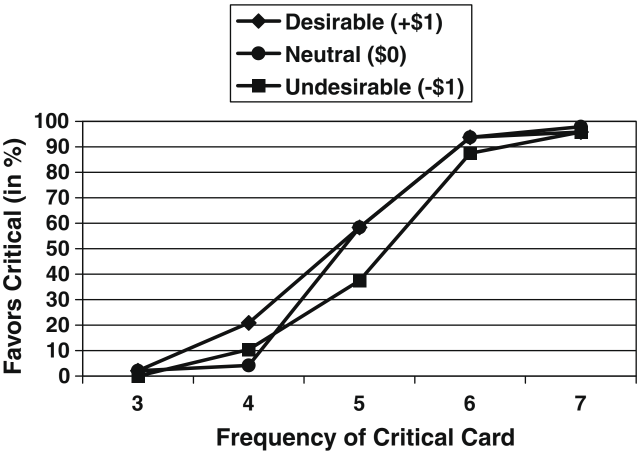

Fig. 8. From the continuous condition of Experiment 4, the percent of trials on which participants' likelihood judgments favored the critical item over the noncritical item, as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1).

Discussion

The results of Experiment 4 suggest that the desirability bias operates the same in a postdiction paradigm as in a prediction paradigm. The results also show that the magnitude of the desirability bias drops substantially when a continuous likelihood judgment rather than a dichotomous prediction is solicited. The findings are again consistent with the biased-guessing account. This account assumes that guessing would play the same role in dichotomous postdiction and prediction, but that guessing or arbitrary optimism would not have the same role in likelihood judgment.

Experiment 5

In Experiments 1–4, we did not directly manipulate rates of guessing. Instead, we tested whether the desirability bias would shrink substantially in conditions in which arbitrary guessing would not be critical determinants of responses (e.g., likelihood judgments; outcome predictions about nonstochastic events, except for incredibly difficult trivia questions). In Experiment 5, we sought more specific evidence of the role of guessing by directly manipulating guessing. We again used our marked-card paradigm, and we again used likelihood judgments as the dependent measure. We reasoned that even though likelihood judgments were shown in Experiment 4 to be relatively insensitive to desirability biases, we would observe a stronger desirability bias if we could sufficiently encourage participants to inject their arbitrary hunches or guesses into their estimates. To do this, we used a scale-juxtaposition method that we have developed for other projects to encourage people to separate their beliefs about the objective likelihood of an event from their more intuitive or gut-level impressions of the likelihood of the event. For each card draw, participants provided two judgments on separate scales appearing on the same screen (see description below). This method, along with strong accompanying instructions, gave participants explicit encouragement to express their guesses on one scale but not the other.

Method

Forty-four undergraduates participated. The design was identical to that of Experiment 1, except for the addition of a within-subject scale factor (assessment scale versus hunch scale). The procedures were also identical except for the differences describe here. Namely, on the computer screens that solicited responses, there were two questions and scales. The first question asked "What is your statistical assessment as to the card that will be drawn?" and was accompanied by a slider scale anchored by "will definitely be Z" on the left, "equal chances of Z and Y" in the middle, and "will definitely be Y" on the right. After the participant responded, the second question appeared below the first (with the first question and scale remaining visible.) It asked "What is your hunch or intuition as to the card that will be drawn?" and was accompanied by a slider scale anchored by "strongly leaning toward Z" on the left, "not leaning toward Z or Y" in the middle, and "strongly leaning toward Y" on the right. Instructions provided at the beginning of the session introduced the distinction between the two questions: "First, we will ask you to take a rational, statistical, and objective point of view and indicate your best assessment of the likelihood of one or another outcome. Next, we will ask you about your hunch, your guess, or your intuitive expectation about what will happen in the card draw. Maybe your intuitive expectations and hunches are similar to your more rational or statistical assessments, but they certainly don't need to be. We are interested in both types of predictions." Given that we were encouraging people to flip between statistical assessments and hunches, we removed the monetary incentives for accuracy.

Results

Responses on both scales were coded from 0% to 100%, as they were for the continuous condition of Experiment 4. Fig. 9a and 9b display the results. Our main prediction was that responses on the assessment scale would not exhibit a desirability bias, whereas responses on the hunch scale would exhibit a significant desirability bias. As is evident from a visual inspection of Fig. 9a and 9b tested more precisely below—this is exactly what we found.

The overall analysis for this study involved a Scale-Type × Desirability × Frequency ANOVA. The most critical result, which supports our main hypothesis, was a significant Scale-Type × Desirability interaction, $F(2, 42) = 15.50$, $p < .001$. The Scale-Type factor also produced a significant main effect and an interaction with frequency ($ps < .001$), but rather than detailing all the results from the overall ANOVA, we will focus on analyses conducted within the levels of scale-type.

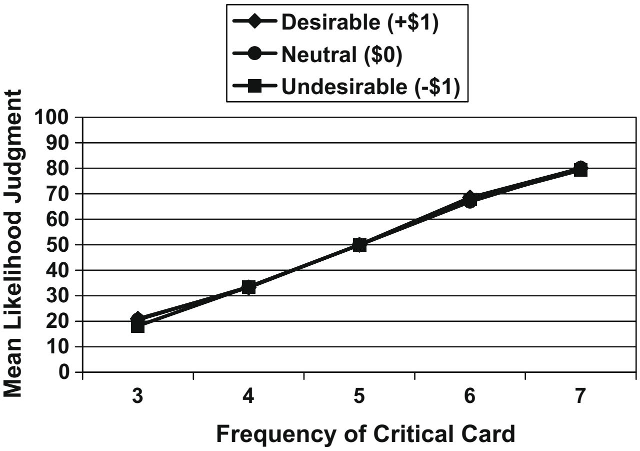

Fig. 9a. From the assessment scale of Experiment 5, average likelihood judgment for the critical card as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1). All three desirability lines are represented yet difficult to distinguish due to their proximity/overlap.

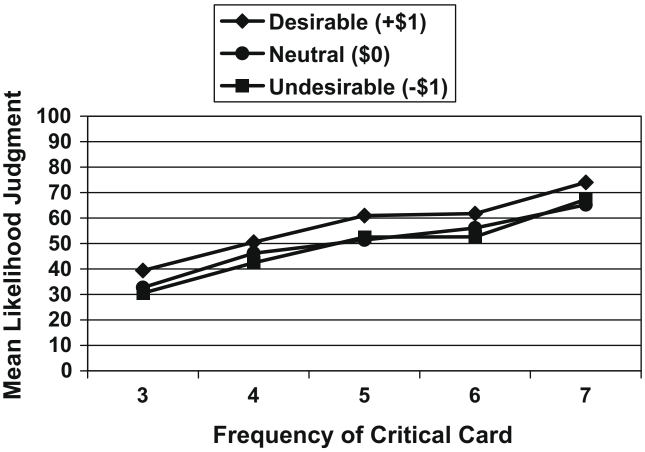

Fig. 9b. From the hunch scale of Experiment 5, average likelihood judgment for the critical card as a function of the frequency of the critical card (out of 10) and whether the drawing of a critical card was desirable (+$1), neutral ($0), or undesirable (−$1).

For the assessment scale, the desirability main effect was, as expected, not significant, $F(2, 42) = 1.18$, $p > .20$. Not surprisingly, the frequency of the critical card was a strong influence on the estimates, $F(4, 40) = 110.62$, $p < .001$. Finally, the Desirability × Frequency effect was not significant, $F(8, 36) = 0.87$, $p > .20$.

For the hunch scale, the desirability main effect was, as expected, significant, $F(2, 42) = 15.82$, $p < .001$. Not surprisingly, the frequency of the critical card was again a strong influence on the estimates, $F(4, 40) = 24.65$, $p < .001$. Finally, the Desirability × Frequency effect was not significant, $F(8, 36) = 0.69$, $p > .20$.

Discussion

The results of Experiment 5 demonstrate that when instructions and scales encourage people to express their hunch or guess—even on a continuous likelihood scale—the resulting estimates will be biased in an optimistic direction. This can be contrasted with the results of Experiment 4, which had demonstrated that when typical likelihood scales and instructions are used, people will not be substantially influenced by outcome desirability in the card paradigm.

Of the experiments in this paper, Experiment 5 is most direct in providing support for the biased-guessing account. When guessing was not encouraged, the desirability bias was essentially absent; when guessing was encouraged, the desirability bias was robust. We should emphasize that nothing about the instructions for the hunch scale suggested to people that they should guess optimistically rather than pessimistically or neutrally. It was conceivable that the results for the hunch scale could have reflected an increase in pessimism (e.g., bracing for negative outcomes) or simply no desirability effect (e.g., if the hunches reflected essentially random fluctuations). Therefore, the fact that the findings from the hunch scale revealed a tendency for guesses to fall in an optimistic direction is instructive, and it is compatible with our position that optimistic guessing was the primary basis for the effects in earlier experiments.

With that said, the results of Experiment 5 do have wrinkles. First, as is evident from Fig. 9b, the overall desirability bias is primarily driven by differences between the +$1 condition and the other two conditions; likelihood judgments did not differ between the −$1 and $0 conditions ($p > .20$). Second, the magnitude of the desirability bias involving hunches in this study is clearly smaller than those observed when participants made outcome predictions about cards in our other experiments. Additional research would be necessary to test whether these features of the results persist in replications and to determine precisely why. However, we do not believe that these features disqualify the conclusion that the results of Experiment 5 are supportive of our biased-guessing account. It is not too surprising that some characteristics of the desirability bias do not perfectly align between experiments that use substantially different dependent variables. Another interesting question for future research is whether the same desirability bias can be found when people see only one likelihood scale, but they are strongly urged to express their hunch. In designing Experiment 5, we assumed (based on preliminary work with juxtaposed scales in our lab) that the act of reporting their careful estimates of objective probability would help participants to distinguish between their objective assessments and their hunches—otherwise their reported hunch would be heavily anchored by what they knew to be the objectively correct answer. However, only additional research can determine whether it was the strong instructions or the combination of strong instructions and the juxtaposed-scales methodology that were critical for eliciting a significant desirability bias in Experiment 5.

General discussion

In the introduction to this paper, we discussed how the existing evidence for the desirability bias was mixed, and that the strongest evidence for some form of desirability bias was localized within a particular paradigm—the marked-card paradigm. Therefore, in the research described here, we sought to gain a better understanding of what underlies the desirability bias in the marked-card paradigm as well as test whether the bias extends to situations slightly different from the marked-card paradigm—namely to cases in which an outcome is nonstochastic and cases in which a likelihood judgment is solicited. Our main hypothesis was that the desirability bias in the marked-card paradigm was due primarily to biased guessing rather than biased evidence evaluation or biased-thresholds (or to experimental artifacts).

In Experiment 1, we detected the desirability bias in our new version of the marked-card paradigm that removed potential artifactual problems. In Experiments 2 and 3, using essentially the same paradigm but with nonstochastic rather than stochastic events, we showed that the desirability bias did not have the same impact on predictions about trivia questions, except for questions that were exceedingly difficult. In Experiment 4, the desirability bias was shown to extend to a postdiction paradigm but not to cases in which likelihood judgments rather than dichotomous postdictions were solicited. Finally, using a novel juxtaposed-scale method in Experiment 5, we showed that even for continuous likelihood judgments, a robust desirability bias could be observed when guessing was encouraged on one of the scales.

This set of findings is consistent with our position that biased guessing is the primary contributor to the robust effects in the classic marked-card paradigm. When participants in the marked-card paradigm face a deck with an equal number of critical and noncritical cards, guessing is essentially required. Even when the deck has unequal numbers of critical and noncritical cards, guessing might still be viewed as necessary by participants—except for those who apply a maximization principle. Therefore, biased guessing can account for large desirability biases regarding 50–50 desks as well as the gradual reduction in the desirability bias as the proportion of critical and noncritical cards becomes more unequal (Experiment 1). Also, the guessing component is applicable to stochastic events regardless of whether the relevant case concerns postdiction or prediction (Experiment 4). For nonstochastic outcomes, however, guessing is usually less relevant (Experiment 2). Participants base their prediction (or postdiction) on whatever their assessment of the evidence suggests; they are naturally reluctant to make a prediction that contradicts their own evidence assessment. If their evidence assessment offers no distinction between two outcomes (as with the exceptionally difficult questions introduced in Experiment 3), entirely arbitrary guessing becomes relevant, which makes predictions vulnerable to a desirability bias. Finally, for making likelihood judgments under typical conditions or instructions (e.g., Experiment 4), entirely arbitrary guessing is not relevant and therefore the desirability bias is minimal. However, when instructions and the juxtaposed-question format encouraged guessing, the desirability bias was robust (Experiment 5).

The biased-evaluation and biased-threshold accounts would have difficulty explaining elements of the overall result pattern. A biased-evaluation account would have particular difficulty explaining why the effects detected on outcome predictions would not extend to likelihood judgments. Assuming that evidence evaluation processes precede a response stage, one would expect any bias in evidence evaluation to manifest on various types of responses, not just outcome predictions. The biased-threshold account would have difficulty explaining why effects detected with stochastic cases (the card conditions) did not readily extend to the nonstochastic cases (the trivia conditions). If the desirability bias is simply due to a shift in response threshold, the bias would have been more evident for the most difficult trivia questions from Experiment 2, not merely the exceedingly difficult questions that we inserted in Experiment 3 as a way of testing guessing.

We should note that a biased-threshold account could be modified or extended in an effort to account for the results of Experiments 1–4. However, such an account would have to include awkward caveats. For example, we could assume that the bias in thresholds is so small that it produces detectable desirability bias only when uncertainty is exceedingly high (to account for the trivia-condition results), but less uncertainty might be required when there is stochasticity in the outcomes (to explain why there is a desirability bias even for 60–40 card decks, not just 50–50 decks). Furthermore, any biased-threshold account would have difficulty with the results of Experiment 5, because decision thresholds are not applicable to judgments on a continuous scale. Therefore, we favor our biased-guessing account and believe it is importantly distinct from a biased-threshold account. To explain the results of Experiment 5 using our biased-guessing account, we need to assume that a subjectively arbitrary component of an expectation can be expressed within a discrete prediction (as in the classic effect) but can also be exhibited as a judgment bias under some unique conditions (such as those set up in Experiment 5).

The big picture on the desirability bias

As we have discussed, our findings make a strong case that biased guessing is a key reason for the classic desirability biases found in the marked-card paradigm. However, what does the set of findings suggest about the desirability bias outside the specific paradigm? First, in terms of generalizing the marked-card results to everyday contexts, our findings suggest that people will often make optimistic predictions when guessing about stochastic outcomes. This is a critical conclusion because many everyday contexts involve predictions about outcomes that are either fully or partially stochastic, such as the case with Bob in the opening vignette, who might attempt to predict if the weather will allow the Blue Angles to fly.

Second, however, there are also many everyday contexts in which people need to make predictions about outcomes for which the relevant uncertainty is epistemic, not stochastic, such as Julie attempting to determine whether her branch is the one being closed (see Kahneman & Tversky, 1982). Our findings suggest that desirability biases might be less strong or absent in such cases, unless a person is so uncertain that she must simply guess about the outcome, rather than let her perceptions of evidence guide her predictions.

Third, our findings have important implications for the question of how outcome desirability impacts (if at all) judgments of likelihood or scaled optimism. The body of published research on this question is far from convincing (see Krizan & Windschitl, 2007a, 2009). Experiments directly examining this issue have produced mixed results (e.g., Bar-Hillel & Budescu, 1995; Bar-Hillel et al., 2008a, 2008b; Klein, 1999, Study 1; Krizan & Windschitl, 2007b; Price, 2000; Vosgerau, submitted for publication). In perhaps a telling sign regarding the published research on this issue, Bar-Hillel and Budescu (1995) entitled their paper describing several tests of the desirability bias (or wishful thinking) as "The Elusive Wishful Thinking Effect," and they entitled a recent follow-up chapter as "Wishful Thinking in Predicting World Cup Results: Still Elusive" (Bar-Hillel et al., 2008b). Also, although there are many plausible mechanisms by which motivations might influence evidence evaluation (see e.g., Armor & Taylor, 1998; Croyle, Sun, & Hart, 1997; Ditto & Lopez, 1992; Edwards & Smith, 1996; Krizan & Windschitl, 2007a; Kunda, 1990), these mechanisms have not been adequately tested in studies in which the dependent variable is likelihood judgment. When some form of likelihood judgment is the dependent variable, there can be factors that enhance pessimism (or mitigate optimism), most notably a tendency to brace for bad news (Butler & Mathews, 1987; Sanna, 1999; Shepperd, Findley-Klein, Kwavnick, Walker, & Perez, 2000; Shepperd, Grace, Cole, & Klein, 2005; Shepperd, Ouellette, & Fernandez, 1996; Sweeny, Carroll, & Shepperd, 2006; Sweeny & Shepperd, 2007; van Dijk, Zeelenberg, & van der Pligt, 2003; see also Armor & Sackett, 2006; Gilovich, Kerr, & Medvec, 1993). Finally, various studies and conceptual perspectives suggest ways in which people are pessimistically biased (e.g., Einhorn & Hogarth, 1985; Mandel, 2008; Pratto & John, 1991; Risen & Gilovich, 2007; Risen & Gilovich, 2008; Weber, 1994; Weber & Hilton, 1990; see also Chambers & Windschitl, 2004). In short, the question of how desires influence scaled optimism is far from settled in the existing literature.

Our findings suggest that the influence of outcome desirability must be understood in two parts. First, people—on average—might exhibit no large-scale optimistic or pessimistic biases in how they evaluate the likelihood of a desired outcome, when those likelihood estimates are solicited in a typical way (such as in Experiment 4). Second, people might simultaneously hold an optimistic assumption about potential outcomes, but this optimism will only be apparent with some types of measures (e.g., outcome predictions, specific likelihood measures that encourage and facilitate the expression of hunches). Returning to the title of this paper, people may sometimes have a way of "going optimistic without leaving realism."

Coda

Despite our conclusion about "going optimistic without leaving realism," we would be remiss if we did not point out the perhaps larger lesson from our results. Namely, any discussion of desirability bias must attend to potential moderators. As illustrated by our own findings, the apparent magnitude of a desirability bias can shift dramatically as a function of the nature of the critical outcome and the type of dependent variable—even when the same amounts of money are used to manipulate desire in these cases. Another potential moderator or set of moderators would be individual differences. Although not discussed above, we included many individual-difference measures in these studies. As it turns out, none were particularly useful for determining who would show an optimistic versus pessimistic tendency (see description in Appendix A). Perhaps there is, in fact, a broad-based tendency for humans to lean—all else equal—in an optimistic direction (see Armor and Taylor; 1998; Lench & Ditto, 2008; Peterson, 2000; Schneider, 2001; Taylor & Brown, 1988). People might typically default to an optimistic orientation given that optimism seems to be required for the fulfillment of goals (Armor & Taylor, 1998), or because an optimistic orientation is more compatible with maintaining a positive mood (Segerstrom, Taylor, Kemeny, & Fahey, 1998) and with being favorably perceived by others (Helweg-Larsen, Sadeghian, & Webb, 2002). However, we believe moderators of these influences could be quite important. For example, testing within a different culture or making a prevention goal (rather than promotion goal) salient could impact the results. Therefore, "going optimistic without leaving realism" provides a good description of what was found within the parameters of our experiments, and it may well reflect a general tendency, but we believe there is much to be learned about what moderators substantially qualify that phrase.

Acknowledgments

This work was supported by Grant SES 07-20349 to Paul D. Windschitl from the National Science Foundation.

We wish to thank Krista Reeve for her help in creating materials for this research.

Appendix A

Although we analyzed for individual-differences correlates within each experiment, the power to detect such correlates within many of our studies was small. Therefore, we also analyzed correlations across some experiments, and below we report the results in three sets: (1) based on participants who made outcome predictions about cards (Experiments 1, 2, and 4; total $N = 52$), (2) based on participants from the trivia conditions, who always made outcome predictions (Experiments 2 and 3; total $N = 69$), and (3) based on participants responses to the hunch scale from Experiment 5 ($N = 44$). Our main interest for these analyses was how the magnitude of the desirability bias—indexed as the difference between the rates of selecting the critical item when it was positive (+$1) versus when it was negative (−$1)—was related to the scores on the standard individual-difference measures.

The measures that we used included the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984), the Rational-Experiential Inventory, which assesses interest and self-perceived ability in relying on rational or experiential thinking (REI; Pacini & Epstein, 1999), the Numeracy Scale (Lipkus, Samsa, & Rimer, 2001), and the Life Orientation Test, which assesses dispositional optimism (LOT-R; Scheier, Carver, & Bridges, 1994). Experiments 1–4 also included a measure of promotion and prevention motivational orientations (RFQ; Regulatory Focus Questionnaire; Higgins et al., 2001), whereas Experiment 5 included the Behavior Inhibition Scale and Activation Scale (BIS/BAS; Carver & White, 1994) and the Belief in Good Luck Scale (Darke & Freedman, 1997).

In selecting these measures, we only included measures for which we could—a priori—articulate at least some rationale for

Table A1
Correlations between desirability bias and various individual-difference measures.

	Card conditions of Experiments 1, 2, and 4	Trivia conditions of Experiments 2 and 3	Hunch scale condition of Experiment 5
Dispositional optimism (from LOT-R)	.14	.12	−.06
Positive affect (from PANAS)	.13	.18	.08
Negative affect (from PANAS)	.10	−.05	−.12
Numeracy	−.29*	−.19	−.15
Need for cognition	−.20	.10	.03
Rational thinking total (from REI)	−.15	.05	.09
Experiential thinking total (from REI)	.26	.11	.25
Promotion focus (from RFQ)	.03	.04	
Prevention focus (from RFQ)	.14	.00	
Behavioral inhibition scale			−.31*
BAS—drive			−.16
BAS—fun seeking			.15
BAS—reward responsiveness			.05
Belief in good luck scale			.24

* Significant at .05 level.

its potential as a moderator of desirability bias. Nonetheless, essentially none of the measures proved to be a substantial moderator of the desirability bias (see Table A1). We will leave it to the reader to interpret patterns of interest, but our overall conclusion was that the standard measures were not helpful in predicting the magnitude (or direction) of people's desirability biases. Some readers might be most surprised by the fact that dispositional optimism measured with the LOT-R did not significantly predict the desirability bias, but other researchers have already documented that the LOT-R often does not do well in predicting optimism about specific events (e.g., Lipkus, Martz, Panter, Drigotas, & Feaganes, 1993).

Appendix B

See Table B1.

Table B1
This table displays the exact means and standard deviations relevant to Figs. 2–9.

	3 or A		4 or B		5 or C		6 or D		7 or E	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Critical predictions in percentage for card paradigm in Experiment 1 (see Fig. 2)										
Desirable (+$1)	13.3	29.7	46.7	44.2	90.0	20.7	93.3	17.6	100.0	0.0
Neutral ($0)	0.0	0.0	3.3	12.9	60.0	38.7	90.0	20.7	96.7	12.9
Undesirable (−$1)	0.0	0.0	16.7	30.9	23.3	32.0	70.0	41.4	80.0	36.8
Critical predictions in percentage for card condition of Experiment 2 (see Fig. 3)										
Desirable (+$1)	6.7	17.6	53.3	39.9	80.0	31.6	93.3	25.8	100.0	0.0
Neutral ($0)	6.7	17.6	30.0	45.5	53.3	35.2	83.3	24.4	96.7	12.9
Undesirable (−$1)	6.7	17.6	30.0	41.4	30.0	36.8	53.3	44.2	83.3	30.9
Critical predictions in percentage for trivia condition of Experiment 2 (see Fig. 4)										
Desirable (+$1)	34.6	45.1	52.6	48.4	51.3	50.8	66.7	47.9	84.6	32.5
Neutral ($0)	30.8	40.4	55.1	50.6	47.4	50.3	57.7	49.7	79.5	35.7
Undesirable (−$1)	20.5	30.9	46.2	51.0	50.0	50.1	60.3	48.0	84.6	32.8
Critical predictions in percentage for trivia condition of Experiment 3 (see Fig. 5)										
Desirable (+$1)	31.7	48.4	50.0	50.4	63.3	49.4	60.0	48.9	76.7	40.2
Neutral ($0)	25.0	45.2	33.3	46.9	48.3	51.1	53.3	52.0	85.0	32.6
Undesirable (−$1)	26.7	45.3	28.3	45.2	31.7	47.9	56.7	45.1	76.7	43.0
Critical predictions in percentage for the dichotomous condition of Experiment 4 (see Fig. 6)										
Desirable (+$1)	29.5	36.7	54.5	43.4	77.3	36.9	100.0	0.0	100.0	0.0
Neutral ($0)	4.5	14.7	9.1	19.7	43.2	38.7	84.1	23.8	100.0	0.0
Undesirable (−$1)	4.5	14.7	15.9	28.4	36.4	44.1	54.5	40.6	77.3	36.9
Likelihood judgments for the continuous condition of Experiment 4 (see Fig. 7)										
Desirable (+$1)	21.0	12.5	43.3	15.5	51.1	6.3	67.5	11.1	81.0	13.5
Neutral ($0)	17.8	9.0	34.5	11.0	50.4	6.9	67.4	11.0	81.0	8.5
Undesirable (−$1)	16.6	8.2	35.8	13.4	47.4	10.1	60.7	18.0	78.6	13.2
Percent of predictions favoring critical option for continuous condition of Experiment 4 (see Fig. 8)										
Desirable (+$1)	2.1	10.2	20.8	35.9	58.3	35.1	93.8	16.9	95.8	14.1
Neutral ($0)	2.1	10.2	4.2	14.1	58.3	38.1	93.8	16.9	97.9	10.2
Undesirable (−$1)	0.0	0.0	10.4	25.4	37.5	36.9	87.5	30.4	95.8	20.4
Likelihood judgments for the assessment scale of Experiment 5 (see Fig. 9a)										
Desirable (+$1)	21.0	12.2	33.2	10.2	50.0	0.7	68.5	10.0	80.0	11.58
Neutral ($0)	20.6	11.3	33.5	10.2	49.9	0.3	66.9	8.5	80.2	9.8
Undesirable (−$1)	18.2	10.5	33.5	9.7	50.0	0.2	67.7	11.3	79.3	11.4
Likelihood judgments for the hunch scale of Experiment 5 (see Fig. 9b)										
Desirable (+$1)	39.4	18.0	50.5	12.2	61.0	15.2	61.7	14.7	74.0	15.5
Neutral ($0)	32.7	16.1	46.2	15.7	51.4	15.1	56.1	15.7	65.2	20.1
Undesirable (−$1)	30.5	17.0	42.4	19.5	52.5	17.6	52.6	16.7	67.3	17.4

Note: percentages were first calculated per participant, and then the mean (or overall) percentages and the standard deviations were computed across participants.

Appendix C

Five examples of the trivia questions used in Study 2.

(1) How much of the world's population is left-handed?—About 25%, About 10%.
(2) Which state accounts for more oil produced in the United States?—Alaska, Texas.
(3) What country sends the most tourists to Australia?—Japan, United States.
(4) What is the most common last name in the US?—Smith, Johnson.
(5) Which state was first to require license plates on cars?—New York, Massachusetts.

Five examples of the new and exceedingly difficult trivia questions used in Study 3.

(1) The first police force was established in Paris in what year? —1676, 1667.
(2) What is the genus of both golden peas and night monkeys? —Aotus, Oenanthe.
(3) In 2000, how many people visited the Eiffel Tower? —6,315,324, 6,423,658.
(4) Who was the first US president inaugurated in American-made clothes?—James Madison, Andrew Jackson.
(5) Who invented the lava lamp?—Andrew Jenkins, Edward Walker.

References

Armor, D. A., & Sackett, A. M. (2006). Accuracy, error, and bias in predictions for real versus hypothetical events. *Journal of Personality and Social Psychology, 91*, 583–600.

Armor, D. A., & Taylor, S. E. (1998). Situated optimism: Specific outcome expectancies and self-regulation. In M. Zanna (Ed.). *Advances in experimental social psychology* (Vol. 30, pp. 309–379). New York: Academic Press.

Balcetis, E. (2008). Where the motivation resides and self-deception hides: How motivated cognition accomplishes self-deception. *Social and Personality Psychology Compass, 2/1*, 361–381.

Bar-Hillel, M., & Budescu, D. V. (1995). The elusive wishful thinking effect. *Thinking & Reasoning, 1*, 71–104.

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

Bar-Hillel, M., Budescu, D. V., & Amar, M. (2008b). Wishful thinking in predicting World Cup results: Still elusive. In J. I. Krueger (Ed.), *Rationality and social responsibility*. Psychology Press.

Biner, P. M., Huffman, M. L., Curran, M. A., & Long, K. R. (1998). Illusory control as a function of motivation for a specific outcome in a chance-based situation. *Motivation and Emotion, 22*, 272–291.

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.

Butler, G., & Mathews, A. (1987). Anticipatory anxiety and risk perception. *Cognitive Therapy and Research, 11*, 551–565.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306–307.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67*, 319–333.

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

Crandall, V. J., Solomon, D., & Kellaway, R. (1955). Expectancy statements and decision times as functions of objective probabilities and reinforcement values. *Journal of Personality, 24*, 192–203.

Croyle, R. T., Sun, Y., & Hart, M. (1997). Processing risk factor information: Defensive biases in health-related judgments and memory. In K. J. Petrie & J. A. Weinman (Eds.), *Perceptions of health and illness*. Amsterdam: Harwood.

Darke, P. R., & Freedman, J. L. (1997). The belief in good luck scale. *Journal of Research in Personality, 31*, 486–511.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. *Personality & Social Psychology Bulletin, 28*, 1379–1387.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality & Social Psychology, 63*, 568–584.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in evaluation of arguments. *Journal of Personality and Social Psychology, 71*, 5–24.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92*, 433–461.

Gal, I., & Baron, J. (1996). Understanding repeated simple choices. *Thinking & Reasoning, 2*, 81–98.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.

Gilovich, T., Kerr, M., & Medvec, V. H. (1993). Effect of temporal perspective on subjective confidence. *Journal of Personality & Social Psychology, 64*, 552–560.

Helweg-Larsen, M., Sadeghian, P., & Webb, M. S. (2002). The stigma of being pessimistically biased. *Journal of Social and Clinical Psychology, 21*, 92–107.

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., & Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology, 31*, 3–23.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*, 247–257.

Irwin, F. W. (1953). Stated expectations as a function of probability and desirability of outcomes. *Journal of Personality, 21*, 329–335.

Irwin, F. W., & Metzger, M. J. (1966). Effects of probabilistic independent outcomes upon predictions. *Psychonomic Science, 5*, 79–80.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*, 143–157.

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology, 63*, 534–544.

Klein, W. M. (1997). Objective standards are not enough: Affective, self-evaluative, and behavioral responses to social comparison information. *Journal of Personality and Social Psychology, 72*, 763–774.

Klein, W. P. (1999). Justifying optimistic predictions with minimally diagnostic information under conditions of outcome dependency. *Basic and Applied Social Psychology, 21*, 177–188.

Krizan, Z., & Windschitl, P. D. (2009). Wishful thinking about the future: Does desire bias optimism? *Social and Personality Psychology Compass, 3*, 227–243.

Krizan, Z., & Windschitl, P. D. (2007a). The influence of outcome desirability on optimism. *Psychological Bulletin, 133*, 95–121.

Krizan, Z., & Windschitl, P. D. (2007b). Team allegiance can lead to both optimistic and pessimistic predictions. *Journal of Experimental Social Psychology, 43*, 327–333.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498.

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.

Lipkus, I. M., Martz, J. M., Panter, A. T., Drigotas, S. M., & Feaganes, J. R. (1993). Do optimists distort their predictions for future positive and negative events? *Personality and Individual Differences, 15*, 577–589.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*, 37–44.

Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition, 106*, 130–156.

Marks, R. W. (1951). The effect of probability, desirability, and "privilege" on the stated expectations of children. *Journal of Personality, 19*, 332–351.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*, 972–987.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*, 407–413.

Peterson, C. (2000). The future of optimism. *American Psychologist, 55*, 44–55.

Peterson, C. R., & Ulehla, Z. J. (1965). Sequential patterns and maximizing. *Journal of Experimental Psychology, 69*, 1–4.

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

Price, P. C. (2000). Wishful thinking in the prediction of competitive outcomes. *Thinking and Reasoning, 6*, 161–172.

Price, P. C., & Marquez, C. (2005). *Wishful thinking in the prediction of a simple repeatable event: Effects of deterministic versus probabilistic predictions*. Unpublished Manuscript. California State University, Fresno.

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.). *Advances in experimental social psychology* (Vol. 20, pp. 297–333). San Diego, CA: Academic Press.

Risen, J. L., & Gilovich, T. (2007). Another look at why people are reluctant to exchange lottery tickets. *Journal of Personality and Social Psychology, 93*, 12–22.

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology, 95*, 293–307.

Rosenthal, R., & Fode, K. L. (1963). Psychology of the scientist V: Three experiments in experimenter bias. *Psychological Reports, 12*, 491–511.

Rothbart, M., & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioral Science, 2*, 38–43.

Sanna, L. J. (1999). Mental simulations, affect, and subjective confidence. *Psychological Science, 10*, 339–345.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A re-evaluation of the life orientation test. *Journal of Personality and Social Psychology, 67*, 1063–1078.

Schneider, S. L. (2001). In search of realistic optimism: Knowledge, meaning, and warm fuzziness. *American Psychologist, 56*, 250–263.

Segerstrom, S. C., Taylor, S. E., Kemeny, M. E., & Fahey, J. L. (1998). Optimism is associated with mood, coping, and immune change in response to stress. *Journal of Personality and Social Psychology, 74*, 1646–1655.

Shepperd, J. A., Findley-Klein, C., Kwavnick, K. D., Walker, D., & Perez, S. (2000). Bracing for loss. *Journal of Personality and Social Psychology, 78*, 620–634.

Shepperd, J. A., Grace, J., Cole, L. J., & Klein, C. (2005). Anxiety and outcome predictions. *Personality & Social Psychology Bulletin, 31*, 267–275.

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology, 70*, 844–855.

Sweeny, K., Carroll, P. J., & Shepperd, J. A. (2006). Is optimism always best?: Future outlooks and preparedness. *Current Directions in Psychological Science, 15*, 302–306.

Sweeny, K., & Shepperd, J. A. (2007). Do people brace sensibly? Risk judgments and event likelihood. *Personality and Social Psychology Bulletin, 33*, 1064–1075.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.

Trope, Y., & Liberman, A. (1996). Social hypothesis testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 239–270). New York: Guilford.

van Dijk, W. W., Zeelenberg, M., & van der Pligt, J. (2003). Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment. *Journal of Economic Psychology, 24*, 505–516.

Vosgerau, J. (submitted for publication). *Optimism and pessimism in subjective probabilities: How prevalent is wishful thinking?*

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measure of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070.

Weber, E. U. (1994). From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin, 115*, 228–242.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity events. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 781–789.

Windschitl, P. D., Martin, R., & Flugstad, A. R. (2002). Context and the interpretation of likelihood information: The role of intergroup comparisons on perceived vulnerability. *Journal of Personality and Social Psychology, 82*, 742–755.

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70%—right?: The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1514–1533.