DISBELIEVED BELIEFS: SUBJECTIVE ESTIMATES OF BIAS IN PROBABILISTIC BELIEFS AND THEIR RELATIONSHIPS TO DESIRE

Michael Martin Siepmann

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania in Partial

Fulfillment of the Requirements for the Degree of Doctor of Philosophy

1999

.....	.....	.....
Supervisor of dissertation	Supervisor of dissertation	Graduate Group Chairperson

To Anna.

Acknowledgments

The research I report in this dissertation was initiated by Karen Steinberg and Jonathan Baron, in collaboration with John Sabini and myself. I took over as principal investigator during the data collection phase of Study 1, and was primarily responsible for data analysis for the three studies and the design and execution of Studies 2 and 3. I am very gratefully to Karen and Jon both for getting the project started and for their generosity in allowing me to take the helm. I am also grateful to Jonathan Baron and John Sabini for their invaluable help both as advisors and as collaborators, and to David Williams and Robert DeRubeis, whose input as members of my dissertation committee has also been invaluable. John Monterosso and Steven Heine contributed some key ideas that were extremely helpful to me, both in clarifying how to interpret data I had already obtained and in planning subsequent studies. Barbara Gault, David Bersoff, Esteban Cardemil, Anna Gardiner, Johanna Renouf, Jeremy Siepmann, and David Siepmann also provided helpful input and discussion, and I am grateful to all of them.

ABSTRACT

DISBELIEVED BELIEFS: SUBJECTIVE ESTIMATES OF BIAS IN PROBABILISTIC BELIEFS AND THEIR RELATIONSHIPS TO DESIRE

Michael Martin Siepmann

Jonathan Baron and John Sabini

It has often been argued or assumed that it is impossible to simultaneously have a belief and disbelieve that belief, or view it as biased in a specific direction. In three studies we asked people what they believed were the probabilities of various outcomes varying in desirability, and then asked them what they would believe if they thought (a) in the way they would ideally like themselves and others to think, (b) in a way that would maximize the accuracy of their beliefs, or (c) in a way that would maximize the effectiveness of their beliefs for goal achievement. Results suggest that conscious disbelief of one's beliefs is neither impossible nor uncommon, that it is related to desire (though differently for different people), and that it may reflect some degree of accurate awareness of biasing effects of desire.

Table of Contents

List of Tables

Introduction

Previous research has provided much evidence that the desire to have a particular belief can bias people's thinking in such a way as to help them form or maintain that particular belief even if the evidence points in a somewhat different direction. This phenomenon has been discussed under many rubrics, such as motivated reasoning (Kunda, 1990), motivated social cognition (Kruglanski, 1996), wishful thinking (Bar-Hillel and Budescu, 1995), desirability bias (Budescu and Bruderman, 1995), self-deception (Silver, Sabini, and Miceli, 1989; Mele 1997), unrealistic optimism (Weinstein, 1980), self-serving attributional bias (Miller and Ross, 1975), akratic belief (Mele, 1994), and positive illusions (Taylor and Brown, 1988). Considerably less attention has been given to the question of whether people think their desires bias their beliefs. In one relatively trivial sense, it is clear that at least some people think that their desires bias their beliefs, namely psychologists who believe theories that say desires can bias beliefs (and who accept that these theories describe their own thinking and not just other people's thinking). However, a more interesting sense of this question is whether people can think that a specific belief of theirs is biased in a specific direction. Logic suggests the answer must be "no". For example, it would seem blatantly self-contradictory to say: "I believe that the probability of rain tomorrow is 40%, and I also believe that my desire to have a picnic tomorrow has caused me to underestimate that probability by 20%". The second clause implies that the speaker really believes the probability of rain is 60%, which of course contradicts the first clause. The seeming logical impossibility of disbelieving one's own beliefs has often been argued (or assumed) to imply psychological impossibility, or at least the need for the disclaimer that the disbelief must be

unconscious. Rather than assume that logical impossibility implies psychological impossibility, we decided to try asking people whether they disbelieved their beliefs.

Before describing our findings, it is worth (a) explaining some terms we will be using, (b) reviewing previous research on desirability bias in probabilistic beliefs, (c) reviewing expressions of the idea that it is impossible to disbelieve one's beliefs, and (d) reviewing previous findings of awareness of bias.

Terminology

We will discuss our terminology with respect to probabilistic beliefs, since that is what our studies are about, but this terminology could in principle be used for other kinds of beliefs.

By "objective bias" we mean any normatively unjustified deviation of a belief from an objectively determined normatively correct belief. For example, if Joe believes the probability of rain tomorrow is .8 but normatively he should believe it is .9, his belief has an objective bias of -.1.

By "subjective bias" we mean a person's subjective estimate of their objective bias. So if Joe thinks that normatively he should believe the probability of rain tomorrow is .7, but he actually believes it is .8, he has a subjective bias of +.1. The idea that people cannot disbelieve their beliefs can also be expressed, then, as the idea that subjective bias must always be zero.

By "objective desirability bias" we mean the relationship between desire and objective bias. For example, someone who tends to underestimate the probabilities of desired outcomes by .1, neutral outcomes by .2, and undesired outcomes by .3 has a positive

objective desirability bias; the more they desire an outcome, the more positive and the less negative their objective bias. This example demonstrates the important point that positive desirability bias does not necessarily imply overestimation of the probabilities of desired outcomes and underestimation of the probabilities of undesired outcomes; desirability bias is a relationship between desire and the degree of over- or underestimation. In the example above, the person underestimates the probabilities of all outcomes, but underestimates less for desired outcomes than undesired outcomes.

By "implicit subjective desirability bias" we mean a systematic positive or negative relationship between desire for a belief to be true and subjective bias. It is "implicit" because the relationship with desire is objectively measured by the experimenter, not explicitly reported by the subject. We call a subject's explicit report of a relationship between their subjective bias and their desire "explicit subjective desirability bias". Table 1 summarizes the definitions of the three types of desirability bias in terms of how they are measured. From now on, we will abbreviate desirability bias as DB, except in headings.

Evidence for Objective Desirability Bias in Probabilistic Beliefs

Some studies providing evidence about the existence of objective DB in probabilistic beliefs have asked subjects to express their probabilistic beliefs directly, while others have asked for other kinds of responses related to probabilistic beliefs, such as binary predictions. Another important way these studies vary is in the type of normative basis they have for calling subjects' beliefs biased.

Objective Probability as the Normative Basis

Bar-Hillel and Budescu (1995) noted that there have been few studies of DB in which

subjects express probabilistic beliefs directly and the normative standard of comparison is an objective probability. They conducted several studies of this type, and only one yielded a DB effect. However, it was the only one in which they both gave subjects a real basis for desire, in the form of monetary consequences, and did not give subjects a monetary incentive for accuracy. An earlier study in this category was conducted by Slovic (1966), who compared probabilistic beliefs to objective probabilities specified by Bayes' theorem at various points during subjects' observation of a sample of 50 poker chips being drawn from a bag of 100 chips, which they were told had either had 30, 40, 50, 60, or 70 red chips. Subjects for whom the number of chips in the bag mattered financially had probabilistic beliefs that deviated more from the Bayesian ideal than those of subjects for whom the number of chips in the bag was inconsequential. However, the effects of desire differed between subjects, with some people showing consistent positive DB and others showing consistent negative DB. Individual differences in DB have also been reported by Irwin and Snodgrass (1966), Irwin and Metzger (1966, 1967), and Cohen and Wallsten (1992). Slovic (1966) pointed out that "these individual differences tended to cancel one another when data were averaged over Ss, thereby reducing the size of between group differences" (p. 28), which suggests that studies that have not considered the possibility of individual differences in DB may have underestimated its importance.

There have been several studies in which the normative standard of comparison is an objective probability, but subjects express their beliefs as binary predictions. Marks (1951) told fifth and sixth graders how many picture cards were in packs of ten cards. At all objective probabilities (.1 to .9), children were much more likely to predict that they would

draw a card with a picture on it when a picture card would earn them a “point” than when it would lose them a “point”. Irwin (1953) found a similar, but weaker, effect with college students. He also asked for ratings of confidence in predictions and found “some evidence that desirability leads to increased confidence as well as greater frequencies of expectation of desired outcome." (p. 334). Crandall, Solomon, and Kellaway (1955) and Irwin and Graae (1968) found that stronger manipulations of desire had stronger effects on predictions, and Irwin and Snodgrass (1966; Irwin and Metzger, 1966, 1967) found that DB remained even when subjects were required to bet on the outcome they predicted, thereby making it against their financial interest to predict objectively less likely, but more desirable events. Budescu and Bruderman (1995) hypothesized that having subjects make several consecutive predictions of whether they would draw a marked card from a single pack, which was reshuffled after each prediction, rather than a single prediction, might eliminate DB, but found that it did not.

Outcomes as the Normative Basis

When the objective probability of events is not known, the actual outcomes can be used as a normative basis for judging DB. Sieber (1974) found that students who believed a multiple choice exam would irrevocably affect their course grade overestimated the probabilities of their answers being correct, more than did students who believed they would have a second chance at the examination. Wright and Ayton (1989) had subjects rate the desirability of 70 personal events, predict whether each event would occur in their lives over the next four weeks, and estimate the probability that their prediction would prove accurate[1]. They also paid their subjects to return once a week over the four week period to record

whether each event had happened. They found that across items, the mean desirability of an item was significantly positively correlated (over .4) with overconfidence on that item (i.e., the mean judged probability minus the proportion of subjects for whom the event had actually happened).

Impossibility of Everyone Being Right as the Normative Basis

Fischer and Budescu (1995) had subjects predict the number of seats to be won by seven parties (or clusters of small politically similar parties) in an election to the Israeli Knesset, and to rate their degree of identification with each party. Although they reported the actual number of seats won by each party, they did not use this outcome information in analyzing DB^2. For four of the seven parties, they found positive relationships between identification and number of seats predicted. The normative basis for calling this a DB is that if people make different predictions about an event that can only turn out one way, then they cannot all be right. If their predictions are related to their desires, then as a group their errors must also be related to their desires, even though some people's predictions may turn out to be right. Similar findings were reported by Babad (1995) with regard to the same election, and by Granberg and Brent (1983) for two party elections. In the financial domain, Olsen (1997) asked investment managers to rate the desirability and probability of various economic events (e.g., "corporate merger activity will diminish"). He computed separate desirability-probability correlations for each event. This correlation was above .2 for all but two of 14 events he asked Chartered Financial Analysts in the U.S.A. about, and averaged .29. Of 10 events he asked Taiwanese investment managers about, the correlation was above .2 for nine, and averaged .36.

The same normative basis applies to Weinstein's (1980) finding that people believe, on average, that their chances of experiencing various future life events are above average for desirable events and below average for undesirable events. Zakay's (1983) finding that people judged the probabilities of positive events more than .1 higher for themselves than for "'someone just like yourself in every respect'" (p. 276), and the probabilities of negative events more than .06 lower, is similar to Weinstein's, but the normative basis for attributing bias is different. While logically there is nothing to prevent all of Zakay's subjects from being right that their prospects are better than those of someone exactly like them in every respect, it is hard to think of a justification for anyone believing this, since if the comparison person does not differ in any respect, what basis can there be for attributing worse prospects to that person?

The Idea That Subjective Desirability Bias Must Be Zero

The idea that Subjective DB must be zero--that someone cannot view a belief she has as biased by desire, thereby disbelieving her own belief--has been discussed in the literatures on self-deception and motivated social cognition. Whether or not Subjective DB must be zero has important implications for approaches to improving people's thinking. Baron (1994) suggested that "If people know that their thinking is poor, they will not believe its results. One of the purposes of a book like this is to make recognition of poor thinking more widespread, so that it will no longer be such a convenient means of self-deception" (p. 45). Baron (1991, 1995) presented evidence that individual differences in standards for thinking are indeed correlated with corresponding individual differences in how people actually think. If people really cannot knowingly violate their own standards, then poor thinking must be

due to either failure to notice that one is violating one's own standards, or to having poor standards, or both. In that case, persuading people to adopt better standards for good thinking, and perhaps encouraging them to monitor more closely whether they are living up to them, should improve thinking. On the other hand, if people knowingly violate their own standards, then merely improving their standards, or their awareness of whether they are living up to them, may only lead people to knowingly violate better standards, and perhaps notice more of their violations, without thinking any differently. If this is the case, then those who wish to improve thinking cannot rely on an automatic link between noticing a violation of a standard and correcting it. It may be necessary to persuade people of the benefits of living up to ideal standards for good thinking, so that when they notice they are violating them, they will be motivated to do something about it.

Self-Deception and the Apparent Impossibility of Conscious Contradictory Beliefs

In discussions of self-deception, there is disagreement about whether self-deception requires simultaneous contradictory beliefs at all (Silver et al., 1989 and Mele, 1997, argue that it does not), but apparent agreement that if it does, they cannot both be conscious, otherwise the deception would not work. Gur and Sackeim (1979) wrote that "it has been noted that when it is assumed that people are necessarily aware of their cognition, the concept of self-deception is paradoxical" (p. 148), and quoted Sartre's (1958) formulation of the apparent paradox:

> The one to whom the lie is told and the one who lies are one and the same person, which means that I must know in my capacity as deceiver the truth which is hidden from me in my capacity as the one deceived. Better yet, I must know the truth very exactly in order to conceal it more carefully--and this not at two different moments, which at a pinch would allow us to re-establish a semblance of

> duality--but in the unitary structure of a single project. How then can the lie subsist if the duality which conditions it is suppressed? (p. 49) (p. 148)

Mele (1997) noted that some theorists conclude from this that self-deception cannot exist: "The very nature of belief precludes one's simultaneously believing that p is true and believing that p is false . . . . self-deception, according to the view at issue, requires being in an impossible state of mind" (p. 92). Other theorists conclude that one of the beliefs must be unconscious. For example, Gur and Sackeim (1979) proposed...

> the following criteria as necessary and sufficient for ascribing self-deception to any given phenomenon:
>
> 1. The individual holds two contradictory beliefs (p and not-p).
> 2. These two contradictory beliefs are held simultaneously.
> 3. The individual is not aware of holding one of the beliefs (p or not-p).
> 4. The act that determines which belief is and which belief is not subject to awareness is a motivated act. (p. 149)

The evidence they offered for the existence of self-deception defined this way involves interpreting physiological responses as indicative of unconscious beliefs that contradict conscious beliefs. Quattrone and Tversky (1984) also provided evidence for self-deception, as defined by Gur and Sackeim (1979), by showing that people increased or decreased the amount of pain they tolerated after being led to believe that a high or a low pain threshold was diagnostic of a healthy heart, but denied having done this intentionally. Of most relevance to the question of whether Subjective DB must be zero was the fact that of nine subjects who admitted having purposefully tried to change the amount of pain they tolerated, only two inferred that they had a healthy heart, compared to 20 of the 29 subjects who denied intentionally changing their pain tolerance. Quattrone and Tversky took this to show that "denying the ulterior motive makes it easier for subjects to make the comforting

diagnosis" (p. 243). This may well be true, but it does not show that it is impossible to think a belief is biased and still hold that belief. Lack of awareness may facilitate bias, but that does not mean that awareness renders bias impossible. It might be, for example, that a certain amount of discrepancy between what one believes and what one knows one should believe is manageable, but that larger discrepancies require suppression of the knowledge of what one should believe. By analogy, one can usually exceed the speed limit by a certain amount without getting stopped by the police, but to test whether one's new Lamborghini can go as fast as advertised without paying a hefty fine one must conceal one's speed from the police. Self-deception (or DB) might work in a similar way.

The size of the discrepancy might not be the only determinant of whether awareness of bias can coexist with bias. Another possible determinant is ambiguity about what one should believe. Greater ambiguity may allow greater discrepancies between what one believes and what one thinks one should believe, since it allows one to think that perhaps one actually should believe what one does believe. A person might say, for example, "I probably should believe X, but who knows, maybe I'm right to believe Y!". A stronger version of this idea would be that disbelieved beliefs may only be possible in the presence of ambiguity about what one should believe. In other words, disbelieved beliefs may be constrained by the condition "p(I should believe Y) > min", where min is some minimum probability, but not by the condition "p(I should believe Y) > p(I should believe X)", where X and Y are two alternative beliefs.

Motivated Social Cognition and the Need for an Illusion of Objectivity

Discussions of DB in the motivated social cognition literature have focused not

mainly on logical paradoxes of contradictory beliefs, but on the processes of evidence gathering and inference by which people may arrive at motivationally biased conclusions. Whereas discussions of self-deception have pondered how someone could believe p and not-p simultaneously, discussions of motivated social cognition have not found it necessary to propose that people ever do such a thing, but have focused rather on how people can control the process of thinking to bias it towards their desired conclusion without noticing that they are doing this. The assumption is that if they noticed that their thinking was biased, they would not believe the conclusions it led them to. For example, Pyszczynski and Greenberg (1987) argued that when motives influence attributions, "they do so in ways that enable one to maintain an illusion of objectivity concerning the manner in which those inferences were derived" (p. 302), and Klein and Kunda (1992) proposed that "people attempt to construct seemingly rational justifications for their desired beliefs . . . . They draw the desired conclusions only if they can construct such justifications" (p. 146, emphasis added).

Klein and Kunda (1992) found that the desire to believe specific things about other people led people to alter their beliefs about more general issues in such a way as to allow them to infer their specific desired belief from their general beliefs. For example, when told that someone had got all the answers right on a history quiz, and asked to judge the role of luck in determining people's scores on such quizzes, subjects who expected the target person to be their partner in an upcoming history quiz game with a cash prize judged luck as less of a determining factor than did subjects who expected the target person to be their opponent in the upcoming quiz. Of course, the greater the role of luck, the less the good score on the previous quiz would predict good performance on the upcoming quiz.

While Klein and Kunda's (1992) results suggest that people try to construct justifications for their desired beliefs, a study by Hsee (1996) provides more direct evidence that the degree to which they are able to do this affects the degree to which they are able to arrive at their desired beliefs. For example, he asked people to imagine that they were a real estate appraiser appraising a condominium based on a description of its features and the features of another condominium they had recently appraised at $100,000. The target condo was either described as identical to the previously appraised one, or as having some advantages and some disadvantages relative to the comparison condo. When the target condo had pros and cons relative to the comparison condo, the appraisals differed considerably depending on whether the subjects were told to imagine that their fiancé was the prospective buyer or was the prospective seller of the condo. When the target condo was described as identical, whether the fiancé was the buyer or the seller made significantly less difference. Presumably, when the condos differed, it was much easier for subjects to construct justifications for pricing the target one differently from the comparison one, in whichever direction their desire to please their fiancé dictated. Hsee's (1996) "elasticity hypothesis" is that when factors one can justify basing a decision on are "elastic", that is, their implications for the decision are amenable to multiple interpretations, one will take advantage of this "elasticity" in the justifiable factors in order to covertly let unjustifiable factors influence one's decision in the desired direction.

These results and others provide impressive evidence that an "illusion of objectivity" helps when cultivating biased beliefs. However, they do not show that without an "illusion of validity" one cannot have biased beliefs at all. While Klein and Kunda's (1992) proposal

that "people attempt to construct seemingly rational justifications for their desired beliefs" (quoted above) has persuasive evidence to support it, their proposal that people "draw the desired conclusions only if they can construct such justifications" (quoted above, emphasis added) is a much bolder claim. That claim requires evidence that people never draw desired conclusions at the same time as realizing that their conclusions are not fully justified, and the claim can be falsified by a single demonstration that people sometimes do this, which we think we provide in the three studies reported below.

So, whether or not self-deception is facilitated when beliefs that contradict one's desired beliefs are unconscious, and whether or not motivated bias can operate more effectively under the cover of an "illusion of objectivity", we are not aware of any empirical evidence that shows it is impossible to believe that a belief one holds is biased by desire--evidence that it is impossible to have a nonzero Subjective DB. Perhaps, because of the assumption that logical impossibility implies psychological impossibility, nobody has gone to any great lengths to show that nonzero Subjective DB is impossible (or possible). For example, with regard to Truly False Consensus Errors (the bias of overprojecting one's own characteristics into one's estimates of others' characteristics), Krueger and Zeiger (1993) (who argue that people are aware that other people have that bias), state explicitly "we assumed that people do not recognize TFCEs in their own judgments. We left this assumption untested because we considered the coexistence of the commission and recognition of TFCEs within the same person patently illogical" (p. 680).

Why Nonzero Subjective Desirability Bias May Be Possible

But logic can be deceptive. The apparently logically nonsensical nature of nonzero

Subjective DB is based on the assumption that the sentence "I believe p and I believe not-p" is as self-contradictory as the sentence "p is true and p is false". However, this depends on what "I believe" means. After all, "I like p and I like not-p" need not be contradictory. It is perfectly intelligible to say, "I like being with people and I like not being with people." Also, "We believe p and we believe not-p" is a perfectly intelligible thing for a group of people to say, meaning that the group includes people with each belief. Would it not also be quite reasonable for a person to say, "I understand that the evidence implies p, but my gut feeling and intuitive sense is that not-p"? If we pressed this person to tell us what she "really believes", would we consider her insane if she responded, "Both. Intellectually I believe p, but at a gut level I believe not-p"? We will defer until the General Discussion detailed consideration of the kinds of theories that might account for a finding of nonzero Subjective DB. For now, having noted that the apparent logical impossibility of disbelieving one's own beliefs may reflect mistaken assumptions about the nature of the entity that says "I believe...", or about what that entity means when it says that, it is worth reviewing empirical findings that point in the direction of awareness of bias being possible.

There are at least three possible types of awareness of bias. The type of most interest here involves simultaneously having a bias and being aware of having it. Whether that type can exist depends on whether awareness of bias automatically and inevitably causes correction of bias. If it does, then although awareness of bias while the bias persists would be impossible, awareness that one had a bias would still be possible. And even if people could never be aware even that they had a bias, they might still be aware that a particular form of bias exists in other people.

Awareness of Bias in Others

Several studies have provided evidence for awareness that others have biased beliefs. Dawes, Singer, and Lemons (1972) showed that people wrongly think that the attitudes of people who disagree with them are more extreme than their own attitudes, a bias called the "contrast effect". They also found that people were aware that other people have this bias. They put together a collection of attitude statements, some of which a target person endorsed and others of which the target person expected someone who disagreed with him or her to endorse. They showed this collection of statements to subjects and asked them to guess which statements the target person endorsed and which the target person thought someone who disagreed with him or her would endorse. At a rate significantly above chance, subjects correctly inferred that the target's own attitudes were the less extreme ones. Many of Dawes et al.'s subjects explicitly "reported that they had used their own version of the contrast effect" (p. 293) in inferring the target's own attitudes, leading Dawes et al. to comment that "a paradox arises: If people naively understand that people tend to exaggerate the differences between themselves and those with whom they disagree, why do they continue to do it?" (p. 293).

Snyder, Stephan, and Rosenfield (1976) rigged a competitive game and randomly assigned subjects to win or lose conditions. After the game, they had subjects both (a) rate the roles of skill and luck in causing their own and their opponent's outcomes and (b) predict their how their opponent would answer these rating questions. Self-serving bias was clearly evident in the ratings; losers attributed the outcomes more to luck and less to skill than did winners. Snyder et al. expected subjects to be unaware of this self-serving bias, even in their

opponents. They thought subjects would expect their opponents to give ratings similar to their own. In fact, subjects' predictions of their opponents ratings were closer to the opponent's actual ratings than to the subjects own ratings, indicating that they were "generally aware of their opponent's egotism" (p. 439). Snyder et al. presumed that this awareness did not extend to subjects' own egotism, and that, if asked, the subjects would have predicted that an "uninvolved bystander" would give ratings similar to their own. That presumption received some support from Vivian and Berkowitz's (1992) finding that subjects who thought they had been randomly allocated into one of two groups (a) rated work produced by their ingroup more favorably than work they believed was produced by the outgroup, (b) predicted that impartial judges would agree with their ratings, and (c) predicted that outgroup members would disagree with their ratings, rating their own work more favorably than the subject's group's work. That is, they expected the outgroup to be biased, but did not see themselves as biased.

Kirkpatrick and Epstein (1992) and Denes-Raj, Epstein, and Cole (1995) studied the ratio bias effect, a tendency to treat a probability as higher if it is expressed as a ratio of larger numbers. (For example, people find a 10 in 100 chance of winning a prize more attractive than a 1 in 10 chance). They found that "most people believe they are more rational than others and, accordingly, assume that others will exhibit a ratio bias effect more than they will" (Denes-Raj et al., 1995, p. 1090).

Krueger and Zeiger (1993) showed that people correctly expected others to show the Truly False Consensus Effect, which consists of overestimating the extent to which other people share one's characteristics. In one study, they told subjects whether a target person

had been willing to walk around campus wearing a sandwich board saying Eat at Joe's, and then asked them to guess the target's estimate of the percentage of people who would be willing to do this. Subjects who were told the target had been unwilling thought the target would give lower estimates than did subjects who were told the target had been willing. In another study, Krueger and Zeiger showed subjects a target person's estimates of the percentages of people who endorse each of 18 MMPI statements, and asked them to guess whether the target person had endorsed each statement him or herself. Across items, the more the target had overestimated (or the less they had underestimated) the percentage of people who endorse the item, the more subjects tended to guess that the target him or herself did endorse the item. Krueger and Zeiger interpreted both of these findings as showing that subjects were aware that people's own characteristics bias their estimates of others' characteristics.

Krueger, Ham, and Linford (1996) provided evidence that people are aware of being victims, but not perpetrators, of the actor-observer effect. This is a tendency for people, as observers, to view other people as consistently displaying the same traits across a wide range of situations, while as actors, they view themselves as displaying traits to varying degrees depending on the situation. As actors, Krueger et al.'s subjects correctly predicted that an observer (their roommate) would rate them as consistent across situations on more traits than the actors rated themselves consistent on. As observers, however, subjects failed to predict that the actor they were rating (their roommate) would rate him or herself as consistent on fewer traits than the observer rated him or her consistent on.

Krueger (1998) found that students' ratings of the self-descriptiveness and desirability

of a set of traits showed what he called a self-enhancement bias; for example, people who rated the trait "quiet" as more self-descriptive than the average person rated it tended also to rate it as more desirable than the average person rated it[3]. The same correlation appeared in students' predictions of their roommates' ratings of the self-descriptiveness and desirability of the traits, and Krueger interpreted this as showing that people are aware that others have this type of self-enhancement bias.

How does the above evidence for awareness of bias in other people bear on the question of whether people can be aware of bias in their own beliefs? To believe that other people's beliefs are affected by bias but that one's own are not, one presumably must either have an answer to the question, "what's so different about me?", or must ignore or fail to ask that question. Of course, attribution of bias to others but not oneself might be a bias in itself. If most people think they are better than average (Weinstein, 1980), then most people may also think that they are one of a small number of unbiased individuals in the world, and they may have little trouble explaining what's so different about them. However, someone who believes other people are biased seems more likely to consider the possibility that he himself is biased, than someone who has no awareness of bias in anyone. If a person does become aware of bias in her own beliefs, the next question is whether she will correct the bias she perceives.

Correction of Perceived Bias

The ability to notice and attempt to correct bias has been explored in the assimilation-contrast literature (e.g., Wegener and Petty, 1995), which focuses on how judgments can be biased by irrelevant contextual factors such as mood, priming effects, and recently accessed

memories. Assimilation occurs when the judgment takes on qualities of the irrelevant factor, for example if someone in a happy mood judges a person as nicer than they would had they been in a neutral mood. Contrast effects occur when the judgment becomes more dissimilar to the irrelevant factor than it would otherwise have been, for example if watching a documentary on hurricanes leads someone subsequently to judge the weather that day as less windy than she otherwise would have. Wegener and Petty (1995) proposed that people have differing naive theories about how various types of contextual influences are likely to bias their judgments. For example, their subjects expected that thinking about the weather in vacation spots like Jamaica, the Bahamas, or Hawaii would have a contrast effect on their judgments of the desirability of the weather in Indianapolis, making it seem less desirable than if they had not thought about the vacation spots. However, they expected that thinking about the weather in those vacation spots would have an assimilation effect on their judgments of how satisfied people in Hawaii and the Bahamas are with their jobs, making them judge those people as more satisfied with their jobs. Wegener and Petty had all their subjects rate the weather in the three vacation locations and then asked some of them to try to make sure that their perceptions of the weather did not influence their next set of ratings. The effects of this bias correction prompt were in accord with people's naive theories of how thinking about the weather in the vacation spots would bias their judgments; compared to subjects who were not prompted to correct bias, those who were prompted rated the weather in Indianapolis as more desirable, and rated people in Hawaii and the Bahamas as less satisfied with their jobs. Wegener and Petty also showed that for a given context and judgment pair (e.g., vacation spot weather as the context and job satisfaction in Hawaii as

the judgment) individual differences in naive theories predicted individual differences in the effects of a bias correction prompt. According to Wegener and Petty, previous models of bias correction in the assimilation-contrast literature have assumed that attempts to correct bias are always attempts to counteract assimilation by making the judgment less like the contextual factor. One thing to note about Wegener and Petty's methodology is that subjects were explicitly told what source of potential bias to attend to. The finding that people can correct when prompted as explicitly as this does not necessarily imply that people ever spontaneously notice and try to correct a potential bias of this sort, though it certainly makes it more plausible that they might.

Indirect evidence that people can correct DB is, perhaps, provided by Bar-Hillel and Budescu's (1995) findings that the wishful thinking effect was "elusive" in several studies in which they explicitly asked people to be as accurate as possible and offered a monetary incentive for accuracy. Bar-Hillel and Budescu did not interpret their findings this way, and since they did not directly assess the effects of an accuracy instruction or incentive, it is admittedly speculative to interpret their findings this way. However, it is interesting that in the one study in which they did find a DB effect, they did not include an accuracy instruction or incentive.

However, there have been several findings suggesting that people either cannot correct DB, or are unwilling to correct it. Babad, Hills, and O'Driscoll (1992) gave some subjects predicting the outcome of a New Zealand election an instruction that "stated that research has shown that people often tend to be emotional, irrational, and biased when dealing with political issues, and warned the the respondent to be as objective and rational

as possible" (p. 465). They found that subjects who were given this instruction "displayed somewhat moderated WT [wishful thinking], but the influence of that instruction was relatively small" (p. 469) and not statistically significant. Babad and Katz (1991) found that wishful thinking was clearly evident in the bets of sports fans, indicating that a financial incentive for objectivity did not eliminate wishful thinking. Similarly, Irwin and Snodgrass (1966) and Irwin and Metzger (1966, 1967) found that requiring subjects to bet on their predictions in a laboratory task did not eliminate DB.

So, there appears to be mixed evidence regarding people's ability to become aware of bias and correct it. Wegener and Petty's (1995) findings show that people understand the idea that irrelevant contextual factors can bias their judgments, and that, at least when prompted, they will adjust their judgments according to their theories of how the specific contextual factor is likely to have biased them. However, it is less clear whether people can reduce DB, or under what circumstances they might be willing and able to do so. The main conclusion we draw from the evidence reviewed above is that there exist circumstances in which people are either unwilling or unable to reduce or eliminate their DB in response to a request to do so or a financial incentive for doing so. Whether there are also circumstances in which people are willing and able to reduce or eliminate DB is an open question.

Simultaneous Bias and Awareness of Bias

The type of awareness of bias of most interest in the present context is awareness of bias while the bias remains in effect. This is what we are labeling subjective bias and, specifically in the case of desire-related bias, Subjective DB. The logical possibility of simultaneous bias and awareness of bias, at least in a domain other than belief, can be

demonstrated by imagining wearing rose colored contact lenses. One could either be aware that one's visual perception was biased to make the world look rose colored, or one might be unaware of this. However, awareness of this visual bias would not stop the world from looking rose colored. Perhaps something analogous is possible with beliefs.

We noted above that Kirkpatrick and Epstein (1992) and Denes-Raj et al. (1995), found that people attributed greater ratio bias to others than they admitted themselves. However, the other side of this coin is that people did admit to some degree of bias themselves. Epstein's (1990, 1994) Cognitive-Experiential Self Theory predicts that people will often admit to bias, even while remaining affected by it. Epstein posits the existence of a two separate conceptual systems, an evolutionarily older experiential system and an evolutionarily newer, and uniquely human, rational system. The two systems operate in parallel, according to very different principles. As a result, people's "estimates of objective probability can be very different from their experience of subjective probability" (Kirkpatrick and Epstein, 1992, p. 536), and people are often aware of disagreements between their rational and experiential systems, for example when they talk of a conflict between their "head" and their "heart". Kirkpatrick and Epstein (1992) found evidence for awareness of ratio bias in comments their subjects made, such as "'I know the odds are the same, but 10 winners seem more hopeful than 1' and 'The odds are 1:10 in both bowls, so there's really no difference, but the chances of picking a winner seem better with the bowl that has 10 winners in it'" (p. 539). In addition, many subjects were willing to pay to choose the 10 in 100 option, despite admitting that they knew the odds were the same in both bowls. Their rational knowledge that the two probabilities were identical apparently did not correct the

ratio bias that produced a compelling intuitive feeling that the chances of winning were higher in the 10 in 100 bowl than in the 1 in 10 bowl. For those who were willing to pay, the intuitive belief proved a more powerful influence on their behavior than the rational belief.

More direct evidence about awareness is provided by Epstein, Lipson, Holstein, and Huh (1992) in their studies of the "if only" effect. They gave people pairs of scenarios in which a negative outcome, such as a car accident, happened in two ways. In one scenario, the outcome happened in a way that invited the thought that "if only" the protagonist had done something differently the negative outcome could have been avoided. They asked each subject (a) how much more or less foolish most people would feel in the "if only" scenario than in the other scenario, (b) how much more or less foolish the subject him or herself would feel in the "if only" scenario, and (c) to "give a strictly logical response . . . . Put your emotions aside, and decide who actually behaved more foolishly in terms of bringing about the unfortunate event that occurred" (p. 332). Because the protagonist could not have anticipated that doing the minor thing differently would have had different consequences than doing it the way he or she did it, there was no rational reason to feel any more foolish in the "if only" scenario. People do tend to feel more foolish in that scenario, however. The evidence for awareness of this "if only" bias was that when asked to give a "strictly logical response", about 46% of subjects correctly indicated that there was actually no difference in how foolishly the two protagonists behaved, but when asked about their own response, only about 11% of subjects said they themselves would feel equally foolish in both scenarios, whereas 79% said they would feel more foolish in the "if only" scenario. In a second study, identical to the first except that subjects were asked for the strictly logical response first

rather than last, 63% correctly identified the logical response, but only 12% said it was how they would respond, and about 83% said they would feel more foolish in the "if only" scenario. So, just after indicating that there is actually no difference in foolishness between the two scenarios, many people admitted that they would feel more foolish in the "if only" scenario. This seems to be a demonstration of simultaneous bias and awareness of bias.

Nevertheless, we designed Study 1 thinking that Subjective DB ought to be be zero; we presumed that people cannot knowingly violate their own standards for belief formation. So our null hypothesis, which we did not really expect to reject, was that Subjective DB would not differ significantly from zero. The alternative hypothesis we had in mind was that Subjective DB would be positive. There is much evidence for positive Objective DB, as reviewed above, so we assumed that if people thought they were biased by their desires, they would think they were positively biased by their desires.

Study 1

In Study 1 we asked subjects to judge the probabilities that each of 16 statements about their future lives were true, and very briefly to write down the reasons behind their answers. We then asked them to reflect on how they would like themselves and others to think, ideally, and to say what probabilities they would have assigned if they thought in that way. Next we asked them, for each statement on which their initial and ideal probabilities differed, to endorse one or more of 12 explanations for the discrepancy, and then to indicate, on a scale from +5 to -5, how happy or unhappy they would feel if they knew each statement were true. We then asked them to judge the probabilities again, telling them to feel free to look at their answers to previous questions, such as their initial and ideal probability

judgments. Finally, they completed a 16 item self-report scale related to DB.

Our main (null) hypothesis was if we found any differences at all between answers to the initial and ideal probability questions, they would merely be random adjustments resulting from additional time to think, and would be unrelated to desire. That is, we expected to find that Implicit Subjective DB was zero. Our alternative hypothesis was that Implicit Subjective DB would be positive, meaning that the more a subject desired a statement to be true, the more his or her initial probability would be higher than his or her ideal probability. Including the final probability question allowed us to assess whether ideal probabilities, if they turned out to differ systematically from initial probabilities, represented revised beliefs (bias correction) or simultaneous awareness of bias while the bias remains in effect. In the former case, final probabilities should match ideal probabilities, whereas in the latter case final probabilities should match initial probabilities.

Methods

Subjects

Subjects were students from the University of Pennsylvania and Philadelphia College of Pharmacy and Science. Fifty-eight subjects completed the Study 1 questionnaire, but three gave nonsensical responses, resulting in usable data from 36 females and 19 males. Their ages ranged from 18 years to 33 years 2 months with a median of 19 years.

Materials

The first part of the questionnaire (see Appendix A) consisted of six questions each of which referred to 16 statements. The first eight of these statements expressed desirable beliefs and the second eight expressed undesirable beliefs. Subjects wrote their responses

on a separate answer sheet, and were asked to answer the questions on each given page before turning to the next page. The first question simply asked how likely each statement was to be true. We will refer to answers to this question as initial probabilities. The second question asked subjects to list very briefly the main reasons for each of the beliefs they had written down. The third question asked what probability subjects would have assigned if they thought in the way they would want themselves and others to think ideally. We will refer to answers to this question as ideal probabilities. The fourth question asked subjects, for each statement on which their initial and ideal probabilities differed, to endorse one or more of 12 listed explanations for the difference. The fifth question asked how subjects would feel if they knew that each statement were true, on a scale from -5 to 5, where -5 means "as unhappy as anything could make me", 5 means "as happy as anything could make me", and 0 means "neither happy nor unhappy". The sixth question asked subjects to answer the first question ("how likely is each statement to be true?") again, and told them to feel free to look at their answers to previous questions, such as the first and third. We will refer to answers to this question as final probabilities.

The second part of the questionnaire (Appendix B) consisted of a self-report scale concerning DB. It comprised four pairs of items regarding how people should think, and four corresponding pairs of items regarding how the subject actually thinks. Within each set of four pairs, two pairs asked directly about DB, and two pairs asked about standards for confidence in belief. Of the two pairs directly about DB, one pair was about confidence in beliefs being influenced by what one wants to be true or false, versus not being influenced by that; the other pair was about facing the truth even if it hurts, versus believing what makes

one feel good, even if it is false. Of the two pairs about standards, one pair asked whether confidence in beliefs is, or should be, influenced by hunches; the other pair asked whether confidence in beliefs depends, or should depend, only on the kinds of evidence that most people would consider relevant. We included the items about standards for confidence because we assumed that the more someone bases their confidence in beliefs on hunches, or on kinds of evidence most people would not consider relevant, the easier that person finds it to become confident in the truth of desired beliefs. Subjects expressed their agreement or disagreement with each statement on a five point scale. We calculated the total score by subtracting each anti-DB item rating from its corresponding pro-DB statement and adding the eight results. In addition, we calculated subscales from the two halves of the scale, one about how people should think, the other about how the subject actually thinks.

Procedure

Subjects came to a lab with regular advertised opening hours and a variety of questionnaires, including the one described above, available to be completed for payment of $6 per hour. Subjects were free to choose which questionnaires to complete and in what order. Subjects who had partially completed a questionnaire when the lab closed were allowed to leave it with the lab supervisor and come back to complete it another day.

Results

Implicit Subjective Desirability Bias

For each subject, we estimated the following regression equation, using 32 probability judgments as cases. "Happy" is the subject's rating of how happy or unhappy they would be if they knew each statement were true, and "Initial" is a dummy variable coded

as 1 for initial probabilities and 0 for ideal probabilities:

$$\text{Probability judgment} = \beta_0 + \beta_1 \text{ Happy} + \beta_2 \text{ Initial} + \beta_3 \text{ Happy} \times \text{Initial} + \text{error}$$

We took the coefficient of the interaction term, β_3, as a measure of Implicit Subjective DB. Positive values of β_3 indicate positive Implicit Subjective DB, because they indicate that desire (i.e., happiness if the statement were true) was a more positive (or less negative) predictor of initial probabilities than of ideal probabilities. In other words, the initial probabilities of subjects with positive values of β_3 were more congruent with their desires than the judgments they said they would have given if they thought the way they would ideally like themselves and others to think. Conversely, negative values of β_3 indicate negative Implicit Subjective DB, because they indicate that desire was a less positive (or more negative) predictor of initial probabilities than of ideal probabilities. In other words, the initial probabilities of subjects with negative values of β_3 were less congruent with their desires than the probabilities they said they would have given if they thought the way they would ideally like themselves and others to think.

The mean Implicit Subjective DB (i.e., β_3) was -0.200, which was not significantly different from zero (SD = 4.069, t (54) = -0.364, p = .717). The mean Implicit Subjective DB coefficients for females and males were also not significantly different from zero, or from each other. The same was true, for all subjects and for males and females separately, when we calculated Implicit Subjective DB using final probabilities in place of initial probabilities. While this could reflect a true Implicit Subjective DB of zero, with variance between subjects being mere noise, another possibility is that Implicit Subjective DB is not zero for all subjects, but is negative for some and positive for others, resulting in a mean

close to zero. (Recall, for example, that Slovic, 1966, found individual differences in Objective DB which "tended to cancel one another when data were averaged over Ss," p. 28.)

To investigate whether the variance in Implicit Subjective DB between subjects was mere noise or reflected individual differences, we assessed the split half internal consistency reliability of our Implicit Subjective DB measure. The mere noise hypothesis predicts a reliability of zero. We ran two sets of within subject regressions to calculate separate Implicit Subjective DB measures based on the odd numbered and even numbered life events. The correlation across subjects between these two measures of Implicit Subjective DB is shown in Table 2, and indicates split half internal consistency reliability of about .8, clearly falsifying the mere noise hypothesis. We obtained similar split half correlations when we used final probabilities in place of initial probabilities, and when we considered male and female subjects separately, using either initial or final probabilities (see Table 2). Apparently for many of our subjects, Implicit Subjective DB was not zero; it was negative for some and positive for others.

Explicit Subjective Desirability Bias

We derived a measure of Explicit Subjective DB by subtracting the number of life events for which a subject endorsed the statement "My answer to question A was influenced more by pessimism" to explain why his or her initial and ideal probabilities differed, from the number of life events for which she or he endorsed the statement "My answer to question A was influenced more by optimism, by what I wished were true". As with Implicit Subjective DB, the mean Explicit Subjective DB of 0.745 was not significantly different

from zero, t (54) = 1.208, p > .2. However, as shown in Table 2, the Explicit Subjective DB measure had a split half internal consistency reliability above .7, and significant positive correlations of about .5 with Implicit Subjective DB measures derived using initial or final probabilities.

Desirability Bias Self-Report Scale

Table 2 shows Cronbach's alphas for the DB self-report scale and subscales, and their correlations with each other and with Implicit and Explicit Subjective DB. The full scale's reliability is over .7, the reliability of the "how I think" subscale is under .6 and that of the "how people should think" subscale is just over .4. The negative correlations of about -.3 between the "how I think" subscale and Implicit Subjective DB are significant, and the nonsignificant correlations with the "how people should think" subscale and the full scale are also negative. However, if the significance threshold were Bonferroni-corrected, none of these correlations would be significant. The DB self-report scale and subscales also have small nonsignificant negative correlations with Explicit Subjective DB.

If it represents a real effect, the -.3 correlation between Implicit Subjective DB and the "how I think" subscale means the following. The more negative a subject's Implicit Subjective DB with respect to the specific life event statements on the questionnaire, the more that subject tended to agree that their confidence in their beliefs is affected by what they want to be true or false, that it is influenced by hunches, that it depends on the kind of evidence they consider relevant but most people would not consider relevant, and that they believe what makes them feel good, even if it is false (rather than facing the truth even if it hurts).

Final Probabilities

Subjects' final probabilities were closer to their initial probabilities than to their ideal probabilities. The overall mean position of final probabilities on a scale where the initial probability is represented by 0 and the ideal probability is represented by 1 was 0.196 (95% confidence interval = 0.118 to 0.274, df = 54). In other words, subjects' final probabilities moved only about 20% of the distance from their initial probabilities to their ideal probabilities.

Discussion

Although neither the mean Implicit Subjective DB nor the mean Explicit Subjective DB differed significantly from zero, our null hypothesis that Implicit Subjective DB would be zero can nevertheless be rejected because we have evidence that for many subjects, both Implicit Subjective DB was different from zero. The overall mean Implicit Subjective DB was not different from zero only because negative and positive Implicit Subjective DB canceled each other out. The high internal consistency reliabilities of our measures of Implicit Subjective DB indicate that individual variation about the mean was not mere noise but reflected differences between subjects.

Of course, differences between subjects are not necessarily meaningful differences. It might be argued that each subject, when asked to give ideal probabilities, had no real basis for giving judgments different from their initial judgments, but assumed that the study required them to do so. Preferring, for some reason, not to give randomly different ideal judgments, they searched for an arbitrary basis for forming ideal judgments and settled on making the ideal judgments differ from the initial judgments in a way related to their desire.

They then settled on a random and arbitrary direction and magnitude for the relationship to desire.

Evidence against that argument is provided by the Explicit Subjective DB measure and its correlation with Implicit Subjective DB. If subjects made their ideal judgments differ from their initial judgments in a way arbitrarily and randomly related to desire, then one would not expect them to report that their initial judgments were more influenced by optimism or pessimism than their ideal judgments. If optimism or pessimism had anything to do with the judgments, then the relationship to desire would not be arbitrary. Subjects would, however, be expected to endorse the statement: "There was no particular reason why my answers were different. The difference between the two answers could just as easily have been in the opposite direction." In fact, the mean number of statements for which subjects endorsed that reason was only 0.673, compared to 2.673 for the optimism reason (t (54) = 4.051, $p < .0002$) and 1.927 for the pessimism reason (t (54) = 3.632, $p < .0007$). Furthermore, the number of endorsements of the optimism reason minus the number of endorsements of the pessimism reason yielded a measure of Explicit Subjective DB which had good internal consistency reliability and was significantly and quite strongly positively correlated with Implicit Subjective DB. This renders implausible the argument that the differences between subjects in Implicit Subjective DB were random or arbitrary.

Does the difference between initial and ideal probabilities really show that subjects disbelieved their own beliefs? Evidence that the ideal beliefs existed simultaneously with the actual beliefs, rather than merely replacing them, is provided by the facts that subjects' final probabilities were much closer to their initial probabilities than to their ideal

probabilities, and that we found essentially identical results for Implicit Subjective DB regardless of whether we calculated it using initial or final probabilities.

If one grants that ideal and actual beliefs existed simultaneously, the question arises whether a conflict between an ideal belief and an actual belief amounts to disbelief of the actual belief? The answer is that it does in a person whose ideal is to have accurate beliefs. Such a person's ideal belief, if it differs from their actual belief, must differ in that they consider it more accurate (otherwise it would not be more ideal). To consider a belief different from one's actual belief to be more accurate than one's actual belief truly amounts to disbelieving one's actual belief.

With this in mind, we identified the approximately one third of subjects who most disapproved of DB on the "how people should think" DB self-report subscale. Their disapproval of DB ranged from -3 to -11 on a scale with a theoretical maximum range of -16 (strongest disapproval of DB) to +16 (strongest approval of DB). The Implicit Subjective DB scores of these 16 subjects were similar to those of the other 39 subjects in internal consistency reliability (split half rs = .868 and .845 respectively) and variability (SDs = 3.631 and 4.206, respectively), so like other subjects these subjects' ideal beliefs differed from their initial beliefs in desire-related ways. Furthermore, their final beliefs were even more in accord with their initial beliefs than were those of other subjects; the mean position of their final probabilities on a scale where the initial probability is represented by 0 and the ideal probability is represented by 1 was 0.066, compared to 0.250 for other subjects. Therefore, these 16 subjects really seem to have disbelieved their own beliefs. Their responses to the DB self-report scale showed that they would ideally like themselves and others to think more

accurately, with less DB; and their final beliefs showed that they did not actually believe what they had just said they would believe if they thought in that more accurate way.

The indication that subjects who admitted to DB on our "how I think" subscale may have tended to have more negative and/or less positive Subjective DB is surprising. While three of the four pairs of items in the "how I think" subscale did not clearly imply positive rather than negative DB (e.g., "My own confidence in my beliefs is affected by what I want to be true or false"), one pair very clearly did (e.g., "I believe what makes me feel good, even if it is false."). So, if anything, one might expect a positive correlation between this subscale and Subjective DB. One possible explanation for why the correlations were negative would be if, on the "how I think" subscale, subjects answered more in accord with how they would like to think than with how they actually think. People with negative Subjective DB are people who indicated that if they thought the way they would ideally like themselves and others to think, their beliefs would be more congruent with their desires. Perhaps when they expressed agreement with items such as "I believe what makes me feel good, even if it is false" , those people meant something more like "I aim to believe what makes me feel good, even if it is false."

Study 2

While Study 1 showed that people can knowingly violate their own ideals for thinking, we noted that the significance of this finding depends on just what people's ideals for thinking are. Nonzero Subjective DB defined relative to a subject's own ideal way of thinking amounts to disbelief of one's own beliefs only in subjects whose ideal is to think accurately. With the help of our self-report DB "how I think" subscale, we were able to

show that nonzero Subjective DB was just as prevalent among subjects whose responses on that scale implied that their ideal was to think accurately. However, a more direct way to demonstrate the existence of nonzero Subjective DB that amounts to disbelief of one's own belief is to ask subjects a different question; instead of asking them what they would believe if they thought in the way they would ideally like themselves and others to think, we asked our Study 2 subjects what they would believe if they thought in a way that would maximize the accuracy of their beliefs. By specifying the ideal relative to which bias is defined, we can ensure that any non-zero Subjective DB we find can be interpreted as disbelief of one's own belief no matter what a subjects's own ideals might be.

We suspected that if some of our Study 1 subjects did not aspire to the ideal of accuracy, the reason might be that they felt beliefs affected by positive or negative DB could be more effective at helping them achieve their goals, for example by increasing their motivation or self-confidence. To explore this possibility we asked our Study 2 subjects, in addition to the question about accuracy, what they would believe if they thought in a way that would maximize the effectiveness of their beliefs at helping them achieve all their goals. This allowed us to find out whether our subjects thought that greater positive or negative DB could help them achieve their goals more effectively. To investigate whether Subjective DB is related to people's ideals for thinking, we included a self-report scale designed to measure subjects' ideals about whether good thinking should produce accurate beliefs, effective action, or good feelings.

We also included Fenigstein, Scheier, and Buss's (1975) Private Self-Consciousness scale, hypothesizing that if, as research suggests, most people actually have positive

Objective DB, then people who are more self-aware will tend to know this about themselves, and may have more positive Subjective DB than less self-aware people do.

Finally, we included Ryff's (1989) scales of Psychological Well-Being. We thought Subjective DB might be related to well being. People with low levels of well-being might have negative Subjective DB, because they know they experience the world more negatively than most people do; people with high levels of well-being might have positive Subjective DB, because they realize they have an unusually positive outlook on life.

Method

Subjects

Ninety nine subjects provided usable data via the Study 2 questionnaire on the world wide web; 33 were male, 65 were female, and one did not report his or her gender or age. The 98 subjects who reported their age ranged from 17 to 26 years old with a median age of 19 years. All but four of the subjects gave addresses to send their payment to; 66 were in Pennsylvania (59 in Philadelphia), 20 were in other U.S. states (California, Massachusetts, Virginia, Illinois, New Jersey, New York, West Virginia, Iowa, Florida, and Hawaii), five were in Canada (Ontario and Quebec), and the remaining four were in England, Australia, Bulgaria, and Turkey.

The questionnaire stated immediately below its title, with the first sentence in bold type, "This study is for undergraduate students only. If you are not an undergraduate student, then I appreciate your interest, but please do not fill out this study because essential parts of it will not apply to you". So we believe that all the subjects were undergraduate students, but we cannot be completely sure of this.

Three people completed the questionnaire twice, and we ignored their second submissions. We also ignored six submissions that were incomplete and lacked names, and two submissions with answers such as "yes" to questions requiring numerical answers.

Materials

The questionnaire was implemented as a world wide web page divided into two frames side by side. The instructions included the request "Please complete this study in the order given, answering each question fully before proceeding to read and respond to the next one." The answer areas appeared in the left frame, and the questions appeared in the right frame one at a time. In the right frame each question was followed by a place to click the mouse to bring up the corresponding answer area in the left frame. In the left frame each answer area had below it a place to click the mouse to bring up the next question in the right frame. If subjects scrolled down beyond their current answer area in the left frame they encountered blank space and a repeating message such as "This is space between the Question D and Question E response areas. Please do not scroll down to Question E until you have finished Question D."

The questionnaire consisted of five questions relating to 16 statements, followed by three personality scales. The 16 statements were identical to those in Study 1, except that "(in the future)" was appended to statements 14 and 15. Subjects typed their answers to the first three questions into three input spaces to the right of each statement. As in Study 1, the first question simply asked how likely each statement was to be true. We will again refer to answers to this question as initial probabilities. The second and third questions asked, in counterbalanced order, what probabilities subjects would have assigned if they thought in a

way (a) “that would lead you to form beliefs that are as accurate (or true) as possible”, and (b) “that would lead you to form beliefs that would maximize your achievement of your goals”. The full text of these questions is shown in Appendix C. We will refer to answers to these questions as accuracy probabilities and effectiveness probabilities respectively. The fourth set of questions asked subjects how strongly their answers to each of the first three questions were influenced by evidence, by intuitions and hunches, by optimism, and by pessimism. Subjects answered these 12 questions on a scale from 0 to 6, where 0 meant "Not at all influenced", 3 meant "Somewhat influenced", and 6 meant "Extremely influenced". The fifth question asked subjects how much they wanted each statement to be true or to be false, on a scale from -5 to 5, where -5 meant, “I want the statement to be FALSE as much as I want anything," 0 meant, "I neither want the statement to be true nor want it to be false," and 5 meant, "I want the statement to be TRUE as much as I want anything."

The first personality measure in the questionnaire was Ryff’s (1989) Psychological Well-Being scale. The version we used had 54 items, which were presented in a random order that was the same for all subjects. The second personality measure was Fenigstein et al.’s (1975) 10-item Private Self-Consciousness scale. The third and last was a 48-item Thinking Ideals scale designed to measure ideals as to whether good thinking should result in true beliefs, effective action, or positive feelings. Each of the three subscales, which we will refer to as Truth, Effectiveness, and Good Feelings subscales, consists of eight pairs of items. The items, all of which are listed in Appendix D, were presented in a random order that was the same for all subjects.

Procedure

Subjects visited a page on the world wide web with a variety of questionnaires, including the one described above, available to be completed for mailed payment based on an estimate of $1 for 10 minutes of very careful reading and responding by a slow reader. The present questionnaire paid $6. Subjects were free to choose which questionnaires to complete and in what order.

Results

Implicit Subjective Desirability Bias Relative to Accuracy and Effectiveness Ideals

For each subject, we estimated two regression equations as follows, each using 32 probability judgments as cases. "Want" is the subject's rating of how much they want each statement to be true or false, "Initial vs. Accuracy" is a dummy variable coded as 1 for initial probabilities and 0 for accuracy probabilities, and "Initial vs. Effectiveness" is a dummy variable coded as 1 for initial probabilities and 0 for effectiveness probabilities:

(a) Probability judgment = β_{A0} + β_{A1} Want + β_{A2} Initial vs. Accuracy + β_{A3} Want × Initial vs. Accuracy + error

(b) Probability judgment = β_{E0} + β_{E1} Want + β_{E2} Initial vs. Effectiveness + β_{E3} Want × Initial vs. Effectiveness + error

We considered the coefficient of the interaction term from equation (a), β_{A3}, to be a measure of Implicit Subjective DB relative to an accuracy ideal, and we considered the corresponding coefficient from equation (b), β_{E3}, to be a measure of Implicit Subjective DB relative to an effectiveness ideal. That is, in (a), "bias" means deviation from the most accurate belief possible, whereas in (b), "bias" means deviation from the most effective belief possible.

The mean Implicit Subjective DB relative to an accuracy ideal (β_{A3}) was 0.732, and

was significantly different from zero, t (98) = 2.906, SD = 2.507, p < .005. The mean Implicit Subjective DB relative to an effectiveness ideal (β_{E3}) was -2.081, and was also significantly different from zero, t (98) = -6.895, SD = 3.003, p < .00001. These two means were significantly different from each other, t (98) = 8.746, SD = 3.200, p < .00001. Males and females did not differ significantly on Implicit Subjective DB relative to an accuracy ideal or on Implicit Subjective DB relative to an effectiveness ideal.

As shown in Table 3, the measures of the two types of Implicit Subjective DB both have split half internal consistency reliabilities over .7, and the correlation between them is under .4.

Explicit Subjective Desirability Bias

We derived a two-item measure of Explicit Subjective DB relative to an accuracy ideal from subjects' ratings of how strongly their initial and accuracy probabilities were influenced by optimism and by pessimism. We formed one item as [rated influence of optimism on initial probabilities] minus [rated influenced of optimism on accuracy probabilities]. Positive values of this item mean a subject said optimism influenced her initial probabilities more than her accuracy probabilities. We formed the other item as [rated influence of pessimism on accuracy probabilities] minus [rated influence of pessimism on initial probabilities]. Positive values of this item mean a subject said pessimism influenced his accuracy probabilities more than his initial probabilities. As shown in Table 3, Cronbach's alpha for this two-item measure is about .35, Cronbach's alpha for the corresponding two-item measure of Explicit Subjective DB relative to an effectiveness ideal is about .7, and the correlation between the two measures of Explicit Subjective DB is

under .4.

Table 3 also shows that Explicit Subjective DB relative to an accuracy ideal is correlated over .5 with Implicit Subjective DB relative to an accuracy ideal, but only about .1 with Implicit Subjective DB relative to an effectiveness ideal. Similarly, Explicit Subjective DB relative to an effectiveness ideal is correlated over .5 with Implicit Subjective DB relative to an effectiveness ideal, but only about .1 with Implicit Subjective DB relative to an accuracy ideal.

The mean Explicit Subjective DB relative to an accuracy ideal was 2.010, which was significantly different from zero, t (95) = 6.105, SD = 3.227, $p < .0001$. The mean Explicit Subjective DB relative to an effectiveness ideal was -1.146, which was also significantly different from zero, t (95) = -3.443, SD = 3.261, $p < .001$. These two means were significantly different from each other, t (95) = 8.456, SD = 3.657, $p < .0001$. Males and females did not differ significantly on Explicit Subjective DB relative to an accuracy ideal or on Explicit Subjective DB Relative to an effectiveness ideal.

Personality Scales

As shown in Table 3, the Thinking Ideals scale, the Psychological Well-Being scale, and the Private Self-Consciousness scale all showed adequate internal consistency reliability in this sample. However, none of these scales correlated significantly with any of the Subjective DB measures, or with each other.

The three subscales of the Thinking Ideals scale all had adequate internal consistency reliability, as shown on Table 4. The strong positive correlation between the Effectiveness subscale and the Feeling Good subscale indicates that the more a subject felt that thinking

is good when it is effective, the more that subject tended to feel that thinking is good when it leads to positive feelings. The negative correlations between the Truth subscale and the other two subscales indicate that the more a subject felt that thinking is good when it leads to accurate beliefs, the less that subject tended to feel that thinking is good when it leads to effective action or to positive feelings.

Before Bonferroni correction, the Truth subscale of the Thinking Ideals scale was significantly negatively correlated with Implicit Subjective DB Relative to Accuracy, $\underline{r} = -.212$, $\underline{n} = 98$, $\underline{p} < .05$, and marginally significantly negatively correlated with Explicit Subjective DB Relative to Accuracy, $\underline{r} = -.185$, $\underline{n} = 96$, $\underline{p} < .08$. If meaningful, these negative correlations would imply that the more a subject felt that thinking is good when it leads to accurate beliefs, the less positive, or the more negative, was their Subjective DB Relative to Accuracy. However, appropriate Bonferroni correction would render these correlations nonsignificant. All other correlations between Thinking Ideals subscales and measures of Subjective DB were nonsignificant.

Discussion

The primary purpose of Study 2 was to test whether we would still find nonzero Subjective DB when we defined bias relative to the ideal of accuracy, rather than relative to subjects' own ideals for thinking. We did.

Recall that we originally designed Study 1 to decide between the null hypothesis that Subjective DB would be zero and the alternative hypothesis that it would be positive. Because much research exists showing that people often have positive Objective DB, we originally thought that either people would deny having any DB, or they would admit to

positive DB. However, the research in question shows positive Objective DB relative to an accuracy ideal, not necessarily relative to people's personal ideals for thinking. So a better version of our original alternative hypothesis is that Implicit Subjective DB relative to an accuracy ideal would be positive.

The results of Study 2 support that hypothesis. On average, our subjects did show positive Implicit and Explicit Subjective DB relative to an accuracy ideal. However, they showed quite strongly negative Implicit and Explicit Subjective DB relative to an effectiveness ideal. In other words, as a group our subjects felt that their actual beliefs were inaccurately congruent with their desires, but that to be maximally effective at helping them achieve all their goals, their beliefs should be even more inaccurately congruent with their desires. To make this more concrete, the mean of 0.732 for Implicit Subjective DB relative to an accuracy ideal (β_{A3}) implies that subjects collectively felt they overestimate the probabilities of propositions they rate +5 in desirability by (5 × 0.732 =) 3.7 percentage points. The mean of -2.081 for Implicit Subjective DB relative to an effectiveness ideal (β_{E3}) implies that subjects collectively felt that it would be maximally effective to overestimate the probability of propositions they rate +5 in desirability by (5 × 2.081 =) 10.4 percentage points more than the 3.7 points by which they think they already do overestimate. Overall, then, our subjects implicitly recommended overestimating the probabilities of strongly desired outcomes by about (3.7 + 10.4 =) 14 percentage points, and underestimating the probabilities of strongly undesired outcomes by the same amount.

As in Study 1, the Explicit Subjective DB measures closely mirrored the Implicit Subjective DB measures, providing more evidence that people are consciously aware of the

relationship between their subjective estimates of bias and their desires.

The lack of strong correlations between the Thinking Ideals scale and subscales and the Subjective DB measures suggests that Subjective DB does not reflect people's ideals for how their thinking should be, which makes it more likely that it does reflect how they actually perceive their thinking to be.

The pattern of intercorrelations among the three Thinking Ideals subscales indicates that the more a subject thought good thinking should achieve accurate beliefs, the less that subject tended to think good thinking should achieve effective action and good feelings. However, the more a subject thought good thinking should achieve effective action, the more that subject thought good thinking should achieve good feelings. This implies that our subjects saw good feelings and effective action as compatible goals, but saw both of these goals as somewhat incompatible with the goal of truth. Perhaps the most interesting aspect of this is the implied view that accurate beliefs are not very helpful in facilitating effective action--a view also implied by the negative correlation between the Truth and Effectiveness subscales of the Thinking Ideals scale.

The lack of significant correlations between Psychological Well-Being and the Subjective DB measures imply that Subjective DB is not related to well-being as a trait. This leaves open the possibility that Subjective DB is an unstable individual difference like mood, rather than a stable personality trait, and might be related to more transitory measures of well-being, such as mood. It may be, however, that Subjective DB is not related to well-being at all. Well-being may be related to how one actually sees the world, not to the way of seeing the world one thinks would be maximally accurate or effective.

Correlations between Private Self-Consciousness and Subjective DB measures might be expected if most people have similar Objective DB, because variation in Subjective DB would then reflect variations in self-awareness rather than differences in Objective DB. Our finding of no significant correlations between Subjective DB and Private Self-Consciousness could lead one to speculate that perhaps people differ more in Objective DB than research on the topic has shown. However, the lack of correlation could have many other causes, and little can be concluded from it.

Study 3

One limitation of Studies 1 and 2 is that it was impossible to compare Subjective DB to Objective DB, since the right answers to the probability judgment questions were not known. In Study 3 we attempted to measure both Subjective DB and Objective DB, by asking for probabilities not of life events, but of outcomes of computer controlled games of chance. Our hypothesis was that Subjective DB would accurately reflect Objective DB, at least to some degree. In other words, we hypothesized that when asked what judgments they would have given if they thought in a way that would maximize the accuracy of their beliefs, subjects would give judgments that showed less Objective DB than their initial judgments. We also hypothesized that this effect would be stronger for people high in Private Self-Consciousness.

We were also interested in exploring possible differences in Objective and/or Subjective DB depending on whether one expects to discover the truth in the near or far future. Our hypothesis about this was that when people expect to discover the truth soon, they may use negative DB in order to reduce the potential disappointment, whereas when

they will not discover the truth for a while, positive DB can make them feel better for the time being without fear that the positive feelings will imminently be disrupted by an undesirable truth. Besides hypothesizing that this may true of Objective DB, we also wondered whether people might share our intuitions and see themselves as having this pattern of positive DB for remote outcomes and negative DB for imminent outcomes. So we asked each subject both about a set of imminent outcomes and a set of remote outcomes.

Study 3 also included a new personality scale intended to measure Subjective DB as a trait, by asking directly about tendencies towards positive or negative DB with regard to beliefs about the self, the future, and the world. If we were to find strong correlations between this measure and our task-specific measures of Implicit and Explicit Subjective DB, it would suggest that the latter reflect stable traits, at least in part. We expected Scheier and Carver's (1985) Life Orientation Test, a measure of optimism in the simple sense of expecting desirable outcomes, to be moderately correlated with the Trait Subjective DB scale, since the more optimistic one is, the more plausible it is that part of one's optimism is due to positive DB, and the more pessimistic one is, the more plausible it is that part of one's pessimism is due to negative DB. However, we did not expect the two scales to be highly correlated because two equally optimistic or pessimistic people could differ in how much of their optimism or pessimism they attribute to DB.

We also included a Composite Actively Open Minded Thinking scale similar to that used by Stanovich and West (1997). We hypothesized that open-minded thinking would be positively related to the absolute value of Implicit and Explicit Subjective DB, on the basis that open-minded people may be more aware of and willing to admit to having biases.

We also included Epstein, Pacini, Denes-Raj, and Heier's (1996) Rational-Experiential Inventory. It consists of a Faith in Intuition scale, which measures the tendency to think in an intuitive or experiential way, and a modified version of Cacioppo and Petty's (1982) Need for Cognition scale, which measures tendency to think in an analytical or rational way. Epstein et al. (1996) show that these two scales measure independent dimensions. They also suggest that "strong experientiality may interfere with logical thinking; that is, people who are strongly experiential tend to accept their heuristic thinking as rational". Based on this, we hypothesized that subjects high in Faith in Intuition would have Subjective DB closer to zero than other subjects, especially if they were also low on Need for Cognition.

We also included the Fenigstein et al. (1975) Private Self-Consciousness scale, hypothesizing that people high on this scale may have more accurate Subjective DB.

Method

Subjects

Fifty four subjects contributed usable data by completing the Study 3 questionnaire on the world wide web; 32 were female and 22 were male. Data from an additional nine subjects were excluded; three did not give addresses for payment, rendering the main manipulation invalid for them, one gave nonsensical probability estimates indicating she either seriously misunderstood the task, or did not take it seriously, and five gave probability estimates which bore very little relationship to the actual probabilities, indicating they did not take the task seriously[4].

Unfortunately, due to computer problems, ages were correctly recorded for only 15

of the 54 subjects. The ages of these subjects ranged from 15 to 48 years old, with a median age of 23 years 11 months.

Forty of the subjects gave addresses for payment in the U.S.A. (nine in Pennsylvania, four each in New Jersey and Virginia, three each in California and Illinois, two each in Colorado, Missouri, and Washington, and one in each of 11 other states), six were in Canada (two each in Alberta and Nova Scotia, and one each in Ontario and Quebec), four were in Germany, and the remaining four were in Holland, Norway, Turkey, and South Korea.

Twenty seven subjects reported being students. (Seven identified themselves as graduate students, four as college students, and one as a highschool student). Six subjects had computer-related occupations, four social or medical services occupations, four administrative or accounting occupations, two were teachers, two were salespersons, and the others were a researcher, a musician, a graphic designer, a retired person, a homemaker, an unemployed person, an editor, a customer service representative, and an automotive technician.

Materials

First part of questionnaire. The questionnaire was implemented as a questionnaire on the world wide web. The purpose of the first part of the questionnaire (the content of which is illustrated in detail in Appendix E) was to present subjects with a probability judgment task in which the correct answer was known to us, but not to the subjects, who had to rely on an inherently ambiguous visual estimation process. Ambiguity was important to allow room for DB to affect judgments. This purpose was served by a computer program implementing a “game of chance” involving a black rectangle on the computer screen with

white dots rapidly appearing and disappearing at random positions on it. When the "game" was played, the computer would randomly choose a point on the rectangle and see if there was a white dot there at that moment. If so, the result of that game was a "hit", otherwise it was a "miss". The computer program could vary the average number of white dots on the rectangle at any moment so as to set the probability of a hit anywhere in the range from 0 to 1,000 chances in 100,000 In the questionnaire, probabilities were always expressed in chances in 100,000.

Subjects were given ten demonstration trials in which they simply viewed the black rectangle set at ten different probabilities of a hit, and then did 20 practice trials in which they estimated the probability of a hit and were then told the actual probability and their percentage error. The sequence of probabilities in the demonstration and practice trials was the same for all subjects, and was derived by generating 30 random fractional numbers, x, between 1 and 19 and setting the probabilities at 20×1.228^{x} in 100,000. This spaced the probabilities approximately equally on a logarithmic scale, so as to make them approximately equally distinguishable from each other, and made them high enough that it was never possible to count the white dots, and low enough that the computer program could display them.

Following the demonstration and practice trials were two sets of 20 initial probability trials. Each of these trials involved estimating the probability of a hit with a monetary consequence for the subject. In half the trials, the subject stood to win money if that game of chance produced a hit. In the other half of the trials, the subject stood to lose money if that game of chance produced a hit. (As the questionnaire explained to them, subjects could lose

up to $3 of their initial $4, and they could lose some or all of any money they won on another game of chance.) In order to approximate equal intervals of desire, the set of wins and losses was derived from a rough approximation of the prospect theory value function (Kahneman and Tversky, 1979) taken from an exercise in Baron (1994, p. 366). For values of x from 1 to 10, the gains were $\$x^2$ (i.e., $1, $4, $9, ... $100) and the losses were $\$0.5x^2$, rounded to the nearest whole dollar. (i.e., $1, $2, $5, ... $50). To ensure attention to the gains and losses, subjects had to enter their probabilities by typing, for example "win49=350", to estimate the chance of winning $49 as 350 in 100,000.

The probabilities used in the two sets of 20 initial probability trials were equal to 20×1.228^x in 100,000, for integers x from 0 to 19. The 20 probabilities and the 20 monetary consequences were paired in such a way that the magnitude of the correlation between probability and consequence was less than .02, and the magnitude of the correlation between probability and absolute value of consequence was less than .1.

The two sets of 20 initial probability trials were identical in all but two respects. First, the randomized order of the probability-consequence pairs was different in each set. Second, in one set the subject would be told the outcomes of the games of chance the same day, after completing the questionnaire, whereas in the other set, the subject would be told the outcomes several weeks later, via email. The order of these two conditions, the "find out today" condition and "find out later" condition, was counterbalanced across subjects.

After completing the two sets of 20 initial probability trials, subjects were asked to take a few moments to consider the question: "How would you think and perceive if both (a) your only goal was to arrive at the most accurate (or true) beliefs possible, and (b) you had

the ability to eliminate all influences on your thinking and perception that conflicted with this goal of accuracy?". In the two sets of 20 accuracy probability trials which followed, subjects were shown the same sequence of games of chance as in the initial probability trials, were reminded of the probability they had estimated initially, and were asked in each case, for example: "If you thought in a way that would lead to the most accurate beliefs possible, how likely would you have thought it is that you will lose $32 on this game?". To ensure attention to the reminder of their initial judgment, it was displayed in parentheses in the input box, so they had to type over it to give their answer.

Playing the games. Before proceeding to the second part of the questionnaire, subjects played the games of chance. The games were displayed in the same order they had appeared in the initial and accuracy probability trials, and for each game subjects were told (truthfully) that when they pressed a key, the computer would pick a random point on the black rectangle and record whether there was a white dot there, that is, whether the outcome of the game was a "hit" or a "miss".

Second part of questionnaire. The second part of the questionnaire consisted of five personality scales, presented one item at a time. The first one was Epstein et al.'s (1996) Rational-Experiential Inventory (REI), which consists of a modified 19-item version of Cacioppo and Petty's (1982) Need for Cognition scale, followed by a 12-item scale measuring Faith in Intuition. Subjects responded to each item on a five-point scale by typing ct for "completely true", st for "somewhat true", ne for "neutral", sf for "somewhat false", or cf for "completely false".

The second scale was a new 12-item scale intended to measure Subjective DB as a

trait. The items are listed in Appendix E. Six of them imply that the respondent tends towards positive DB, and six imply that the respondent tends toward negative DB. Within each of those two sets of six items, three items concern bias in beliefs about desirable things and three concern biases in beliefs about undesirable things. Each of these four sets of three items includes one item about the self, one item about the world, and one item about the future. Subjects responded to each item on a six-point scale by typing agst for "agree strongly", agmo for "agree moderately", agsl for "agree slightly", disl for "disagree slightly", dimo for "disagree moderately", or dist for "disagree strongly".

The third scale was composed of several scales combined into a 40-item Composite Actively Open Minded Thinking scale similar to the one used by Stanovich and West (1997). The items taken from Stanovich and West (1997) were nine items measuring absolutism, nine items measuring dogmatism, three items measuring categorical thinking, and ten items measuring flexible thinking. We were unable to obtain permission to include two subscales which Stanovich and West (1997) had included from the Revised NEO Personality Inventory (Costa and McCrae, 1992). However, we included an additional nine items written by Sá, West, and Stanovich (1998) to measure "the extent to which people identify their beliefs with their concept of self" (p. 21). A theoretical paper by Cederblom (1989) inspired Sá et al. to develop these items, which they included in a Composite Actively Open Minded Thinking scale similar to the one used by Stanovich and West (1997). We presented our set of 40 items in a random order that was the same for all subjects.

The fourth scale was the 10-item Fenigstein et al. (1975) Private Self-Consiousness scale. The response scale for this and for the Composite Actively Open Minded Thinking

scale was the same as for the Trait Subjective DB scale.

The fifth and last scale was Scheier and Carver's (1985) Life Orientation Test, a measure of optimism consisting of eight real items and four filler items intended to disguise the purpose of the scale somewhat. Subjects responded to each item on a five-point scale by typing sa for "strongly agree", ag for "agree", ne for "neutral", di for "disagree", or sd for "strongly disagree".

Third part of questionaire. Following the personality scales were four self-report items intended to measure Explicit Subjective DB with respect to the probability judgments in the first part of the questionnaire, both before and after the accuracy instruction, and both for the imminent outcome condition and the remote outcome condition. For example: "In my FIRST look at the games that I will learn the outcomes of in a few weeks (Set A), my judgments of the chances of winning or losing by getting a 'hit' were biased [BLANK]." To fill in the blank, subjects chose from a list of nine options ranging from "extremely optimistically" through "neither optimistically nor pessimistically" to "extremely pessimistically".

Three open-ended questions were included to explore subjects' thoughts about the pros and cons of DB. One of them was, "I think it is [BLANK] in my best interest to believe that the future will be better (or less bad) than it probably really will. It is in my best interest when [BLANK] because [BLANK]. It is not in my best interest when [BLANK] because [BLANK]." To fill in the first blank, subjects chose from a list of five options; "almost always", "usually", "sometimes", "not usually", or "never". The second and third blanks were rectangular areas into which subjects could type written responses. The other two items

were identical except that where the first item said "better (or less bad) than", the second item said "worse (or less good) than", and the last item said "exactly as good or as bad as".

Finally, there was a rectangular area in which subjects could type any comments or feedback about the questionnaire, and then a form asking for name, address for payment, etc. After completing that form, subjects were told the outcomes of the "find out today" games of chance, and reminded that they would learn the outcomes of the other games in a few weeks, by email.

Procedure

Subjects visited a page on the world wide web with a variety of questionnaires, including the one described above, available to be completed for mailed payment. Subjects were free to choose which questionnaires to complete and in what order. The present questionnaire paid an initial $4 which could potentially change to an amount from $1 to $204, depending on the outcomes of the games of chance. The actual outcomes resulted in two subjects being paid $1, 58 subjects being paid $4, one subject being paid $8, and one subject being paid $68.

Results

Objective Desirability Bias

To assess Objective DB before and after the accuracy instruction we estimated two regression equations for each subject, one based on the subject's 40 initial probabilities and the other based on the subject's 40 accuracy probabilities. The equation was as follows, where "Probability judgment" is a logarithmic function of the judged probability of a hit, "Actual" is a logarithmic function of the actual probability of a hit[5], "Desire" is the square

root of either the potential number of dollars to be gained or of twice the potential number of dollars to be lost, and "Later vs. Today" is a dummy variable coded as 1 for games of chance the subject would learn the result of in a few weeks, and as 0 for games of chance the subject would learn the result of at the end of the questionnaire:

$$\text{Probability judgment} = \beta_0 + \beta_1 \text{ Actual} + \beta_2 \text{ Desire} + \beta_3 \text{ Later vs. Today} + \beta_4 \text{ Desire} \times \text{Later vs. Today} + \text{error}$$

The coefficient of the Desire term, β_2, was our measure of Objective DB in the "find out today" condition, where "Later vs. Today" is 0. The sum of the coefficients of the Desire term and the interaction term, $\beta_2 + \beta_4$, was our measure of Objective DB in the "find out later" condition, where "Later vs. Today" is 1.

The mean Objective DB in the "find out today" condition (β_2, based on initial probabilities) was 0.045, SD = 0.173, t (53) = 1.915, one-tailed p = .03. The mean Objective DB in the "find out later" condition ($\beta_2 + \beta_4$, based on initial estimates) was 0.001, SD = 0.199, t (53) = 0.044, one-tailed p = .48. The difference between the two conditions was significant, SD = 0.158, t (53) = 2.049, $p < .05$. In neither condition did the mean Objective DB differ significantly between males and females.

The split half internal consistency reliability of the Objective DB measure in the "find out today" condition was .392, and that of the Objective DB measure in the "find out later" condition was .775, so there were individual differences in Objective DB.

Reduction in Objective Desirability Bias

The main hypothesis for Study 3 was that Implicit Subjective DB would be a somewhat accurate reflection of Objective DB. This hypothesis implies that the Objective

DB in subjects' accuracy probabilities should be closer to zero than the Objective DB in subjects' initial probabilities. The absolute value of Objective DB, transformed by raising to the power 0.2 to create a more normal distribution, was indeed lower following the accuracy instruction, both in the "find out today" condition ($\underline{M}_{initial}$ = 0.536, $\underline{M}_{accuracy}$ = 0.509, $\underline{SD}_{difference}$ = 0.098, $\underline{t}$ (53) = 2.019, one-tailed $\underline{p}$ = .024) and in the "find out later" condition ($\underline{M}_{initial}$ = 0.553, $\underline{M}_{accuracy}$ = 0.512, $\underline{SD}_{difference}$ = 0.146, $\underline{t}$ (53) = 2.023, one-tailed $\underline{p}$ = .024). In neither condition did the mean change in Objective DB after the accuracy instruction differ significantly between males and females. This reduction in the strength of relationship between estimation error and desire following the accuracy instruction cannot be explained as an artifact of subjects basing their estimates increasingly closely on the actual probabilities, because their estimates in fact became less strongly related to the actual probabilities following the accuracy instruction; the mean coefficient of the Actual probability (β_1) dropped from 0.878 for the initial probabilities to 0.859 for the accuracy probabilities ($\underline{SD}_{difference}$ = 0.058, $\underline{t}$ (53) = 2.380, $\underline{p}$ = 0.021).

Private Self-Consciousness. Our hypothesis that Implicit Subjective DB would be a more accurate reflection of Objective DB among people high in Private Self-Consciousness was not supported. That hypothesis predicts a positive correlation between Private Self-Consciousness and the amount of reduction (following the accuracy instruction) in the absolute value of Objective DB. This correlation (with the absolute value of Objective DB again raised to the power 0.2) was .161 in the "find out today" condition ($\underline{n}$ = 54, $\underline{p}$ > .2) and -.159 in the "find out later" condition ($\underline{n}$ = 54, $\underline{p}$ > .2). The Private Self-Consciousness scale had a Cronbach's alpha of .786 in this sample.

Implicit Subjective Desirability Bias

To derive measures of Implicit Subjective DB in the "find out today" and "find out later" conditions, and across both conditions, we estimated regression equations for each subject as follows, where "Desire" is the square root of either the potential number of dollars to be gained or of twice the potential number of dollars to be lost, and "Initial vs. Accuracy" is a dummy variable coded as 1 for initial probabilities and 0 for accuracy probabilities:

$$\text{Probability judgment} = \beta_0 + \beta_1 \text{ Desire} + \beta_2 \text{ Initial vs. Accuracy} + \beta_3 \text{ Desire} \times \text{Initial vs. Accuracy} + \text{error}$$

For each subject, we estimated one regression equation like this using the 40 probability judgments in the "find out today" condition, another using the 40 probability judgments in the "find out later" condition, and a third using all 80 probability judgments (ignoring the "find out today" versus "find out later" distinction). In each case, the coefficient of the interaction term, β_3, was our measure of Implicit Subjective DB.

None of the three measures of Implicit Subjective DB differed significantly from zero. The mean in the "find out today" condition was 0.022 (SD = 0.119, t (53) = 1.354, p = .18), the mean in the "find out later" condition was -0.016 (SD = 0.201, t (53) = -0.583, p > .5), and the mean collapsing across the two conditions was 0.003 (SD = 0.1424, t (53) = 0.153, p > 0.8). The same was true when males and females were considered separately. However, these measures of Implicit Subjective DB had satisfactory split half internal consistency reliabilities of .669 for "find out today" measure, .665 for the "find out later" measure, and .787 for the measure based on all the estimates.

Explicit Subjective Desirability Bias

The mean responses to the four questions asking subjects to rate the degree of optimistic or pessimistic bias in their probability judgments did not differ significantly from the neutral response ("neither optimistically nor pessimistically"), and the neutral response was modal for all four questions. There were no significant differences in mean responses between the four questions. It therefore seemed reasonable to form a measure of Explicit Subjective DB with respect to the complete set of 80 probability judgments by summing the responses to the four questions. This measure of Explicit Subjective DB is considerably more explicit than the measures used in Studies 1 and 2, because the questions asked about optimistic or pessimistic bias, whereas the questions in the earlier studies asked about influence by optimism versus pessimism. This four-item scale had a Cronbach's alpha of .947. Its mean did not differ significantly from zero (M = -0.074, SD = 5.690, t (53) = -0.096, p = .92), or by gender. It correlated positively, but not significantly, with Objective DB in the "find out today" condition (r = .153, n = 54, p = .27), with Objective DB in the "find out later" condition (r = .225, n = 54, p = .10), and with the measure of Implicit Subjective DB based on all 80 probability judgments (r = .183, n = 54, p = .19).

Trait Subjective Desirability Bias

The 12-item Trait Subjective DB scale had a Cronbach's alpha of .732 (with the negative DB items reverse scored). The six-item positive and negative DB subscales had Cronbach's alphas of .706 and .788 respectively. We devised the Trait Subjective DB scale with the intention of measuring a unidimensional construct; we assumed that people would have either positive, neutral, or negative Trait Subjective DB. However, the correlation between the positive and negative Trait Subjective DB subscales was only -.126, and 12 of

the 54 subjects (almost a quarter) were above the median on both subscales. This suggests that positive and negative Trait Subjective DB are somewhat independent, and may be two traits rather than one.

Neither the positive nor the negative Trait Subjective DB subscale was significantly related to Implicit Subjective DB ($r = -.015$, $p > .9$, and $r = .182$, $p > .18$, respectively), to Explicit Subjective DB ($r = .102$, $p > .4$, and $r = -.089$, $p > .5$, respectively), to Objective DB in the "find out today" condition ($r = -.126$, $p > .3$, and $r = .053$, $p > .7$, respectively), or to Objective DB in the "find out later" condition ($r = .087$, $p > .5$, and $r = .154$, $p > .2$, respectively). None of these correlations differed significantly between males and females.

The Life Orientation Test had a Cronbach's alpha of .906 in this sample. The correlation between the positive Trait Subjective DB subscale and the Life Orientation Test was .486. Though quite high, this is well below either scale's reliability, and is compatible with the hypothesis that the two scales measure distinct though overlapping constructs. The correlation between the negative Trait Subjective DB subscale and the Life Orientation Test was -.661. In magnitude, this is not very far below the reliability of the negative Trait Subjective DB subscale (.788), which suggests the two scales may measure approximately the same construct.

Actively Open Minded Thinking

The 40-item Composite Actively Open Minded Thinking scale had a Cronbach's alpha of .844. Contrary to our hypothesis that more open-minded people would be more aware of and willing to admit to bias, this scale was not significantly related to the absolute value of Implicit Subjective DB (raised to the power 0.2 for a more normal distribution; r

= .094, $p > .4$), to the sum of the absolute values of the four Explicit Subjective DB items (transformed the same way; $r = -.143$, $p > .3$), or to positive Trait Subjective DB ($r = -.195$, $p > .15$). It was, however, significantly negatively correlated with negative Trait Subjective DB, $r = -.379$, $p < .005$. None of these correlations differed significantly between males and females.

Rational-Experiential Inventory

The Need for Cognition and Faith in Intuition scales of the Rational-Experiential Inventory had Cronbach's alphas of .840 and .856 respectively, and the correlation between them was .149, slightly higher than the .08 correlation reported by Epstein et al. (1996), but compatible with the idea that they are essentially independent. We performed regressions predicting the absolute value of several types of subjective DB using Faith in Intuition, Need for Cognition, and their interaction as predictor variables. Our hypothesis that people high in Faith in Intuition would deny bias predicts a negative Faith in Intuition coefficient. Our hypothesis that this would be especially true for people low in Need for Cognition predicts a positive interaction coefficient.

When the dependent variable was the absolute value of Implicit Subjective DB (raised to the power 0.2 for a more normal distribution), the overall regression model was not significant, $F(3, 50) = 0.107$, $p > .9$, $R^2 = .006$, and neither were any of the terms.

When the dependent variable was the sum of the absolute values of the four Explicit Subjective DB items (also raised to the power 0.2), the overall regression model was marginally significant, $F(3, 50) = 2.474$, $p < .08$, $R2 = .129$, as was the coefficient of Faith in Intuition, $\beta = -0.128$, $t(1) = -1.746$, $p < .09$, and the coefficient of the interaction between

Faith in Intuition and Need for Cognition, $\beta = 0.002$, $t\ (1) = 1.880$, $p < .07$. The signs of these two coefficients are in the hypothesized directions.

When the dependent variable was the positive Trait Subjective DB subscale, the overall regression model was marginally significant, $F\ (3, 50) = 2.388$, $p < .08$, $R^2 = .125$, but none of the terms approached significance.

When the dependent variable was the negative Trait Subjective DB subscale, the overall regression model was significant, $F\ (3, 50) = 3.893$, $p < .02$, $R^2 = .189$, but none of the terms approached significance.

Open Ended Questions

Subjects expressed a wide range of opinions regarding how often positive DB, negative DB, and zero DB are in their best interest. The numbers of subjects who thought positive DB was "almost always", "usually", "sometimes", "not usually", or "almost never" in their best interest were, respectively, 11, 14, 17, 4, and 8. For negative DB, the corresponding numbers were 4, 4, 21, 13, and 12, and for zero DB, the numbers were 15, 17, 16, 2, and 4. Treated as a scale with "almost never" equal to 1 and "almost always" equal to 5, the mean responses for positive, zero, and negative DB were 3.296 (more than "sometimes"), 3.685 (less than "usually"), and 2.537 (less than "sometimes"). The mean for zero DB was marginally significantly higher than the mean for positive DB ($SD_{difference} = 1.619$, $t\ (53) = 1.766$, $p < .09$), which was significantly higher than the mean for negative DB ($SD_{difference} = 2.074$, $t\ (53) = 2.691$, $p < .01$).

Subjects' thoughts about when positive and negative DB are and are not in their best interest, and why, seemed to us to identify three types of costs and benefits; affective costs

and benefits at the time of having the biased belief, behavioral costs and benefits of having the biased belief, and affective costs and benefits when the biased belief is replaced by knowledge of the truth. Many subjects said that, before the truth is known, positive DB can improve mood and reduce anxiety, perhaps helping one to get through tough times. One person wrote that "it is never better to believe that the world will not be better in the future [because] I like to have hope". Another suggested positive DB may also help improve health through the "power of the mind to heal". There was wide agreement that positive DB can increase the chances of unpleasant surprises when one's biased belief is replaced by knowledge of the truth, and that negative DB can increase the chances of pleasant surprises. For example, one person said negative DB is in his best interest when, "[I] want to convince myself I did really good", because, "you say you'll do bad when you know you won't, then when you do good you feel better about yourself." Another subject indicated that, because she knows that negative DB can reduce unpleasant surprises, it can also reduce anxiety "when I'm nervous" before knowing the truth because "if it goes badly, I expected it. if it goes well BONUS POINTS! (no sudden let down)". However, most subjects saw only affective costs of negative DB before the truth is known. Particularly when one cannot do anything to make the future better, some said, negative DB can cause one to dwell pointlessly on how bad the future might be when one would be better off enjoying the present while one can. One subject expressed this particularly emphatically: "i can't imagine why this would be a wise belief under any circumstances [because] i'd commit suicide if i went around wallowing in this type of self-pity".

Behaviorally speaking, there were differing opinions on the motivational effects of

positive and negative DB. Some people suggested positive DB can motivate one to act, creating a positive self-fulfilling prophecy. Others warned that it can lead one to set unrealistically high goals, ultimately causing failure and discouragement. Some said positive DB can lead one complacently to ignore the need to act (or the need to act more carefully) to improve the chances of a positive outcome, or to prepare to cope with a negative outcome. Some subjects mentioned that positive DB can increase, and negative DB can reduce, the tendency to overspend or take unwise financial risks. Several people saw negative DB as a helpful motivating force, especially when one has the ability to improve a bad situation but might not have sufficient motivation to do so unless one believes the situation to be worse than it really is. The idea that negative DB can help one set conservative or realistic expectations was also mentioned by a number of people. However, negative DB was seen as having several behavioral costs. Excessive pessimism may annoy other people and damage friendships and romantic relationships. It may also cause negative self-fulfilling prophecies by impairing one's motivation to improve a situation. In the workplace, particularly, one person noted that "negative thinking does not help for promotion or maintaining a job" and another said that although negative DB is in his best interest when "money is involved or when I'm dating someone [because] I'm more protected from monetary losses (1st case) rejection (2nd case)", it is not in his best interest when "I have to transmit confidence about my work [because] if I don't trust myself who would?".

As to the costs and benefits of having zero DB, many responses amounted to redescriptions of costs and benefits of positive and negative DB. However, some were more distinctly about zero DB. Some distinct affective advantages mentioned were emotional

equanimity, being able to accept and enjoy things as they are without wasting time or energy worrying or hoping or trying to change what one cannot. Some behavioral advantages involved making realistic plans and better decisions, thereby achieving a greater sense of control and fewer feelings of guilt about decisions. One person pointed to both affective and behavioral benefits when she said zero DB was in her best interest "when I predict grades" because "realistic goals that can be reached and excelled upon feel great!". Another person, who acknowledged that "sometimes if you stretch the truth to yourself it helps ease the real news when it comes!", nevertheless felt that zero DB was in her best interest when thinking about "something that involves someone else's welfare" because one "shouldn't lead people to believe things that aren't true". Another person said zero DB was in his best interest when "my work is involved" because "it's good that people think that I have my feet on the ground". Regardless of the costs and benefits of zero DB, some subjects suggested that it is often difficult to achieve, especially in the presence of uncertainty. For example, one person said that when "a situation is predictable ... it is most productive to think realistically" but that when "a situation is unpredictable ... it is most productive to have hope." Another said zero DB was in her best interest when "I have the time to analyze most of the possible events that could happen". One person said zero DB was in his best interest "nearly always but it never works out that way". Another found it "hard to think of a time when it's not" in his best interest because "reality IS whether we like it or not." To the question of when zero DB is <u>not</u> in his best interest, he responded "??? that way lies neurosis [because] what is is, what is not is not".

Discussion

The main hypothesis for Study 3, that Objective DB would be closer to zero following the accuracy instructions, was supported in both the imminent and remote outcome conditions. In other words, Subjective DB was to some degree an accurate reflection of Objective DB, and when prompted to try, subjects were able to reduce the Objective DB in their judgments to some extent. However, contrary to our expectations, people higher in Private Self-Consciousness were no more able than others to reduce the Objective DB in their judgments.

Whereas we expected subjects might show positive Objective DB for remote outcomes and negative Objective DB for imminent outcomes, on average our group of subjects showed zero Objective DB when they would learn the outcomes of their games of chance in a few weeks, but positive Objective DB when they would learn the outcomes the same day, and the difference between the two conditions was significant. A possible explanation is that the outcomes in this study were much less consequential, and more improbable, than the kinds of outcomes we derived our intuitions about the imminent-remote effect from, such as job offers or grants. People often spend considerable time thinking about their chances of receiving a job offer or a grant even when the outcome will not occur for several weeks or more, so positive DB can probably make them feel better at that point. They also know that when the outcome occurs, they will have a significant emotional response one way or the other, so when the outcome is imminent, they will have reason to fear possible disappointment. The outcomes in this study, however, were perhaps too minor and improbable to excite subjects' desires when they would not be revealed for several weeks. When they would be revealed at the end of the questionnaire, they may have induced

enough desire to cause the positive DB we found, but not enough fear of disappointment to cause the negative DB we predicted.

Whereas in Study 2 the means of both Implicit and Explicit Subjective DB relative to an accuracy ideal were significantly positive, in Study 3 the means of Implicit and Explicit Subjective DB did not differ significantly from zero. This might be due to the fact that the outcomes in Study 3 were of minor importance compared to the life event outcomes used in Study 2. In the case of Explicit Subjective DB, the mean near zero in Study 3 might also be due to the more explicit wording of the Explicit Subjective DB items in Study 3. It seems likely that many people are considerably more willing to describe their beliefs as influenced by optimism or pessimism than to describe them as optimistically or pessimistically biased.

As in Study 2, the Implicit Subjective DB coefficients, as well as the Explicit Subjective DB measure, had satisfactory split half internal consistency reliabilities, replicating the finding of individual differences in Subjective DB relative to an accuracy ideal. The correlation between Explicit and Implicit Subjective DB was positive, as in Studies 1 and 2, but was weaker and did not reach significance. Again, this may result from the presumably smaller range of desires. Explicit Subjective DB also showed positive but nonsignificant correlations with Objective DB.

The 12-item Trait Subjective DB scale, which we thought would measure a unidimensional construct, turned out to contain two dimensions. The strength of the correlation between the negative DB subscale and the Life Orientation Test raised doubts as to whether the two scales measure substantially different constructs. The positive DB subscale's correlation with the Life Orientation Test suggested that it overlaps with the Life

Orientation Test, but also measures something else--presumably admission of positive DB. Why might admission of positive DB be more distinct from optimism than negative DB? It might be that pessimistic people almost always think they suffer from negative DB, and optimistic people almost never do, whereas optimistic people, and perhaps also pessimistic people, vary in whether they think they have positive DB. One reason this might be the case is that pessimism is more commonly discussed in terms of pathology than is optimism. Depression, low self-esteem, and lack of self-confidence are extremely widespread and commonly discussed problems, whereas mania, for example, is not. People may therefore be more accustomed to attributing dispositional pessimism to a bias than to attributing dispositional optimism to a bias. Secondly, motivational biases themselves may support the same pattern; when one feels pessimistic, it is comforting to think that things are not really as bad as they seem, whereas when one feels optimistic, it is comforting to think that things are just as good as they seem.

The lack of significant correlations between the Trait Subjective DB subscales and Implicit or Explicit Subjective DB or Objective DB suggests either that the Trait Subjective DB subscales did not successfully measure a general subjective DB trait, or that subjective DB as a trait is not a major determinant of subjective DB with respect to a specific set of beliefs, which might be more influenced by unstable state factors such as mood.

The lack of relationship between the Composite Actively Open Minded Thinking scale and Implicit and Explicit Subjective DB might also be due to the latter being more of a state variable than a trait variable. However, the nonsignificant negative correlation between the Composite Actively Open Minded Thinking scale and positive Trait Subjective

DB, where we predicted a significant positive correlation, does suggest that our hypothesis that more open-minded people would be more aware of and willing to admit to bias is probably wrong. While we might be right that more open minded people are more aware of and willing to admit to their biases, they are no doubt also committed to trying to reduce their biases, and they may take this into account when deciding how much bias to admit to. Since negative Trait Subjective DB did not prove to be clearly distinct from dispositional optimism as measured by the Life Orientation Test, the significant negative correlation between the Composite Actively Open Minded Thinking scale and negative Trait Subjective DB should probably be interpreted as implying that actively open minded people tend to be dispositionally optimistic.

Our hypothesis that people high in Faith in Intuition would deny DB, especially if they were also low in Need for Cognition, received a hint of support with respect to Explicit Subjective DB, though not with respect to Implicit Subjective DB or Trait Subjective DB. This hint deserves to be followed up in future research.

The open-ended thoughts subjects provided on the costs and benefits of positive, negative, and zero DB, and their ratings of how often these are in their best interest show clearly that whether or not people can be aware of <u>having</u> nonzero DB, they are very aware of sometimes <u>wanting to have</u> nonzero DB. Our subjects willingly told us all sorts of reasons why it can sometimes be a good idea, and sometimes a bad idea, to have one's beliefs distorted by negative or positive DB. Also, many subjects shared our intuition that negative DB can be motivated by a desire to reduce the potential for disappointment.

General Discussion

The argument or assumption that people cannot disbelieve their own beliefs, that subjective bias and hence Subjective DB must be zero, appears to be incorrect. We found that many people's Subjective DB differed from zero, both when measured implicitly and when asked about explicitly. Many people are apparently quite willing, if asked, to admit that beliefs they have just expressed are biased relative to their personal ideals for thinking, relative to the ideal of accuracy, or relative to the ideal of effectiveness. The bias they admit to is, in most cases, related to their desires. While many of those with nonzero Subjective DB thought their beliefs were biased in the direction of congruence with their desires, there were also people who thought their beliefs were biased in the direction opposite to their desires. When we defined bias relative to subjects' personal ideals for thinking, in Study 1, these two groups were about equally balanced, so at the group average level of analysis, Subjective DB indeed appeared to be zero. However, when in Study 2 we defined bias relative to the ideal of accuracy, in order to avoid confounding individual differences in ideals with individual differences in estimates of bias, subjects with positive Subjective DB outweighed those with negative Subjective DB sufficiently for the mean Subjective DB to be positive. In contrast, relative to the ideal of effective goal achievement, the mean Subjective DB was strongly negative.

In principle, someone could have nonzero Implicit Subjective DB without being explicitly aware of it; that is, their estimates of bias and their desires could be positively or negatively related without their being aware of the relationship. However, we found quite strong positive correlations between Implicit Subjective DB and Explicit Subjective DB, suggesting that many or most people are aware not just that their beliefs are biased, but that

their beliefs are biased in desire-related ways.

In Study 3 we found that Implicit Subjective DB was, to some degree, an accurate reflection of Objective DB, in that when we asked subjects what they would believe if they thought in a way that would maximize accuracy, the beliefs they said they would have showed less Objective DB than their initial beliefs.

Neither Implicit nor Explicit Subjective DB with respect to specific beliefs, whether about life events or computerized games of chance, was strongly related to any of the personality measures we tried, including measures intended to measure Subjective DB as a trait. It may be related to mood or some other state variable, but further research will be needed to test this hypothesis.

If Subjective DB is related to mood, that would suggest the possibility that estimates of bias are related to desire not because people think that desire biases their beliefs, but because they think mood biases their beliefs. As we discussed above, Wegener and Petty (1995) have suggested that people have naive theories of how irrelevant contextual factors bias their judgments. Both desire and mood could be considered irrelevant contextual factors, and Implicit Subjective DB could result from subjects seeing desire, mood, or both, as sources of bias. If people see desire as a source of bias (but not mood), then the relationship between desire and estimates of bias should be independent of mood (unless people think mood moderates the biasing effects of desire). If people see mood (but not desire) as biasing their judgments via an assimilation effect, then good, neutral, and bad moods should be associated with positive, zero, and negative Implicit Subjective DB, respectively. If people see mood (but not desire) as biasing their judgments via a contrast

effect, then good, neutral, and bad moods should be associated with negative, zero, and positive Implicit Subjective DB, respectively. Distinguishing among these possibilities could be a challenging but fruitful goal for future research.

The Thinking Ideals scale in Study 2 revealed that subjects who thought good thinking produces accurate beliefs didn't think good thinking produces effective action or good feelings, and vice versa, suggesting that, like Taylor and Brown (1988), our subjects viewed positive illusions as more effective than accurate beliefs at helping them achieve their goals. On a similar note, of our 54 subjects in Study 3, about 78% felt that positive DB was in their best interest "sometimes", "usually", or "almost always". However, subjects also saw objectivity as valuable; about 88% felt that having zero DB was in their best interest "sometimes", "usually", or "almost always".

Theoretical Implications

Several existing theories can either account for our findings or suggest ways of accounting for them.

Cognitive-Experiential Self Theory

Epstein's (1994) Cognitive-Experiential Self Theory proposes that "there are two major systems by which people adapt to the world: rational and experiential" and that "people have constructs about the self and the world in both systems. Those in the rational system are referred to as beliefs and those in the experiential system as implicit beliefs or, alternatively, as schemata" (p. 715). In his Table 1 (p. 711), Epstein listed 11 differences between the two systems. Four differences are particularly relevant to our findings. The experiential system is "Affective: Pleasure-pain oriented (what feels good)" whereas the

rational system is “Logical: Reason oriented (what is sensible)”, so DB would be expected in the experiential but not in the rational system. The experiential system is “Self-evidently valid: ‘Experiencing is believing’”, whereas the rational system “Requires justification via logic and evidence”, making it possible for a belief in the experiential system to be disbelieved by the rational system. The experiential system is “Slower to change: Changes with repetitive or intense experience”, whereas the rational system “Changes more rapidly: Changes with speed of thought”, so logical contradictions between the two systems are likely to persist rather than to be automatically resolved. The experiential system is “Experienced passively and preconsciously: We are seized by our emotions”, whereas the rational system is “Experienced actively and consciously: We are in control of our thoughts”, so the rational system may be able to notice and take account of DB in experiential beliefs yet be unable to change those beliefs. Indeed, Epstein (1994) concluded from various examples of everyday thinking and behavior that “even when people know their thinking is irrational, they often find it more compelling than their rational reasoning" (p. 712). Also, as discussed in our Introduction, Epstein and his colleagues (e.g., Epstein et al., 1992) have reported some empirical evidence that bias can persist even when people acknowledge its irrationality.

A variation on Cognitive-Experiential Self Theory, suggested by David R. Williams (personal communication, March 26, 1999), is that the rational system may not always generate its beliefs independently of the experiential system, but may often derive its beliefs from the experiential system’s beliefs, adjusted in accord with metacognitive naive theories about the kinds of errors the experiential system makes. For example, rather than coming to an independent belief about the probability of getting a specific job offer, the rational

system may infer from the level of desire (and any other conditions that bias the experiential system, according to its naive theories) that the experiential system has probably misestimated the probability by a specific amount. The rational system can then derive its own belief by adjusting the experiential system's belief.

Picoeconomics

Another theory that may be able to account for our results is Ainslie's (1992, 1997) theory of Picoeconomics. Ainslie considers belief to be a behavior controlled, like other behaviors, by reinforcement, and he explains DB, or self-deception, the same way he explains addictions and other self-defeating behaviors, in terms of an ongoing battle between a person's shorter term and longer term interests. In the case of belief, the shorter term interest is in having beliefs that give immediate pleasure, and the longer term interest is in tying one's beliefs to reality[6]. If DB is like addictive behavior, then one might expect people to be aware of the conflict between the biased belief they indulge in, and the belief they think they "should" have. Just as a chocolate addict who is trying to lose weight may devour chocolate while at the same time thinking "I shouldn't be doing this!", perhaps people can believe something they want to believe at the same time as thinking "I shouldn't be believing this!".

Knowing Self Deception and Pragmatic Hypothesis Testing

Silver et al. (1989) argued, as did Mele (1997), that self-deception does not require contradictory beliefs. However, they also suggest that it is nonparadoxically possible to know that one's desires are biasing one's thinking, and yet still be self-deceived. They model the process of self-deception as an interaction between a defense attorney arguing for the

desired belief and an impartial jury trying to decide what to believe, in a court lacking both a prosecuting attorney to argue against the desired belief, and a judge to rule on the probative value of evidence. A jury in such a court would be fully aware that the defense lawyer was biased and was trying to deceive the jury, but without the help of a prosecutor and a judge they would be unable to form an unbiased verdict. A real jury, if put in such a position, might sensibly refuse to form any verdict at all until a proper trial had been conducted, but in "the court of your own mind" (p. 220) this is not an option since it would mean never forming any beliefs on issues one has desires about. In our role as internal juror, then, "we may know that we ought to be skeptical, but not how skeptical" (p. 220). Silver et al. (1989) discuss several reasons why we are usually not skeptical enough, rather than too skeptical. However, for present purposes the key feature of this model is that it allows that people can know that DB has affected their process of forming specific beliefs and yet be unable to correct the DB. This model does not in itself account for nonzero subjective DB, since it does not suggest that people can think that DB has biased a belief in a specific direction. Because the jury does its best to correct for the defense attorney's bias, even while realizing that its ability to do this is severely limited, the model seems to imply that the jury always ends up with a single belief which it considers its best estimate of the truth.

However, this model can perhaps be extended to give an account of nonzero subjective DB. Because the jury knows it should be skeptical, but is uncertain exactly how skeptical it should be, it might quite reasonably decide this on pragmatic grounds. Trope, Gervey, and Liberman's (1997) Pragmatic Hypothesis Testing model of wishful thinking suggests that wishful thinking may result from the fact that people "often see failure to detect

that a desired hypothesis is true (false rejection) as more costly than failure to detect that it is false (false acceptance)" (p. 112). Putting this probabilistically, people may often see overestimation of the probability that a desired hypothesis is true as less costly than underestimation, so on pragmatic grounds they may in effect tell their internal jury to be less rather than more skeptical about desired beliefs. For example, if Jack wants to believe he will be promoted, his internal jury might, after hearing from his internal defense attorney, decide that depending on how skeptical one chooses to be, the probability that Jack will be promoted could be estimated at anywhere from .4 to .9. If Jack feels that overestimating the probability is unlikely to do him much harm, but underestimating it could do him substantial harm, then it would seem reasonable for him to opt to be minimally skeptical and believe there is a .9 chance that he will be promoted. If he thinks overestimating the probability could be disastrous and underestimating it will do little harm, he should opt to be maximally skeptical and believe there is a .4 chance that he will be promoted. By this account, then, our subjects' initial probabilities and accuracy probabilities were both within a range their internal juries deemed reasonable, but the accuracy probabilities were closer to the "more skeptical" end of that range.

This model implies that disbelieved beliefs are only possible in the presence of ambiguity about what one should believe. It also suggests that desire itself can create (or exacerbate) ambiguity about what one should believe. However, when there is no room for ambiguity, there should be no disbelief of belief, according to this model. This idea is compatible with our findings, since our studies involved propositions with ample room for ambiguity about what one should believe.

Naive Theories and Flexible Correction of Bias

As we mentioned, Wegener and Petty (1995) claimed that people have differing naive theories of how various irrelevant contextual factors bias their judgments, and will, if they suspect such bias is occurring, attempt to correct it by altering their judgments in the direction and to the extent specified by these naive theories. Another contributor to the assimilation-contrast literature, Strack (1992), who also argued that people try to correct bias through what he called a “process representativeness check”, suggested that "such a theory-based adjustment . . . will not necessarily change an internal representation; it may exert its corrective effects solely at the response level". In support of this, he cites a finding from a diploma thesis by Almut Kübler. Subjects were asked to explain why a target person might provide nursing care for a relative. Some subjects were then warned that having generated this explanation could upwardly bias their subsequent judgment of the likelihood that the target person actually would provide nursing care. All subjects were then asked to judge the likelihood of the target person actually providing nursing care, and the target person’s emotional stability and degree of fulfillment in life. The subjects who were warned about the upwardly biasing effect of having generated the explanation gave lower estimates of the probability, apparently attempting to correct for the bias they had been warned of. However, the warning had no effect on the related judgments of emotional stability and fulfillment. Strack (1992) takes this to mean that in response to the bias warning, people did not correct the representation they had formed of the target person, but merely adjusted their response to the probability question. As a result, the biasing effects on the representation still affected judgments of emotional stability and fulfillment.

Conclusion

Conscious disbelief of one's beliefs is apparently neither impossible nor uncommon. It is often related to desire, sometimes positively and sometimes negatively, and it seems to reflect some degree of accurate awareness of biasing effects of desire. It may be possible because the supposedly unitary person that says "I believe..." actually contains more than one entity capable of having beliefs. For example, people may contain Epstein's (1994) rational and experiential systems, Ainslie's (1992) long and short range interests, or Strack's (1992) response level and internal representation. Another reason disbelief of one's beliefs may be possible is suggested by Silver et al.'s (1989) "internal jury" model, combined with Trope et al.'s (1997) pragmatic hypothesis testing model. Beliefs may be selected on pragmatic grounds from a set of possible beliefs that an "internal jury" considers equally plausible. Thus, a persons' belief in one pragmatic situation may contradict the same person's belief in another pragmatic situation. According to this model, then, what lies behind the surface manifestation of disbelieved beliefs is the internal jury's uncertainty about what to believe. Indeed, disbelieved beliefs may well only be possible in the presence of ambiguity about what one should believe.

How does our finding that people can disbelieve their beliefs affect the utility of the concept of belief? If someone's statement that they believe p does not allow one to conclude that they do not believe not-p, does the concept of belief lose its meaning? We think not, for two reasons. First, our findings do not imply that people can simultaneously believe any arbitrary pair of mutually contradictory beliefs; they merely imply that people can simultaneously believe some pairs of mutually contradictory beliefs. The greater the

discrepancy between two beliefs, and the less ambiguity there is about what should be believed, the less plausible it is that someone would simultaneously believe both of them. For example, someone's statement that they believe there is a 60% probability that the stock they have just bought will double in value within a year allows one to infer that it is highly unlikely that they also believe the probability is 1%. Second, our findings do not imply that it is impossible to find out which propositions a person believes and which they do not; they only imply that it may take more than a single question to do the job. For example, asking both what someone believes and what they would believe if they thought in an accuracy-maximizing way, as we did, may be sufficient. For example, if someone believes the probability they will get divorced is 5% and thinks they should believe it is 10%, then our findings provide no reason not to conclude confidently that they do not believe it is less than 5% or more than 10%. Therefore, the concept of belief retains utility and meaning, though admittedly it does lose a degree of simplicity and perhaps elegance.

A practical conclusion can be drawn from this research. Public health workers and others who have tried to get people to eat healthily, stop smoking, exercise regularly, and so on, may have a lot to teach those of us interested in improving the quality of people's thinking. If people can knowingly have biased beliefs just as they can watch T.V. while thinking "I really should be exercising", then it will not be enough merely to teach people what good thinking is and how to recognize whether they are doing it. It will also be necessary both to persuade them that good thinking is worthwhile, and to help them develop effective strategies for resisting the temptation to think poorly.

References

Ainslie, G. (1992). Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person. Cambridge, UK: Cambridge University Press.

Ainslie, G. (1997). If belief is a behavior, what controls it? Behavioral and Brain Sciences, 20(1), 103-104.

Babad, E. (1995). Can accurate knowledge reduce wishful thinking in voters' predictions of election outcomes? Journal of Psychology, 129(3), 285-300.

Babad, E., Hills, M., & O'Driscoll, M. (1992). Factors influencing wishful thinking and predictions of election outcomes. Basic and Applied Social Psychology, 13(4), 461-476.

Babad, E., & Katz, Y. (1991). Wishful thinking--against all odds. Journal of Applied Social Psychology, 21(23), 1921-1938.

Bar-Hillel, M., & Budescu, D. V. (1995). The elusive wishful thinking effect. Thinking and Reasoning, 1(1), 71-103.

Baron, J. (1991). Beliefs about Thinking. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.) Informal Reasoning and Education (pp. 169-186). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron, J. (1994). Thinking and Deciding (2nd ed.). New York, NY: Cambridge University Press.

Baron, J. (1995). Myside bias in thinking about abortion. Thinking and Reasoning, 1(3), 221-235.

Budescu, D. V., & Bruderman, M. (1995). The relationship between the illusion of control and the desirability bias. Journal of Behavioral Decision Making, 8(2), 109-125.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42(1), 116-131.

Cederblom, J. (1989). Willingness to reason and the identification of the self. In E. Maimon, D. Nodine, & O'Conner (Eds.) Thinking, reasoning, and writing (pp. 147-159). New York, NY: Longman.

Cohen, B. L., & Wallsten, T. S. (1992). The effect of constant outcome value on judgments and decision making given linguistic probabilities. Journal of Behavioral Decision Making, 5(1), 53-72.

Costa, P. T., & McCrae, R. R. (1992). Revised NEO personality inventory. Odessa. FL: Psychological Assessment Resources.

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.

Dawes, R. M., Singer, D., & Lemons, F. (1972). An experimental analysis of the contrast effect and its implications for intergroup communication and the indirect assessment of attitude. Journal of Personality and Social Psychology, 21(3), 281-295.

Denes-Raj, V., Epstein, S., & Cole, J. (1995). The generality of the ratio-bias phenomenon. Personality and Social Psychology Bulletin, 21(10), 1083-1092.

Epstein, S. (1990). Cognitive-experiential self theory. In L. Pervin (Ed.) Handbook of personality theory and research. New York: Guilford Press.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. American Psychologist, 49(8), 709-724.

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. Journal of Personality and Social Psychology, 62(2), 328-339.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. Journal of Personality and Social Psychology, 71(2), 390-405.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. Journal of Consulting and Clinical Psychology, 43(4), 522-527.

Fischer, I., & Budescu, D. V. (1995). Desirability and hindsight biases in predicting results of a multi-party election. In J.-P. Caverni, M. Bar-Hillel, F. H. Barron, & H. Jungermann (Eds.) Contributions to Decision Making - I (pp. 193-211). Amsterdam: Elsevier.

Granberg, D., & Brent, E. (1983). When prophecy bends: The preference-expectation link in U.S. presidential elections, 1952-1980. Journal of Personality and Social Psychology, 45(3), 477-491.

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. Journal of Personality and Social Psychology, 37(2), 147-169.

Hsee, C. K. (1996). Elastic justification: How unjustifiable factors influence judgments. Organizational Behavior and Human Decision Processes, 66(1), 122-129.

Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. Journal of Personality, 21, 329-335.

Irwin, F. W., & Graae, C. N. (1968). Tests of the discontinuity hypothesis of the effects of independent outcome values upon bets. Journal of Experimental Psychology, 76(3), 444-449.

Irwin, F. W., & Metzger, M. J. (1966). Effects of probabilistic independent outcomes upon predictions. Psychonomic Science, 5(2), 79-80.

Irwin, F. W., & Metzger, M. J. (1967). Effects of independent outcome-values of past events upon subsequent choices. Psychonomic Science, 9(12), 613-614.

Irwin, F. W., & Snodgrass, J. G. (1966). Effects of independent and dependent outcome values upon bets. Journal of Experimental Psychology, 71(2), 282-285.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263-291.

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(4), 534-544.

Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. Journal of Experimental Social Psychology, 28(2), 145-168.

Krueger, J. (1998). Enhancement bias in descriptions of self and others. Personality and Social Psychology Bulletin, 24(5), 505-516.

Krueger, J., Ham, J. J., & Linford, K. M. (1996). Perceptions of behavioral consistency: Are people aware of the actor-observer effect? Psychological Science, 7(5), 259-264.

Krueger, J., & Zeiger, J. S. (1993). Social categorization and the truly false consensus

effect. Journal of Personality and Social Psychology, 65(4), 670-680.

Kruglanski, A. W. (1996). Motivated social cognition: Principles of the interface. In E. T. Higgins, & A. W. Kruglanski (Eds.) Social psychology: A handbook of basic principles (pp. 493-520). New York: Guilford Press.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480-498.

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

Mele, A. R. (1994). Self-control and belief. Philosophical Psychology, 7(4), 419-435.

Mele, A. R. (1997). Real self-deception. Behavioral and Brain Sciences, 20(1), 91-136.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? Psychological Bulletin, 82(2), 213-225.

Olsen, R. A. (1997). Desirability bias among professional investment managers: Some evidence from experts. Journal of Behavioral Decision Making, 10(1), 65-72.

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. Advances in Experimental Social Psychology, 20, 297-340.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. Journal of Personality and Social Psychology, 46(2), 237-248.

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of

psychological well-being. Journal of Personality and Social Psychology, 57(6), 1069-1081.

Sá, W. C., West, R. F., & Stanovich, K. E. (1998). The domain specificity and generality of belief bias in reasoning and judgment. Manuscript submitted for publication, Ontario Institute for Studies in Education, University of Toronto.

Sartre, J. P. (1958). Being and nothingness: An essay on phenomenological ontology. London: Methuen and Co.

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. Health Psychology, 4(3), 219-247.

Sieber, J. E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. Journal of Personality and Social Psychology, 30(5), 688-694.

Silver, M., Sabini, J., & Miceli, M. (1989). On knowing self-deception. Journal for the Theory of Social Behaviour, 19(2), 213-227.

Slovic, P. (1966). Value as a determiner of subjective probability. IEEE transactions on human factors in electronics, HFE-7, 22-28.

Snyder, M. L., Stephan, W. G., & Rosenfield, D. (1976). Egotism and attribution. Journal of Personality and Social Psychology, 33(4), 435-441.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. Journal of Educational Psychology, 89(2), 342-357.

Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In L. L. Martin, & A. Tesser (Eds.) The construction of social judgments (pp. 249-275). Hillsdale, NJ: Erlbaum.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin, 103(2), 193-210.

Trope, Y., Gervey, B., & Liberman, N. (1997). Wishful thinking from a pragmatic hypothesis-testing perspective. In M. S. Myslobodsky (Ed.) The mythomanias: The nature of deception and self-deception (pp. 105-131). Mahway, NJ: Lawrence Erlbaum Associates.

Vivian, J. E., & Berkowitz, N. H. (1992). Anticipated bias from an outgroup: An attributional analysis. European Journal of Social Psychology, 22(4), 415-424.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. Journal of Personality and Social Psychology, 68(1), 36-51.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. Journal of Personality and Social Psychology, 39(5), 806-820.

Wright, G., & Ayton, P. (1989). Judgmental probability forecasts for personal and impersonal events. International Journal of Forecasting, 5, 117-125.

Zakay, D. (1983). The relationship between the probability assessor and the outcomes of an event as a determiner of subjective probability. Acta Psychologica, 53(3), 271-280.

Appendix A: Study 1 Questionnaire, Part One

Beliefs (bel4) (Allow plenty of time for this one.)

Part I

Please use the answer sheet. Do not write on this form.

Answer question A about all 16 statements before going on to question B.

A. How likely is each statement to be true? Answer in percent.

Statements

1. I will own my own home.
2. My first job after I complete my education will have a starting salary of at least $30,000.
3. My first job after I complete my education will have a starting salary of at least $50,000.
4. I will receive an acceptable job offer between now and the time I receive my degree.
5. My grade point average will be higher the next time I get my grades.
6. I will make an investment that will double in real value in five years.
7. My work will be recognized with an award.
8. I will live past 80.
9. I will get divorced (in the future).
10. I will have a heart attack before age 50.
11. I will be fired from a job (in the future).
12. I will be sterile.

13. I will break a bone in the next year.

14. I will be diagnosed as having a sexually transmitted disease.

15. I will be treated for alcoholism.

16. I will spend 6 months unemployed and looking for a job, some time after I graduate.

(The following was on the next page.)

Now answer questions B and C about all 16 statements before going on.

B. Consider the reasons you have for your answer to question A -- all the reasons, including feelings, hunches, and other kinds of evidence. Briefly (in a couple of words at most) list the main reasons for each belief you've written down.

C. This is the hardest question. Please think about how you would want yourself and others to think ideally. How should people take into account the evidence they have? Their perceptions? Their feelings and hunches? Their emotions? You do not need to write down what you think about this.

Now re-answer question A, but instead of giving the probability of each statement, say what probability you would assign if you thought in this ideal way. Or, putting this another way, imagine someone else who thought in this ideal way but who had exactly the same reasons, including evidence, hunches, etc., that you have, with the same strength. What probability would this person assign to each statement?

(The following was on the next page.)

Now answer question D about all 16 statements before going on.

D. Look at your answers to questions A and C. If the answers are the same, leave this

blank. If the answers are different, please indicate which of the following are true. (You can indicate more than one.)

1. There was no particular reason why my answers were different. The difference between the two answers could just as easily have been in the opposite direction.
2. My answer to question A was influenced more by evidence.
3. My answer to question C was influenced more by evidence.
4. My answer to question A was influenced more by intuitions or hunches.
5. My answer to question C was influenced more by intuitions or hunches.
6. My answer to question A was influenced more by optimism, by what I wished were true.
7. My answer to question A was influenced more by pessimism.
8. I don't really know much about this.
9. My answer to question A reflects my personal experience, which may not be the same as other people's experience.
10. My answer to question A reflects a personal bias I admit I have.
11. I wasn't paying attention.
12. Some other reason. (Explain.)

(The following was on the next page.)

Now answer questions E and F about all 16 statements.

E. Think about how you would feel if you knew that each statement were true. Rate your feeling on a scale from -5 to 5, in which -5 means ``as unhappy as anything could make me," 5 means ``as happy as anything could make me," and 0 means ``neither happy nor

unhappy."

F. Answer question A again. Feel free to look at your answers to previous questions, such as A and C.

Appendix B: Study 1 Questionnaire, Part Two (Desirability Bias Self-Report Scale)

Part II

Now indicate whether you agree or disagree with each of the following statements, using the following scale:

AA - strongly agree

A - agree, but not strongly

N - neither agree nor disagree

D - disagree, but not strongly

DD - strongly disagree

1. People's confidence in their beliefs should never be influenced by what they want to be true or false.
2. People's confidence in their beliefs should (at least) sometimes be influenced by what they want to be true or false.
3. Confidence in beliefs should never be influenced by hunches.
4. Confidence in beliefs should (at least) sometimes be influenced by hunches.
5. Confidence in beliefs should depend only on the kind of evidence that most people would consider relevant.
6. Confidence in beliefs should (at least) sometimes depend on things that most people would not consider relevant.
7. People should be willing to face the truth even if it hurts.
8. People should believe what makes them feel good, even if it is false.
9. My own confidence in my beliefs is affected by what I want to be true or false.

10. My own confidence in my beliefs is not affected by what I want to be true or false.
11. My confidence in my beliefs is influenced by hunches.
12. My confidence in my beliefs is not influenced by hunches.
13. My confidence in my beliefs depends only on the kind of evidence most people would consider relevant.
14. My confidence in my beliefs depends on the kind of evidence I consider relevant but most people would not consider relevant.
15. I am willing to face the truth, even if it hurts.
16. I believe what makes me feel good, even if it is false.

Thank you.

Appendix C: Second and Third Questions From Study 2 Questionnaire.

Question B (or C, depending on counterbalancing)

Please take a few moments to consider the following question:

How should people think in order to arrive at beliefs that are as accurate (or true) as possible? How should they take into account the evidence they have? Their perceptions? Their intuitions and hunches? Their desires and emotions? Then please answer the following question:

How would you have answered Question A ("How likely is each statement to be true?") if you thought in this way that would lead you to form beliefs that are as accurate (or true) as possible? Or, putting this another way, imagine someone else who thought in this way but who had exactly the same reasons, including evidence, perceptions, intuitions, hunches, desires, and emotions, that you have, with the same strength. How would this person have answered Question A?

Question C (or B, depending on counterbalancing)

First, please think about how people's beliefs may affect their achievement of their goals--all of their goals, including practical (or objective) goals and emotional (or subjective) goals.

Then take a few moments to consider the following question:

How should people think in order to arrive at beliefs that maximize their achievement of their goals? How should they take into account the evidence they have? Their perceptions? Their intuitions and hunches? Their desires and emotions?

Then please answer the following question:

How would you have answered Question A ("How likely is each statement to be true?") if you thought in this way that would lead you to form beliefs that would maximize your achievement of your goals? Or, putting this another way, imagine someone else who thought in this way but who had exactly the same reasons, including evidence, perceptions, feelings, hunches, desires, and emotions, that you have, with the same strength. How would this person have answered Question A?

Appendix D: Thinking Ideals Scale From Study 2 Questionnaire.

The following items are numbered in the randomized order in which they appeared in the questionnaire, but for clarity they are listed here grouped by subscale and item pair.

18. People should believe whatever makes them most effective, even if it's not true. (Effectiveness 1a)

15. People should <u>not</u> believe whatever makes them most effective, unless it's also true. (Effectiveness 1b)

14. I try to have beliefs that help me to be effective, even if they're not true. (Effectiveness 2a)

28. I do not try to have beliefs that help me to be effective, unless they're also true. (Effectiveness 2b)

4. If believing something will help people achieve their goals, this means they should believe it. (Effectiveness 3a)

45. If believing something will help people achieve their goals, this does <u>not</u> necessarily mean they should believe it. (Effectiveness 3b)

23. I believe whatever will help me achieve my goals, even if the evidence points the other way. (Effectiveness 4a)

41. I do not believe whatever will help me achieve my goals, unless the evidence also supports it. (Effectiveness 4b)

30. Effective action is probably the most important goal of thinking. (Effectiveness 5a)

46. Effective action is one of the less important goals of thinking. (Effectiveness 5b)

5. The best type of thinking is that which most often helps you achieve your practical goals. (Effectiveness 6a)

36. The type of thinking that most often helps you achieve your practical goals is actually not the best type of thinking. (Effectiveness 6b)

24. I particularly admire thinkers who make things happen, even if they sometimes believe things that are probably not true. (Effectiveness 7a)

6. I do not admire thinkers who make things happen if they also believe a lot of things that are probably not true. (Effectiveness 7b)

33. Children should be taught first and foremost to think in ways that help them achieve practical goals. (Effectiveness 8a)

22. In teaching children how to think, helping them to achieve practical goals should not be the top priority. (Effectiveness 8b)

35. People's beliefs should be influenced by what they want to be true or false. (Feelings 1a)

7. People's beliefs should not be influenced by what they want to be true or false. (Feelings 1b)

3. I aim to let my beliefs be affected by what I want to be true or false. (Feelings 2a)

20. I aim to prevent my beliefs from being affected by what I want to be true or false. (Feelings 2b)

42. If believing something makes people feel good, this means they should believe it. (Feelings 3a)

47. If believing something makes people feel good, this does not necessarily mean

they should believe it. (Feelings 3b)

19. I believe whatever makes me feel good, even if the evidence points the other way. (Feelings 4a)

37. I do not believe whatever makes me feel good, unless the evidence also supports it. (Feelings 4b)

40. Emotional well-being is probably the most important goal of thinking. (Feelings 5a)

16. Emotional well-being is one of the less important goals of thinking. (Feelings 5b)

17. The best type of thinking is that which most often helps you to feel good. (Feelings 6a)

43. The type of thinking that most often helps you to feel good is actually <u>not</u> the best type of thinking. (Feelings 6b)

38. I particularly admire thinkers who take a positive view of things, even if they sometimes believe things that are probably not true. (Feelings 7a)

32. I do not admire thinkers who take a positive view of things if they also believe a lot of things that are probably not true. (Feelings 7b)

2. Children should be taught first and foremost to think in ways that help them look on the bright side of life. (Feelings 8a)

39. In teaching children how to think, helping them to look on the bright side of life should <u>not</u> be the top priority. (Feelings 8b)

27. People should always be willing to face the truth. (Truth 1a)

9. People should not always feel obliged to face the truth. (Truth 1b)

25. I always try to face the truth. (Truth 2a)

29. Sometimes I do not try to face the truth. (Truth 2b)

48. If the evidence indicates that something is true, this means people should believe it. (Truth 3a)

12. If the evidence indicates that something is true, this does not necessarily mean people should believe it. (Truth 3b)

1. I always believe whatever the evidence shows is most likely to be true. (Truth 4a)

13. I do not always believe whatever the evidence shows is most likely to be true. (Truth 4b)

21. Truth is probably the most important goal of thinking. (Truth 5a)

31. Truth is one of the less important goals of thinking. (Truth 5b)

26. The best type of thinking is that which helps you arrive at the most accurate beliefs. (Truth 6a)

8. The type of thinking that helps you arrive at the most accurate beliefs is actually not the best type of thinking. (Truth 6b)

11. I particularly admire thinkers who seek the truth, whatever it might be. (Truth 7a)

44. I do not particularly admire thinkers who seek the truth, whatever it might be, without regard for the possible consequences of knowing it. (Truth 7b)

34. Children should be taught first and foremost to think in ways that help them discover the truth, whatever the truth may be. (Truth 8a)

10. In teaching children how to think, helping them to discover the truth, whatever it may be, should not be the top priority. (Truth 8b)

Appendix E: First Part of Study 3 Questionnaire.

Games of Chance Study

This study has been tested and found to work fine with Netscape Navigator 3 and 4 on a Pentium 133Mhz PC with 48 MB of RAM. We recommend using one of these browsers. With Microsoft Internet Explorer 4, it was too slow to be usable on the above PC, and one participant has reported using MS IE4 on a 266Mhz PC and encountering a problem mid-way through which prevented her from being able to finish the study and submit her responses. (Click here to download a Netscape browser. If you are using a browser that should work and encounter problems, check that you have both JavaScript and Java enabled in your browser preferences.)

Time Involved

Completing this study should take less than an hour.

Payment for Participating in this Study

You will start with a balance of $4. As part of the study, you will play 40 games of chance, 20 of which involve a chance of winning money, and 20 of which involve a chance of losing money. You can only lose at most $3 of your initial $4, but you can potentially lose all of anything you win. For example, if you win $50 on one game and lose $60 on another, that's a net loss of $10. But since you can only lose at most $3 of your $4, you would end up with $1. The expected average amount participants in this study will be paid is between $6 and $7. The most anyone could possibly get paid is $204. You will get paid at least $1. To participate you must accept that you may end up being paid only $1.

I have read and understood the above and I wish to participate. [Subject must click

on that sentence to proceed.]

(If you would like to ask any questions about this study before deciding whether to participate, please email Michael Siepmann at siepmann@psych.upenn.edu.)

Games of Chance Study

Please resize this window to make it fill all or almost all of your screen.

Before starting, please note:

Please be careful not to resize or minimize this window, or change your screen resolution after starting the study. Depending on your browser, this might cause your answers to be lost.

If you get disconnected from the internet while doing the study, then if possible, just reconnect without minimizing this window. (If your internet service disconnects you after some period such as an hour, then consider reconnecting now to ensure you won't get disconnected.)

It would be a good idea to close any unnecessary programs you may have running. (This will free up your computer's memory and minimize the chances of problems due to conflicts between your browser and other programs.)

A bug in Netscape 3 may cause the cursor (a vertical line like this: |) to appear in a box labeled 'here' rather than in the box you type your answers into. Do not worry about this. Just type your answer and it will appear in the box it is supposed to, despite the | being in the wrong place.

Minimize glare on your screen--for example close blinds or curtains on windows. An important part of what you will be looking at will be small white dots on a black

background, and reflections or glare could make it harder to see them.

Please avoid distractions or interruptions while doing this study. For example, please do not have music, radio, or TV turned on while doing it, please avoid talking with others while doing it, and please do not have an email program running that might notify you of new mail, etc. Thank you.

When this window fills all or almost all of your screen, please click here to load the questionnaire.

The black rectangle below can be used to play games of chance. As you can see, white dots are constantly appearing and disappearing at random positions all over the rectangle. To play a game of chance, the computer looks in a randomly chosen position in the rectangle to see if there is a white dot there at that moment. (Please type 'ok' and press TAB...)

If there is a white dot there, the outcome of the game is a HIT. If there is not a white dot there, the outcome is a MISS. The chance of getting a HIT can be varied by changing the number of white dots. The more white dots the rectangle is showing, the greater the chance of a HIT. ('ok' or 'back' and TAB...)

Later on, you will be looking at the rectangle when it is showing various different numbers of white dots. You will be estimating the chance of a HIT. First, however, I will show you what the rectangle looks like when the chance of a HIT is set at several different values. Please observe carefully. ('show me' or 'back'...)

Demonstration 1 of 10. The chance of a HIT right now is: 89 in 100,000. ('ok'...)

Now I will ask you to estimate the chance of a HIT. During this practice phase, I will

give you feedback on your estimates. ('ask me'...)

Practice 1 of 20. What do you estimate is the chance of a HIT, out of 100,000? (Type your answer then press TAB...)

Feedback for Practice 1 of 20. True chance: 198/100,000. Your estimate: 100/100,000. (49.5% too low.) ('ok' and TAB...)

This study includes two sets, named A and B, of 1 games of chance. In some of the games in each set, if you get a HIT you will win some money. In the other games in each set, if you get a HIT you will lose some money. In all games, a MISS neither wins nor loses you anything. ('more'...)

In each set of games, $100 is the most you can actually be paid. That is, if your Set A wins minus losses total $100 or more, we will pay you $100. The same goes for Set B. So in this study you could win up to $200 in addition to the $4 you start out with. ('more' or 'back'...)

In total, across both sets of games, you cannot lose more than $3 of the $4 you start out with. So, losses can reduce your $4 down to $1. And losses can wipe out wins in the same set. For example, if in Set A you won $50 on one game, but lost $60 on another Set A game, the end result of Set A would be that you would lose $3 of your initial $4. ('more' or 'back'...)

However, losses in one set cannot wipe out wins in the other set. For example, if you won $50 in Set A and lost $60 in Set B, you would receive your $50 from Set A, and your loss from Set B would be deducted from your initial $4, reducing it to the minimum of $1. So, the total you would receive would be $51. ('more' or 'back'...)

Next you will be asked to estimate the chance of getting a HIT and winning or losing money on each game. Later, when you play the games, the computer will record but not reveal to you whether you got a HIT on each game. You will find out whether you got a HIT and won or lost money on each Set A game several weeks from now, by email. 'start' or 'back'...

"Which game is this?" A-1. "What if I get a HIT?" Then you lose $32. "When will I find out?" Several weeks from now, by email. How likely do you think it is that you will lose $32 on this game? (Type "lose32=[your answer]".)

You have finished estimating the chances for the Set A games. The Set B games are similar to the Set A games, except that you will find out whether you got a HIT and won or lost money on each Set B game at the end of this questionnaire. 'start'...

"Which game is this?" B-1. "What if I get a HIT?" Then you lose $13. "When will I find out?" At the end of this questionnaire. How likely do you think it is that you will lose $13 on this game? (Type "lose13=[your answer]".)

You have finished estimating the chances for the Set B games. Now please take a few moments to consider this question: How would you think and perceive if both (a) your only goal was to arrive at the most accurate (or true) beliefs possible, and (b) you had the ability to eliminate all influences on your thinking and perception that conflicted with this goal of accuracy? 'have done so'...

You will now see the Set A and B games a second time. This time, please say what estimate you think you would have given if you had thought and perceived in the way you considered a moment ago--that is, in a way that would lead to the most accurate (or true)

beliefs possible. 'ok'...

First, please answer this for the Set A games--the ones for which you will find out whether you got a HIT and won or lost money on them several weeks from now, by email. 'start'...

"Which game is this?" A-1 (SECOND LOOK). "What if I get a HIT?" Then you lose $32. "When will I find out?" Several weeks from now, by email. If you thought in a way that would lead to the most accurate beliefs possible, how likely would you have thought it is that you will lose $32 on this game? You said 50 before. (Type "lose32=[your answer]".)

Now please answer the same question for the Set B games--the ones for which you will find out whether you got a HIT and won or lost money on them at the end of this questionnaire. 'start'...

"Which game is this?" B-1 (SECOND LOOK). "What if I get a HIT?" Then you lose $13. "When will I find out?" At the end of this questionnaire. If you thought in a way that would lead to the most accurate beliefs possible, how likely would you have thought it is that you will lose $13 on this game? You said 50 before. (Type "lose13=[your answer]".)

Thank you. Now it is time to play the games. This will end Part I of this questionnaire. Part II will follow after you play the last game. ('play'...)

About to play Game A-1. If you get a HIT on this game, you will lose $32. The computer will record the result now, but you will find out by email several weeks from now whether you got a HIT and lost $32 on this game. Press TAB to play this game.

Appendix F: Study 3 Trait Subjective DB Scale.

1. Bad news seems to have more impact on my beliefs about the world than reason says it should. (Negative DB / Bad / World)
2. I'm prone to believing that fewer good things will happen than, in fact, probably will. (Negative DB / Good / Future)
3. I anticipate bad outcomes more often than past experience suggests I should. (Negative DB / Bad / Future)
4. If I encounter evidence that I have undesirable qualities, I tend to ignore it. (Positive DB / Bad / Self)
5. I believe unpleasant things about myself despite an absence of evidence that they're true. (Negative DB / Bad / Self)
6. I tend to believe that the world is a happier and better place than it probably really is. (Positive DB / Good / World)
7. I believe more positive things about myself than are probably really true. (Positive DB / Good / Self)
8. My beliefs about the world ignore many good things that are in fact true about it. (Negative DB / Good / World)
9. My image of the world does not fully acknowledge many bad things I know are in fact true of it. (Positive DB / Bad / World)
10. Feedback telling me I have good qualities seems to have no effect on my beliefs about myself. (Negative DB / Good / Self)
11. I am more optimistic than is justified by the evidence. (Positive DB / Good /

Future)

12. Even when an unpleasant event is really quite likely, I tend to expect that it won't actually happen. (Positive DB / Bad / Future)

Footnotes

[1]This is mathematically equivalent to asking subjects to estimate the probability of each event occurring. For example, to predict that an event will not happen, and to estimate the chance of this prediction proving accurate as 80%, is equivalent to estimating the probability of the event occurring as 20%.

[2]The analysis they could have done using the outcome information would be to compute within-subject correlations (or regression coefficients), with the seven parties as items, between level of identification and predicted minus actual number of seats, and test whether this correlation (or regression coefficient) was significantly positive.

[3]Whether this should really be called a bias is open to question. Krueger (1998) asked subjects, for each trait, "'How desirable or undesirable do you feel it is for people to be or act this way?'" (p. 508), which seems to be a question about their personal preferences, not about something objective. Another way of describing what he calls self-enhancement bias, then, is that people prefer people similar to themselves over people different from themselves.

[4]More specifically, in the within-subject regression equation described in the Objective DB section of the Results section, the subjects who were dropped had coefficients of the Actual probabiliy (β_1) of less than 0.5, and were outside values on a stem and leaf plot of the distribution of this coefficient.

[5]We transformed the actual and judged probabilities using the inverse of the exponential function we used to generate the actual probabilities, described in the Method

section.

[6]Ainslie's (1992) view on why it is in our longer term interest to tie our beliefs to reality is fascinating and highly counterintuitive. In very rough outline, he thinks we value reality mainly for its unpredictability. He argues that, for reasons he goes into, we depend on unpredictable stimuli as part of a self-control strategy to counteract our myopic tendency to choose smaller but earlier rewards over larger but later rewards. Because the brain can self-administer pleasure, that myopic tendency would, if not controlled, lead us to exhaust our capacity to experience pleasure, in much the same way that, for example, overfishing can exhaust the capacity of the oceans to supply fish.

Table 1

Measurement of objective, implicit subjective, and explicit subjective desirability bias

	Type of desirability bias (DB):		
	Objective	Implicit Subjective	Explicit Subjective
Actual belief:	Ask subject	Ask subject	-
Normative belief:	Know	Ask subject	-
Desire:	Ask subject	Ask subject	-
Desirability bias*:	Calculate	Calculate	Ask subject

* i.e., relationship between bias (i.e., actual belief minus normative belief) and desire.

Table 2

Intercorrelations and reliabilities of Study 1 measures

Measure	1	2	3	4	5	6
1: Implicit Subjective DB (initial vs. ideal)	.838					
2: Implicit Subjective DB (final vs. ideal)	.972**	.804				
3: Explicit Subjective DB	.526**	.454**	.753			
4: DB self-report scale, full	-.238	-.215	-.134	.724[a]		
5: DB self-report subscale about how subject thinks	-.299*	-.290*	-.110	.937**	.563[b]	
6: DB self-report subscale about how people should think	-.126	-.092	-.141	.911**[c]	.709**[d]	.419[e]

Note. n = 55. No significant differences between males (n = 19) and females (n = 36) except where noted. Figures on the main diagonal are internal consistency reliabilities. (For Implicit and Explicit Subjective DB, these are split half correlations; they were similar for male and female subjects, all differing by less than .02. For the DB self-report scale and subscales, these are Cronbach's alphas calculated from eight and four variables respectively.) Correlations of magnitude .266 or greater are significant at the .05 level and correlations of magnitude .345 or greater are significant at the .01 level.

[a]Female alpha = .737, male alpha = .644.

[b]Female alpha = .425, male alpha = .678.

[c]Female $\underline{r}$ = .960, male $\underline{r}$ = .771, $\underline{Z}$ = 3.030, $\underline{p}$ < .05.

[d]Female $\underline{r}$ = .844, male $\underline{r}$ = .440, $\underline{Z}$ = 2.504, $\underline{p}$ < .05.

[e]Female alpha = .495, male alpha = .136.

* $\underline{p}$ < .05

** $\underline{p}$ < .01

Table 3

Intercorrelations and reliabilities of Study 2 measures

	1	2	3	4	5	6	7
1: Implicit Subjective DB relative to Accuracy Ideal	.840[a]						
2: Explicit Subjective DB relative to Accuracy Ideal	.537[b]*	.357[b]					
3: Implicit Subjective DB relative to Effectiveness Ideal	.336[a]*	.093[b]	.719[a]				
4: Explicit Subjective DB relative to Effectiveness Ideal	.109[b]	.364[b]*	.540[b]*	.710[b]			
5: Thinking Ideals	.124[b]	.169[c]	-.005[b]	.162[c]	.878[b]		
6: Psychological Well-Being	.061[d]	.081[c]	-.095[d]	-.006[c]	.105[c]	.931[d]	
7: Private Self-Consciousness	.020[a]	-.002[b]	-.016[a]	.020[b]	.128[b]	.052[d]	.687[a]

Note: On main diagonal, rows 1 and 3 are split half correlations, and rows 2, 4, 5, 6, and 7 are Cronbach's alphas on 2, 2, 24, 54, and 10 variables, respectively.

* $p < .001$ (otherwise, $p > .05$).

[a] $n = 99$

[b] $n = 96$

[c] $n = 94$

[d] $n = 97$

Table 4

Intercorrelations and reliabilities of Study 2 Thinking Ideals subscales

	1	2	3
1: Truth subscale	.704[a]		
2: Effectiveness subscale	-.265[a]*	.843[b]	
3: Feeling good subscale	-.391[c]**	.691[d]***	.799[d]

Note: Cronbach's alphas on eight variables are on main diagonal. Similar patterns of correlations were found when male and female subjects were considered separately.

* $p < .01$

** $p < .001$

*** $p < .0001$

[a] $n = 98$

[b] $n = 99$

[c] $n = 96$

[d] $n = 97$