Designing Information for Remediating Cognitive Biases in Decision-Making

Yunfeng Zhang
Computer and Information Science
University of Oregon, Eugene, OR 97403
zywind@cs.uoregon.edu

Rachel K. E. Bellamy, Wendy A. Kellogg
IBM T. J. Watson Research Center
PO Box 218, Yorktown Heights, NY 10598
{rachel, wkellogg}@us.ibm.com

ABSTRACT
Software is playing an increasingly important role in supporting human decision-making. Previous HCI research on decision support systems (DSS) has improved the information visualization aspect of DSS information design, but has somewhat overlooked the cognitive aspect of decision-making, namely that human reasoning is heuristic and reflects systematic errors or cognitive biases. We report on an empirical study of two cognitive biases: conservatism and loss aversion. Two remediation techniques recommended by previous research were tested: the expected return method, an actuarial-inspired approach presenting objective metrics; and bootstrapping, a technique successful in improving judgment consistency. The results show that the two biases can occur simultaneously and can have a huge impact on decision-making. The results also show that the two debiasing techniques are only partly effective. These findings suggest a need for more research on debiasing, and indicate some directions for exploring debiasing techniques and building decision support systems.

Author Keywords
decision making; decision support system; intelligent assistance; multiple-cue probability learning; cognitive bias; conservatism; loss aversion

ACM Classification Keywords
H.1.2 Models and Principles: User/Machine Systems—Human Information Processing; H.4.2 Types of Systems: Decision support

INTRODUCTION
Software is playing a more and more important role in supporting human decision-making. Various decision support systems (DSS) are being deployed in domains such as personal finance, medical diagnosis, public policy making, and organizational strategic planning. Despite this increasing prevalence, researchers have yet to fully understand how to design efficient DSS that truly help people make better decisions. Information design for a DSS is particularly challenging because the system has to: a) present large amounts of data in a way that helps decision makers correctly understand the data and its implications, and b) understand the potential biases in human decision making and present information and tools to help people overcome such biases.

Previous research in HCI has focused on the data visualization aspect of information design for DSS, and has applied various visualization techniques to achieve effective presentation of rich data. Appropriately designed information displays (e.g., [16]) not only support inspection of individual data points, but also facilitate comparisons across data and allows the user to gather more insights for decision making. Information displays for DSS may also include visualizations about the uncertainties of acquired information and risks associated with various decisions (e.g., [3]), which may help users make decisions according to their own risk preferences.

While information visualization is a familiar topic to HCI, designing information for remediating cognitive biases may seem foreign and distant. Psychologists have studied various forms of cognitive biases since 1970s, and have found that these biases can appear throughout many professions and cause substantial economic damage. For example, it has been found that due to people’s extreme aversion to losses, investors often choose bonds over stocks, missing potential financial gains by a factor of 100 [18]. As well, doctors are found to suffer from several biases in making diagnosis including the confirmation bias, which lead to inaccurate diagnosis and unnecessary and sometimes even harmful medical tests [1]. Given the prevalence of cognitive biases and the potential damages they may cause, it seems important for HCI researchers to take on the responsibility of designing DSS’s that remediate cognitive biases.

Recently, some HCI research has begun to focus on the issue of cognitive biases. For example, [22] explored the idea of presenting peer and majority investment decisions to older adults to help them overcome loss aversion. [11] showed that various biases such as the default bias (people tend to choose the default option) can be channeled to cultivate healthy dietary lifestyles. Similarly, [17] showed that some biases can be recruited to make users follow the recommendations of a DSS more often. The appearance of these studies shows that the cognitive aspects of human decision-making is of increasing interest to HCI research.

Work on cognitive biases in HCI is still rare however, and there appears to be a gap between how the HCI and how cognitive science approaches to studying cognitive biases. Both approaches have their merits. The HCI approach tends to focus on the practical aspect such as what can be
done with cognitive biases, whereas the cognitive science research tends to study when and how cognitive biases occur. While the HCI approach addresses more practical issues, the cognitive science approach leads to more fundamental and more generally applicable theories.

This study tries to take a middle ground between the HCI and the cognitive science approach to studying cognitive biases, and we aim to achieve three goals: 1) To design an experimental paradigm that captures many critical aspects of real-world decision-making. This paradigm should create opportunities for exposing cognitive biases and should serve as a platform for testing debiasing methods. 2) To understand when and how some biases occur in this new decision-making paradigm, and how the biases may interact to influence performance. 3) To test two popular debiasing methods (discussed next) that were proposed by previous research and find out whether they can succeed in remediating biases.

The Design of the Experiment
Real-world decision tasks typically consist of two stages: judgment and decision-making. In the judgment stage, the decision maker needs to evaluate the situation and predict the outcome of one or more events based on many task cues. For example, doctors make diagnoses based on a patient’s many test results, and mutual fund managers evaluate a stock based on its price, company earnings and other factors. In the decision-making stage, the decision maker must choose one of many alternative decisions based on the cost, benefit, and risks associated with each decision. For example, doctors sometimes need to decide whether to prescribe tests that allow a precise diagnosis but may also cause harm to the patient, and stock investors need to decide which stocks to invest in given their risks and potential return.

Though these two stages are well recognized as integral parts of a decision task, they often have been studied separately in psychology. Research that focuses on the judgment stage typically employs a multiple-cue probability learning (MCPL) paradigm [7], in which each event outcome is probabilistically associated with several task cues and the participants need to learn the probabilistic associations and predict the outcome in the test trials. Research that focuses on the decision making stage usually presents the participants with small gambles in which each trial has a safe option and a risky option from which to choose. Typically, the safe option offers stable, small returns and the risky option offers large returns but sometimes incurs losses. These two paradigms have been found to reliably produce many cognitive biases observed in real-world decision tasks, and are hence regarded as good vehicles for studying decision-making.

Our experiment combines the MCPL and the gamble-like decision-making paradigms to study judgment and decision-making as an integrated process. We first trained the participants in an MCPL task, and then in the decision-making stage asked them to make decisions among pairs of alternatives. The stimuli in the two stages were generated with the same statistical process so that the participants could use the skills they learned from the MCPL stage to evaluate the prospects of the alternatives in the decision-making stage. This design permits us to assess any interaction between the two stages, including how biases in judgment may affect decision-making. By reusing two experimental paradigms that have been well tested in psychological research, we can strip away many nuisance factors that might affect decision-making and study more clearly when and how biases occur and interact. By combining the two paradigms, we are taking one step further towards studying real-world decision-making situations where both the judgment and decision-making stages are often present, and understanding which debiasing techniques can remediate biases in these scenarios.

Cognitive Biases
This study examines two cognitive biases that frequently appear in real-world decision-making: conservatism and loss aversion. Conservatism describes how people adjust their beliefs insufficiently in light of new information. This bias affects judgment accuracy, and is often manifested as an underestimation of high probabilities and overestimation of low probabilities [2, 5] because of the overuse of the middle of the probability scale. Conservatism can lead to inaccurate judgments for critical real-world decision tasks. For example, it is regarded as a likely cause for why investors tend to under-react to changes in stock fundamentals such as earnings and dividends [8].

The other bias studied here, loss aversion, has been shown to have a very strong and robust influence on people’s decisions [19]. Loss aversion describes people’s tendency to disproportionately prefer avoiding losses to acquiring gains. For example, people generally do not want to take a gamble that has a 90% chance of losing $10 and 10% chance of winning $100, even though the long-term average return, or expected return, of the gamble is $1 positive. This bias leads to suboptimal decisions when risks are involved, and its effect has been observed in many important instances of personal, corporate, and public decision-making.

In the present study, we expect to observe conservatism in the judgment stage because the participants need to estimate probabilities, and loss aversion in the decision-making stage because the participants need to choose from two alternatives, one of which is more likely to bring losses than gains.

Debiasing Techniques
Two debiasing techniques were tested in this study to see whether they can alleviate conservatism and loss aversion: a) Using a bootstrap model to provide suggestions to participants to improve the accuracy of their judgments, and b) showing participants the expected return of each alternative in the decision task to lead them to optimal decisions.

The bootstrap technique has been shown in many MCPL studies [21, 4] to increase probability judgment accuracy by creating a quantitative model of experts’ previous forecasts. It is found in these studies that human judgments, including experts’, are not only biased (e.g., show conservatism), but also inconsistent [15]. However, efficient rules can be extracted from an expert’s past judgments through modeling techniques such as linear regression, and it is found that these
bootstrapped models can achieve higher judgment accuracies than their human counterparts. We thus apply this technique in our experiment to see whether the participants would follow the suggestions of the bootstrap model and whether the bootstrap model can correct the conservatism bias.

The second technique, showing participants the expected return of the different options, has been found to help reduce loss aversion in older adults when the expected return accurately reflects the long-term average return [14]. We are thus interested whether this method will have an effect for our experiment.

The exploration of these debiasing techniques will provide lessons for the information design of decision support systems. For example, if the bootstrap technique is shown to work, then a good information design for medical DSS should provide doctors suggestions that are constructed with the bootstrap method. Similarly, if the expected return method is shown to work, then a good design for a stock DSS should show expected return in addition to the current popular metrics such as risk scores. If, however, these debiasing techniques do not work as intended, we can still see where the techniques failed, which will help in future design and exploration of debiasing methods.

**EXPERIMENT OVERVIEW**

The experiment was designed as an investment game, in which the participant played the role of a venture capitalist who wants to maximize profit by investing money in startup companies. The experiment was divided into two stages: the probability judgment stage and the decision-making stage. In the probability judgment stage, the participant was trained and tested in judging a fictitious startup company’s probability of success given multiple cues in a profile table as shown in Figure 1. The goal of this experimental stage was to (a) examine whether conservatism would appear in a multiple-cue probability learning task, and (b) train the participant with the basic judgment skill needed to make good investment decisions in the second, decision-making stage.

In the decision-making stage, the participant was presented with two candidate companies in each trial, and was asked to choose one to invest money in. The two candidate companies always included a safe option and a risky option. The safe option had a high success likelihood but low potential returns, whereas the risky option had a low success likelihood and high potential returns. The participant would only gain the returns when the invested company succeeded. Thus, a good investment decision required an accurate judgment of each company’s success rate, which the participants were trained to do in the first stage, and a sensible investment strategy that balanced risks and profits. In this decision-making stage, we tested the aforementioned two debiasing techniques: (a) teaching the participants to use the expected return method to decide which company to invest in, and (b) using a bootstrap model to provide participants with suggestions about companies' success likelihoods. These techniques were assigned to two participant groups, and the results were compared with a control group to determine if they could effectively remediate cognitive biases.

This section presents the overall procedure of the experiment. The next two sections will present more detailed experimental design and the results of each experimental stage.

**Apparatus and Procedure**

Figure 1 shows an example of the company profiles used in the experiment. Every profile included ratings on a 1–5 scale for four attributes: leadership ability, proprietary technology, market condition, and competitor strength. Higher ratings in the first three attributes suggested higher success likelihood, whereas higher ratings in competitor strength indicated lower success likelihood. These ratings were randomly generated from a multivariate normal distribution, and they determined a company’s success likelihood through the following equations:

\[
\text{Odds Ratio} = -2.4 + 0.4L + 0.6P + 0.2M - 0.4C
\]

\[
\text{Success Likelihood} = \frac{\text{e}^{\text{Odds Ratio}}}{1 + \text{e}^{\text{Odds Ratio}}}
\]

where L, P, M, and C represent the four attribute ratings. The success likelihood then stochastically determined whether a company would succeed or fail.

Many company profiles were generated and were divided into nine groups by rounding their success likelihoods to the nearest decile. That is, the profiles were grouped into nine deciles from 10% to 90% at 10% intervals. Profiles that fell in [0, .05] and [.95, 1] were discarded because they only covered 5% range. The probability judgment stage and the decision-making stage then randomly drew stimuli from the nine decile groups.

The experiment was implemented as a web application and hosted on Amazon EC2. After obtaining the participant’s informed consent, instructions were presented that explained the tasks and particularly how, of the four attributes, only the competition strength was negatively associated with the success likelihood. To ensure that participants understood the instructions, several multiple choice questions were presented before each stage began, all of which needed to be answered correctly before participants could start a stage.

**Participants**

Seventy-two participants (33 females; mean age = 33, range 19–61) were recruited from Amazon’s Mechanical Turk website. Restrictions were set so that only people in the U.S. and people who had more than 95% approval rate (meaning that more than 95% of their prior tasks were approved by
the task requesters) could participate in the experiment. The participants spent on average about an hour to complete the experiment. Each participant received a base compensation of $4, and up to $6 or $9 bonus depending on the experimental condition that the participant was assigned to. The exact payoff scheme will be discussed in the next two sections.

THE PROBABILITY JUDGMENT STAGE

Task Procedure and Design
The first experimental stage, the probability judgment stage, contained training blocks and test blocks. In the training blocks, the participant was presented with a company profile in each trial and three buttons “fail”, “not sure”, and “succeed”. The participant clicked on one of these buttons to indicate her guess of the outcome, and was then shown the actual binary outcome, “succeed” or “fail”. The participant was expected to learn from these outcomes the associations between ratings and success likelihood.

In the test blocks, the participant was presented with a company profile in each trial and 11 radio buttons for indicating her estimation of the company’s success likelihood ranging from 0% to 100% at 10% intervals. The participant was rewarded 5 cents if the estimated success likelihood was within 10 percentage points of the actual success likelihood, 3 cents if within 30 percentage points, and 1 cent if within 50 percentage points. The participant was, however, not informed about the exact reward scheme, but was told that she would receive more rewards for more accurate estimations, and was shown only the reward she earned after each trial. This limited feedback was to ensure that the participant could not infer the success likelihood from the feedback. This is important because our goal was to understand how well people could learn probabilities by observing the stochastic binary outcomes of the training stage, rather than from feedback associated with the actual success likelihood.

There were two training blocks and two test blocks in the order of training, test, training, and test. Each training block contained 54 trials, and each test block contained 27 trials. The trials were balanced across the nine decile groups. That is, there were equal number of trials with company success likelihood of each decile.

Results of the Probability Judgment Stage
Figure 2 shows for the training blocks the proportion of trials in which the participants chose “succeed”, “fail”, or “not sure”. Error bars represent the 95% confidence intervals (CI) of the participant mean.

Figure 3 shows for the test blocks participants’ estimated success likelihood, or subjective success likelihood (SSL), as a function of the OSLs across the nine decile groups. The gray diagonal line shows what the ideal response would look like if the SSLs were perfectly aligned with the OSLs. The fact that the SSLs are close to the diagonal line suggests that participants were able to predict success and failures somewhat reliably just by examining the company ratings.

More importantly, Figure 3 shows that though SSLs were strongly correlated with OSLs, there were also strong biases in participants’ SSLs. For OSL < 60%, SSL was generally higher than OSL, whereas for OSL > 60%, SSL was generally lower than OSL. This is consistent with the conservatism bias discussed in the introduction. The participants clearly overestimated low probabilities and underestimated high probabilities.
Figure 4. The objective coefficients (light gray) used in Equation 1 for generating the OSLs, and the participants’ regression coefficients (dark gray) extracted after applying linear regression to participants SSLs. Error bars represent the 95% CI.

Figure 4 compares the objective coefficients used in Equation 1 for generating the OSLs with the regression coefficients extracted from participants’ SSLs. Participants’ coefficients were obtained by applying logistic regression using SSLs as the response variable and the four attributes plus an intercept as the predictors. As can be seen in the graph, participants’ coefficients followed the objective coefficients closely, suggesting that the participants somewhat accurately learned the underlying weights assigned to the different attributes. Particularly, there was no significant difference between participants’ and objective coefficients for the leadership attribute, between participants’ and objective coefficients for the market-condition attribute, those in the first test block, and the competition-strength attribute. As can be seen in the graph, participants’ coefficients followed the objective coefficients closely, suggesting that the participants somewhat accurately learned the underlying weights assigned to the different attributes. Particularly, there was no significant difference between participants’ and objective coefficients for the market-condition attribute, between participants’ and objective coefficients for the expected return (ER) of the risky option varied to create two conditions: the high risk condition in which the success likelihood was very low (10% or 20%), and the low risk condition in which the success likelihood was medium (50% or 60%).

In summary, in this first stage of the experiment, participants seemed to have achieved good judgment accuracy after the first training block, and going through the second training block did not improve their accuracy significantly. Participants’ average SSLs for each of the nine deciles in the second test block were virtually the same as those in the first test block, $t(16) = −0.63, p = .51$. It appeared that with the instructions and the first 54 training trials, participants soon learned to estimate the probabilities but were not able to improve with more practice.

In summary, in this first stage of the experiment, participants learned to predict the success likelihood of a company based on the four attribute ratings. They learned the different weights assigned to the attributes, but their learned weights did not completely match the actual weights. The participants exhibited a strong tendency to overestimate low probabilities and to underestimate high probabilities. Such a bias is consistent with those observed in [6, 5], but it is perhaps the first time that this conservatism bias is observed in a experience-based, multiple-cue probability judgment task. The next section will examine the second stage of the experiment, the decision-making stage, to see whether there were biases in participants’ investment decisions, and whether our debiasing techniques helped remediate biases.

**THE DECISION-MAKING STAGE**

**Task Procedure and Design**

In the decision-making stage, each trial presented the profiles of two fictitious startup companies, and the participant needed to choose one in which to invest $1 million (in the game’s virtual currency). In addition to the attribute ratings, the company profiles shown in this stage also contained the potential profit of a company that would be gained if the participant chose to invest in the company and if it were to succeed. The potential profit ranged from $1.8 million for the safest investments to $29 million for the riskiest ones. If the chosen company failed, however, the participant would lose the $1 million investment. Each million dollar gain was worth 4 cents in compensation, and participants earned on average $3.6 bonus in this stage across 48 trials.

The design of this decision-making task followed a typical paradigm used in decision research: every decision was a choice between a risky and a safe option. While many decision studies use certainty candidates (i.e., a 100% probability of delivering a reward) as the safe option, we used high probability candidates instead. Such a design is perhaps closer to real-world decision tasks, where events are rarely 100% certain. Also, we were interested to see whether the same kind of cognitive biases such as loss aversion exist in this slightly modified paradigm.

The success likelihood of the risky option was varied to create two conditions: the high risk condition in which the success likelihood was very low (10% or 20%), and the low risk condition in which the success likelihood was medium (50% or 60%). Table 1 shows the design of these two conditions. The success likelihood of the safe option was maintained at 80% or 90%, while that of the risky option varied.

Another factor manipulated was the expected return (ER) of the risky option. The expected return was calculated by the following equation:

$$ ER = P_{success} \times V_{profit} + (1 - P_{success}) \times V_{loss} \quad (3) $$

where $P_{success}$ is the success likelihood, $V_{profit}$ is the potential profit that would be gained if the company succeeds, and $V_{loss}$ is the investment that would be lost if the company failed, which was always $1 in this experiment. Note that the profiles given to the participants contained $V_{profit}$ rather than ER. To change ER, $V_{profit}$ had to be varied based on

<table>
<thead>
<tr>
<th>Success Likelihood</th>
<th>Risk Level</th>
<th>Risky</th>
<th>Safe</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Risk</td>
<td>10% or 20%</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>Low Risk</td>
<td>50% or 60%</td>
<td>80%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Table 1. Two risk-level conditions. The success likelihood for the risky option varied depending on the risk-level condition. The success likelihood for the safe option was always 80% or 90%.

\[1\] Significance level is set at 0.01 due to multiple comparisons.
the $P_{\text{success}}$ of a given company, which is why $V_{\text{profit}}$ varied substantially from $1.8$ million for the most safe option to $29$ million for the most risky option.

Table 2 shows exactly how the expected return of the risky option was varied to create different risk-premium conditions. Risk premium refers to the difference between the ER of the risky option and the ER of the safe option. The ER of the safe option was maintained at $1.5$ million, while the ER of the risky option was set to $2$, $1.5$, and $1$ million to generate positive, neutral, and negative risk premiums. The goal of changing risk premium was to examine whether participants were sensitive to changes in expected return, and whether they could learn to exploit the risky option when the risk premium was positive and to avoid the risky option when the risk premium was negative.

The third factor of the decision-making task controlled which debiasing techniques were used to help participants make investment decisions. Three conditions were created: control, ER (expected return) aided, and ER-Bootstrap aided. The control condition did not provide any debiasing aid, and hence the decisions could only be made based on the companies’ attribute ratings and potential profit.

In the ER aided condition, participants were provided with instructions about how to use the expected return method and with a worksheet widget to help them easily calculate ER. To ensure that participants understood the ER method, multiple choice tests were administered after the instructions, which had to be answered all correctly before proceeding to the task. The worksheet widget was designed for ease to use. Participants only needed to enter their estimated success likelihood for each company and the worksheet would calculate the ER using Equation 3 and present it. The goal of these measures was to increase the adoption rate of the strategy and reducing the effort needed to apply the strategy. However, we did not want the participants to blindly follow this strategy. Indeed, because the effectiveness of ER depends on the accuracy of participants’ estimated success likelihood and because there were biases in participants’ judgments, ER may not be the best investment strategy. Thus, the participants were told that investing in the company with the higher expected return is in general a good strategy but it does not guarantee higher profit.

The ER-Bootstrap aided condition used the same ER procedure as above, and in addition provided suggestions for the success likelihood needed by the ER worksheet widget. These suggestions were calculated using the bootstrap method. That is, they were calculated based on the participant’s past SSLs rather than the OSLs. Specifically, a logistic regression model was constructed for each participant using the success likelihood estimations the participant provided in the probability judgment stage. (This is the same procedure used to calculate the coefficients for Figure 4). Then, the regression model was used to predict what the participant’s SSL would be given the profiles in the current decision task. The participant could choose to enter the suggested likelihood into the ER worksheet, or provide their own estimation. By providing participants with suggested SSLs, we hope to increase the accuracy of the success likelihood that went into the calculation of ER, and thereby potentially increase the effectiveness of the ER method.

Participants were randomly assigned to one of these three decision-aid conditions, resulted in 24 participants per condition. In the ER and ER-Bootstrap aided conditions, the participant completed two blocks of decision trials, with 24 trials per block. These 24 trials were balanced across the risk premium factor and the risk level factor. In the control condition, the participant first completed two blocks without decision aids, which served as the baseline to be compared with the other two experimental conditions. The control group then did two additional blocks with the ER aid, which repeated the first two blocks but with the trials presented in a different order. The purpose of the two additional blocks for the control group was to use the ER worksheet to collect participants’ SSLs for the profiles of the decision-making stage (the other two conditions already collect SSLs through the ER worksheet). The decisions that the control group made in these two additional blocks were not analyzed, but the SSLs were needed for the data analyses presented next.

Results of the Decision-Making Stage

Figure 5 shows the average of the SSLs the participants entered into the ER worksheet across the six OSLs used by the risky and safe options. The solid line shows the SSLs from the non-bootstrap conditions, which includes the ER condition and the second two blocks of the control condition. The dashed-line shows the ER-Bootstrap condition. As can be seen, the same conservatism bias—underestimation of high probabilities and overestimation of low probabilities—observed in the judgment stage appeared across all conditions in the decision-making stage. Furthermore, it seems providing suggested success likelihood calculated with the bootstrap method did not help remediate the conservatism bias in that the dashed line was not very different from the solid line, $\chi^2(1) = .12, p = .73$.

A further examination of the data shows that the participants used the suggested success likelihood very often when provided, which indicates that the ineffectiveness of the bootstrap method is due to the method itself rather than participants’ failures to follow the suggestions. In fact, the SSLs in the ER-Bootstrap condition matched the suggested success likelihoods very well, $R^2 = .886$. Whereas for the other two conditions, the $R^2$ between SSLs and the offline-bootstrapped likelihood suggestions was only .561. Thus, it seems that the bootstrap method really could not help remediate the conservatism bias. This is an important...
neutral finding that has not come up in previous multiple-cue probability learning studies (e.g., [12, 9, 21]), in which the bootstrap predictions were considered much more accurate and consistent than human judgments.

Figure 6 shows the proportion of risky choices the control group made across the risk-premium and risk-level conditions. As can be seen, within each panel, as the risk premium increases from negative to positive, participants were more and more likely to choose the risky option, \( \chi^2(2) = 14.8, p < .001 \). This suggests that the participants were somewhat aware of the expected return of the two alternatives in each trial, even without the help of an ER worksheet. However, it does not mean that the participants’ strategy approached the optimal. In fact, the data indicate a strong bias towards loss aversion. For example, for the positive risk premium conditions, the optimal strategy is to always invest in the risky option, whereas the participants only invested in the risky option on about 40% of the trials.

Comparing the two panels of Figure 6, it can be seen that participants were generally more likely to invest in the risky option when its risk is low (right panel, 50% or 60% success likelihood, as summarized in Table 1) than when its risk is high (left panel, 10% or 20% success likelihood), \( \chi^2(1) = 39.3, p < .001 \). This result is surprising considering that due to conservatism, the participants had overestimated the success likelihoods of the high risk profiles, which means that from their point of view, the ER of these risky profiles should be even higher than their actual ER, and yet the participants were still more likely to choose the safe option. In other words, the tendency to be loss averse is so strong that overestimating the success likelihoods of the high risk profiles did not make them more attractive.

Figure 7 compares the proportion of risky choices of the control condition with those of the ER and ER-Bootstrap conditions. It seems from this graph, the two debiasing techniques—helping participants use the ER method and providing participants with suggestions for success likelihood estimates—had little effect in remediating participants’ biases. The decision aid factor did not have a significant effect, \( \chi^2(2) = .07, p = .97 \).

Figure 8 shows further where the ER method succeeded and failed. Similar to Figure 7, Figure 8 also presents the proportion of risky choices across the risk premium and risk level conditions, but in this graph, the risk premium and the risk level were measured from the participants’ point of view. Here, the subjective risk premium was calculated using the expected return obtained with participants’ subjective success likelihoods (SSL), unlike in Figure 7 where the risk premium was calculated using the objective success likelihoods (OSL) as in Table 2. There is no neutral risk premium condition in this graph because participants’ subjective ER for the two companies in each trial were rarely the same. The subjective risk level was similarly designated based on the SSL rather than the OSL. The subjective risk is designated high when the SSL for the risky option is below 40%, and is designated low otherwise. This separation at 40% ensures that the two risk levels has about equal number of trials.

Figure 8 shows that the ER technique had an effect when the risk premium is negative, but the technique did not help remediate loss aversion. Specifically, participants became more risk averse when it was advantageous (subjectively) to be so. This can be seen in the graph that when the subjective risk premium is negative, the ER and ER-Bootstrap conditions had smaller proportion of risky choices than that of the control condition, \( \chi^2(1) = 12.7, p < .001 \). When the subjective risk premium is positive, however, the debiasing techniques did not have a strong effect (\( \chi^2(1) = 1.16, p = .28 \)), suggesting that the participants were still strongly affected by their loss aversion even when they knew that the risky option had a higher expected return.

Figure 5. The average SSL that participants entered in the worksheet in the decision-making stage for the six OSLs used by the decision task. The gray diagonal line represents the ideal response without bias. Error bars represent the 95% CI.
DISCUSSION

Cognitive Biases

The results from this experiment show that like in many other probability judgment tasks, in a multiple-cue task, the conservatism bias can also appear. Remarkably, our participants seemed to have successfully learned consistent associations between company attributes and success likelihood by just observing many probabilistic outcomes, although their learned associations did not completely match the actual weights assigned to the attributes. The conservatism bias, specifically in this experiment the overestimation of low probabilities and the underestimation of high probabilities, was thus not due to a failure of learning consistent attribute-likelihood associations, but was a robust cognitive phenomenon.

The results of our decision-making task showed that participants had a strong cognitive bias towards avoiding losses. Even when the participants were aware that the risky option had a higher ER, they still tended to invest in the safe option (as seen in Figure 8) to avoid potential loss.

Evaluation of the Debiasing Techniques

Neither of the two debiasing techniques—improving the consistency and accuracy of participants’ subjective success likelihood (SSL) with the bootstrap model, or teaching and helping people use the ER method to make investment decisions—worked particularly well for remediating cognitive biases. The bootstrap method did improve the consistency of the SSLs as the participants generally followed the suggested SSLs, but it did not improve their accuracy. It was clear that the bootstrap model reproduced
participants’ conservatism bias instead of correcting it. It seems that for this experiment, the claim that “a biased model consistently applied is an improvement over a biased and inconsistent human” ([10], p. 328) does not apply, partly because the participants were already very consistent without the suggested SSLs. In real-world tasks, however, because of other factors such as redundant attributes and insufficient training, improving judgment consistency with the bootstrap method may be useful. However, correcting conservatism likely requires additional procedures.

The ER method helped participants recognize when the risky option was a bad choice (when its expected return was lower than that of the safe option), though this usefulness was modulated by the accuracy of participants’ SSL. Once the participants recognized that the risky option had a negative premium, participants tended to avoid choosing it (as shown by Figure 8 in which the risk is measured subjectively). However, because of overestimation of small probabilities, when the risk was high, participants often overestimated the risky choice’s ER and could not notice that the risky choice actually had negative premium (as shown by Figure 7 that ER did not help people avoid negative-premium risks). Thus, it is important to correct the conservatism bias in designing a decision support system to help people avoid making a decision whose risk outweighs benefit.

The ER method, however, had little effect in correcting loss aversion when the risky option had a positive risk premium. It was found that even when participants knew that the risky option had a positive risk premium (meaning that they would gain more in the long run by always choosing the risky option), they still tended to avoid the risks. This suggests that the bias of loss aversion was not due to incomplete information about the choices or a lack of good decision strategy, but an inherent cognitive preference towards certainty and an aversion towards loss.

**Directions for Future Research**

Our findings have important implications for future research on information design of decision support systems (DSS) that strive to remediate cognitive biases. Particularly, our findings show that the suggestion of many prior experience-based decision-making studies (e.g., [13])—that loss aversion can be corrected by improving the accuracy of people’s judgments—is likely incorrect. Because loss aversion is independent of conservatism, the design of a DSS should consider and address both biases. If these biases are corrected, our results suggest that it would impact decision-making in two ways: correcting conservatism bias would help people avoid risks when the risk outweighs the benefit, and correcting loss aversion would help people take risks more often when the benefit outweighs the risk.

Our results also show that attempting to debias decisions through modeling of the problem domain and experts’ decision rules, as is done in many current DSS’s [10], is unlikely to adequately address cognitive biases. The core function of many DSS’s is to implement a computational model of the problem domain which can produce predictions given certain inputs. For example, in oil field acquisition, oil companies typically run geological models with initial drilling and seismic data as well as geologists’ suggestions as model inputs to guide decisions. Such models are very important and helpful in improving decision-making in that they alleviate the burden on decision makers of heavy computational details and allow accurate insights into the implications of various decisions. However, there are still two stages where cognitive biases can creep in: 1) inputting parameters for the domain model, and 2) making decisions after acquiring model predictions. In the present study, the ER equation (Equation 3) is a simple version of a domain model in that it implements a rational strategy and it alleviates the participants from the burden of computational details. As we have seen, the parameter inputs required by the ER—success likelihood estimations—were indeed affected by the conservatism bias, and at the decision-making stage, the participants’ decisions were also affected by loss aversion even though they had accurate predictions calculated by the ER worksheet widget. These results show that if the goal of designing a DSS is to help people make better decisions, then we should not stop at merely presenting model predictions to the decision maker; the DSS should oversee the whole process of a decision task, including the initial human parameter-inputs, and the final decision-making stage.

Our results also show that the combination of the more theory-oriented, cognitive-science approach and the more practical oriented, HCI approach can lead to fruitful findings, and future research on cognitive biases perhaps should continue with this integrated approach. The cognitive science approach inspired a clear, minimal experimental design that allowed us to discover the separation of the two biases and their interaction effect. This then lays the theoretical ground for designing debiasing techniques. The HCI approach encouraged us to test existing debiasing methods and allowed us to discover their flaws. The final established experimental paradigm allows robust reproduction of cognitive biases, and can be reused in tests of other debiasing methods.

**CONCLUSION**

Through a novel experimental design that combined a multiple-cue probability learning (MCPL) paradigm with a two-alternative decision-making paradigm, we showed that cognitive biases, specifically conservatism and loss aversion, can impact people’s performance in complex decision-making tasks. Particularly, it was found that these two biases can coexist in decision making, which is somewhat surprising given that previous theories [19] could not account for this result. Our results show that loss aversion is likely an independent phenomenon from the biases in probability judgment, and it seems that loss aversion dominates the error in probability judgment such that overestimating small probabilities did not make people more risk seeking.

Our exploration of the two debiasing techniques, bootstrap and expected return (ER), did not achieve satisfactory results, but it revealed where the true obstacle lies. The bootstrap technique did help participants to be more consistent, as advocated by MCPL studies, but these studies overlooked the effect of biases in human probability judgments. The
ER method helped participants make more rational decisions only when their probability judgments were accurate and the rational decision was in favor of loss aversion. When the recommendation of the ER method was against people’s bias, however, it was often ignored. This type of selective adoption coincides with some real-world cases. Shell geologists once (correctly) analyzed that a field contained over one billion barrels of oil, but this number seemed too good to be true, so they revisied their analysis to just “over 200 million barrels” [20], resulting in a costly decision (not to invest). This goes to show how strong cognitive biases can be, and how much impact they may have economically.

The work presented here is but a first step towards building cognitive systems that can reliably improve human decision making, particularly in real-world contexts. Nevertheless, it demonstrates that cognitive biases can have as significant an impact as other human factors and should be considered in information design to support decision-making. Doing so well presents a significant challenge to the HCI community and ongoing research.

REFERENCES


20. Whaley, J. The first UK giant oil field. GEO ExPro 7, 3 (2010).
