Misestimating Probability Distributions of Repeated Events

Oleg Urminsky (oleg.urminsky@chicagobooth.edu)
University of Chicago, 5807 S. Woodlawn Ave
Chicago, IL 60637 USA

Abstract

This paper examines people's subjective beliefs about probability distributions arising from repeated events, such as the number of heads in ten coin flips. Across elicitation methods and decision scenarios, people express beliefs that are systematically biased relative to the actual distribution, over-estimating the tails and under-estimating the shoulders of the distribution. While experts are relatively more accurate than novices, both show significant bias.

Keywords: Judgment; Decision Making; Probability Distributions; Subjective Belief; Intuitive Statistics.

This paper poses a simple but fundamental question. When people know the probabilities of specific events, how accurate are their beliefs about the distribution of repeated events, relative to the actual probability distributions? For example, how well can people predict the distribution of the number of heads in ten coin flips, relative to the binomial distribution?

Study 1

Study 1 collected data on people's prospective beliefs about the Binomial distribution, with 10 outcomes and an equal probability of both outcomes: B(10 , 0.5). In the study, each participant read one of four real-world scenarios and then indicated their beliefs about the distribution using one of three different elicitation methods. Participants also estimated a distribution of height as a control task, and were paid an incentive for accuracy in the tasks.

Method

Adult online participants completed 821 surveys. Participants were told that they were eligible to receive a bonus payment, such that providing the most accurate answer would earn them $1, providing an answer that was no better than guessing at random would earn $0, and answers of intermediate accuracy would earn the corresponding intermediate amount.

Each participant initially read one of four estimation scenarios, summarized below:

1. *Coin Flip Game.* You would flip a coin 10 times and show the experimenter the result each time. Each time that it comes up heads, you win $1, and each time it comes up tails, you get nothing.
2. *Survey Sampling.* In a population exactly half of the people prefer Coke to Pepsi, half prefer Pepsi to Coke, and no one is indifferent. You conduct a survey with 10 people, and there is no sampling bias (everyone has an equal probability of completing the survey).
3. *Soccer Practice.* You would kick a soccer ball into a goal, with the difficulty of the game adjusted for you personally. You would kick the ball from far away enough that, on any given kick, you have a 50% chance of getting the ball in. You would try 10 kicks.
4. *Estimating Height.* Thinking only of men in the U.S. who are 18 years old or older, how common do you think each of the heights below is? Indicate the proportion of men in the U.S. 18 or older who are in each height range.

Participants then estimated 11 quantities which added up to 100: probability of earning from $0 to $10 in (1), the number of people out of 100 earning from $0 to $10 (1), probability of surveying from 0 to 10 people who prefer a given soda in (2), probability of making between 0 to 10 of the kicks in (3), or the proportions of heights in each of 11 intervals in (4).

The quantities were estimated either with an adjustable histogram, by filling in 11 numeric values which were forced to add up to 100 or by choosing between one of six predefined histograms. The order of the response choices or histograms was fully counterbalanced.

After the primary task, participants who had estimated one of the non-height tasks (scenarios 1-3) then also did the height estimate task. The study therefore had a 5 (scenario tasks) x 3 (elicitation methods) between-subjects design, as well as a repeated measure (tasks 1-3 followed by the height estimation) for a subset of participants.

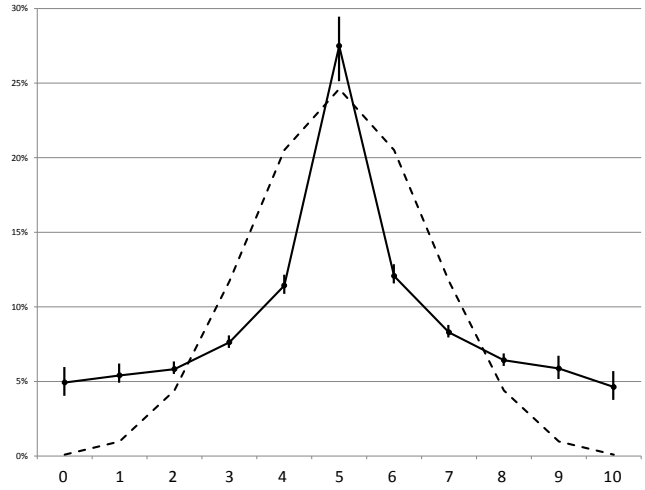
Participants answered several demographic questions and finished the survey. Later that week, participants' bonuses were calculated based on their accuracy and paid.

Results and Discussion

Accuracy of beliefs for elicited distributions.

The first three scenarios, which are represented by the same binomial distribution (under the assumption that each of the ten outcomes is independent) are analyzed first. Combining all the data for the open-ended elicitation methods (histogram and numeric), we have 418 completes. The mean estimates for each outcome are plotted below, in Figure 1, using a solid line, along with the binomial distribution using a dashed line, for comparison.

Figure 1: All Continuous Binomial Estimates



Error bars represent 95% confidence intervals.

As can be seen in Figure 1, the estimates diverge substantially from the binomial distribution. The estimates have significantly more mass in the tails and peak of the distribution, and less mass in the shoulders of the distribution.

The mean of the average estimated distribution was not biased (5.04 vs. 5, t=0.72, p=.47). However, on average, the estimated distributions had a significantly higher variance than the binomial (4.93 vs. 2.5, t=14.3, p<.001) and more kurtosis (.09 vs. -.20, t=3.54, p<.001).

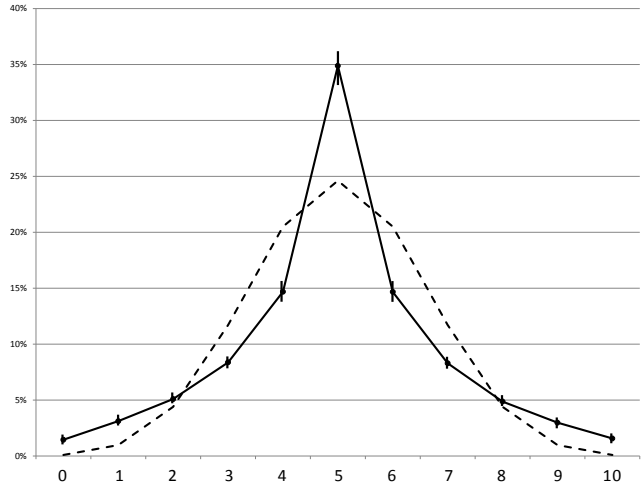
There are several interpretations of the difference between peoples' estimates and the actual binomial distribution that are important to consider. One possibility is that uninformed or unmotivated participants may have used a simple heuristic to solve the problem, perhaps even one they did not really believe. Could heterogeneity among participants, and specifically a preponderance of clearly unrealistic distributions, account for the observed divergence between the estimates and the binomial distribution?

One possibility is that they guessed randomly or fundamentally misunderstood the question. Participants were coded as monotonic if their responses were monotonically decreasing from a maximum at 5, on both sides of the distribution. A majority of participants (n=291, 71%) gave distributions satisfying this criteria. The mean of the average estimated distribution was not biased (5.00 vs. 5, t=0.18, p=.86). However, on average, the estimated distributions still had a significantly higher variance than the binomial (4.17 vs. 2.5, t=10.7, p<.001) and more kurtosis (.15 vs. -.20, t=4.23, p<.001).

Another heuristic would be to provide a uniform (or near-uniform) distribution. This was the case for approximately 15% of the total participants. Some participants (6% of the total) simply put all the mass on one outcome.

Eliminating non-monotonic, near-uniform and single-point distributions yielded 231 responses, 55% of the sample. The mean of just these distributions is shown below in Figure 2, plotted with a solid line, with the binomial distribution plotted using a dashed line for comparison.

Figure 2: Subset of Monotonic Binomial Estimates



Error bars represent 95% confidence intervals.

The results in Figure 2 are similar to those in Figure 1, suggesting that the observed divergence from the binomial distribution cannot be explained by a subset of participants providing non-monotonic distributions. The estimates have significantly more mass in the tails and peak of the distribution, and less mass in the shoulders of the distribution.

The mean of the average estimated distribution was not biased (5.00 vs. 5, t=0.64, p=.53). However, on average, even these estimated distributions pre-screened for the plausibility of their shape had a significantly higher variance than the binomial (3.55 vs. 2.5, t=9.14, p<.001) and more kurtosis (.42 vs. -.20, t=7.00, p<.001).

A final possibility to consider is that individual estimates were based on the actual amount plus random error (Erev, Wallsten and Budescu 1994). This could explain overestimation of the low probabilities in the tails, due to truncation of estimates at zero. However, this would not explain the lack of underestimation and actual overestimation of the highest probability outcome.

Accuracy of beliefs for chosen distributions.

The results shown thus far are based on having participants generate probability distributions, using either a graphical or numeric interface. It is possible that the observed misestimation might be specific to elicited, and therefore constructed, distributions. To test this, a total of 239 participants were instead shown six distributions (order counterbalanced and distributed across the scenarios) and asked to choose which was the most accurate, second most accurate and least accurate. Choices of the most and least accurate are shown below in Figure 3.

Figure 3: Choices Among Histograms

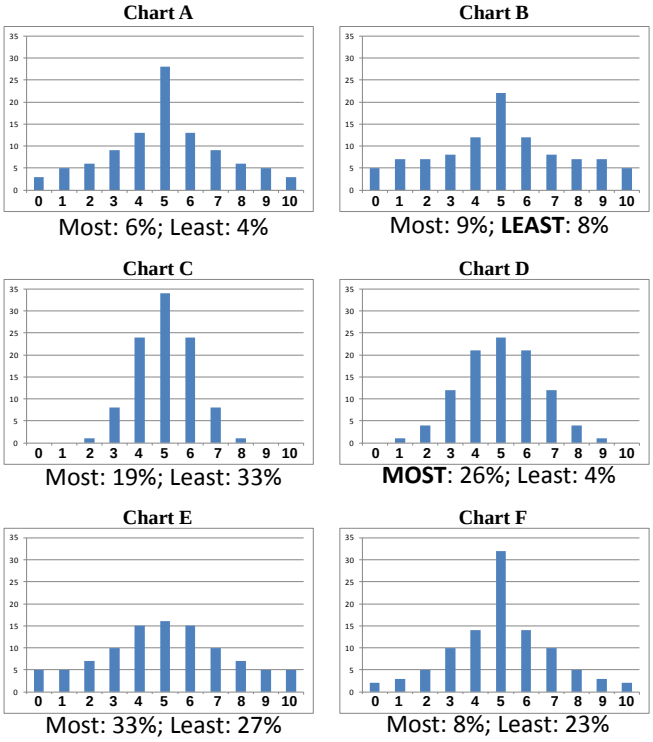


Chart D was the most accurate, but was only chosen as the most accurate by 26% of the participants. While this rate is significantly above chance (Z=3.14, p=.002), it's a fairly low proportion. Only 12% correctly identified the second most accurate chart (Graph C) and only 8% identified the least accurate chart (Graph B), both significantly below chance (Z=2.38, p=.02 and Z=4.63, p<.001, respectively).

Treating each person's choice for the most accurate distribution as their estimate, the average estimated distribution was calculated. The average of the chosen distributions significantly overestimated the share of the tails (0, 1, 2, 8, 9 and 10), and significantly underestimated the share of the shoulders (3, 4, 6 and 7). However, the proportion of the most likely outcome (5) was unbiased. The average variance was significantly higher than in the true distribution (4.21 vs. 2.5, t=12.09, p<.001). The average kurtosis, however, was lower than in the true distribution (-0.33 vs. -0.2, t=8.13, p<.001), unlike for elicited distributions.

Heterogeneity of distributions.

In this study, both the elicitation methods and the underlying scenarios were varied, in order to test the robustness of the results. A regression analysis, shown below, was conducted to test whether these task differences affected the accuracy of the estimated distribution, as measured by the root mean-squared error between the estimated and actual distribution. The baseline categories for the regression were choosing among the six graphs in the coin probability task.

Figure 4: Regression on Error of Estimate (RMSE)

	β	SE	t	p
Constant	9.52	1.55	6.16	.000
Elicitation Method:				
Histogram	17.81	1.60	11.15	.000
Numeric	18.61	1.58	11.79	.000
Scenario:				
Coin - frequency	0.60	1.83	0.33	.744
Soccer	2.88	1.84	1.56	.120
Survey	3.64	1.85	1.97	.049

Having participants choose from among a set of sample graphs was the most accurate elicitation method, while having participants generate a histogram or provide numeric estimates were both significantly less accurate. There was no significant difference in accuracy between the histogram and numeric elicitation methods.

The results were largely consistent across scenarios. Participants were the most accurate when estimating the probabilities of each outcome in the coin flip scenario. Accuracy was similar, but not improved, when the scenario was reframed in frequentist terms, and they instead estimated the proportion of a 100 people who they expected to have each of the outcomes. There was also no significant difference with estimates in the soccer scenario.

Participants were the least accurate in the survey scenario, primarily due to an even higher over-estimation of the middle option (an equal number of Coke and Pepsi preferers). This may be due to the salience of the fact that the proportions are equal in the population, which is emphasized in the scenario.

An additional analysis was conducted to compare estimates in the two coin scenarios (a luck domain) with the estimates in the soccer scenario (a skill domain). Prior research on both sequential predictions and estimates of the randomness of sequences (see Oskarsson et al 2009 for a review) has documented negative recency in beliefs about random or luck-based outcomes (i.e., the gambler's fallacy, Tune 1964) and positive recency in beliefs about skill-based outcomes (i.e., the hot hand, Gilovich et al 1985). In the current context, a spontaneously applied negative recency belief should generally result in under-estimating the tails of the distribution (i.e. believing that 10 heads in a row are even less likely to occur than is true). Conversely, positive recency beliefs should result in higher estimates of the tails of the distribution, as players who are doing particularly well or badly will accumulate many or few goals.

While the participants' estimates were not made in sequence, and are therefore not a direct test of recency beliefs, both the coin flip and soccer scenarios yielded estimates that were more consistent with a positive recency belief. The lack of underestimation of the tails, and therefore of "runs" in the data, for coin flips is particularly surprising given the large literature on misperceptions of randomness. These results call into question the generality of beliefs in negative recency and use of the representativeness heuristic (Tversky and Kahneman 1971) when anticipating random outcomes.

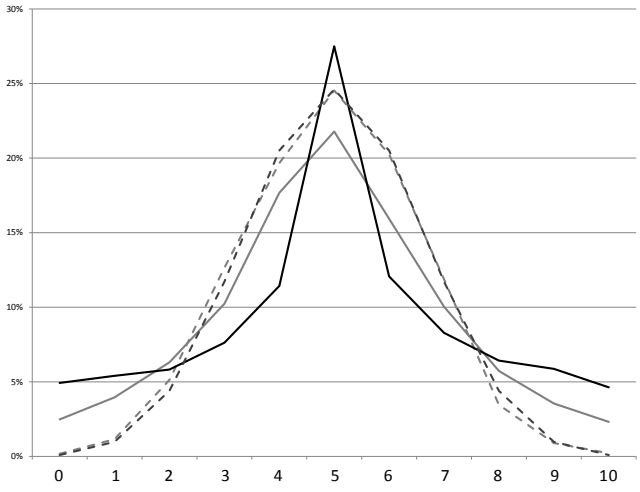
Estimated distribution of height as a control.

The results presented thus far suggest a persistent distortion in people's beliefs about the distributions of anticipated repeated events. It is not clear, however, whether the bias occurs due to a specific bias in compounding repeated probabilistic events, or a more general difficulty in reasoning about distributions in general. Some recent work (Goldstein and Rothschild 2014) has suggested that people's beliefs about experienced distributions of events can be highly accurate, particularly when elicited through a graphical interface.

To test this, the current study included a distribution estimation task based on commonly experienced knowledge and involving no statistical inference. Participants were asked to estimate the proportion of men in the United States aged 18 or older who are in one of 11 categories: either below 5'1", between 5'1" and 6'7" (asked in nine 2" incremented categories) or 6'7" and above. Based on a sample of 3981 male adults, these categories yield a discrete distribution that is quite similar (within 1 percentage point for all categories) to the binomial distribution with 11 outcomes used here.

A total of 539 participants estimated the height distribution (either using a histogram or estimating the amounts directly). The averages of their estimates are shown below in Figure 4 (gray line) and compared with the actual distribution of height (gray dashes), estimates of binomial processes (black line, from Figure 1) and the binomial distribution (gray line).

Figure 5: Height vs. Repeated Probability Estimates



Data for height is shown in gray, binomial processes in black. Estimates are solid lines and actual values are dashed lines.

A total of 539 participants estimated the height distribution (either using a histogram or estimating the amounts directly). The averages of their estimates are shown below in Figure 5 (gray line) and compared with the actual distribution of height (gray dashes), estimates of binomial processes (black line, from Figure 1) and the binomial distribution (gray line).

As can be seen in the figure, the estimates of height are much more similar to the actual distribution for height than are the estimates of binomial processes. Based on the 418 participants who made an estimate for both one of the binomial processes (scenarios 1-3) and the height distribution, the estimates of the height distribution were significantly more accurate (RMSE=19.8 vs. 29.6, t=9.80, p<.001). This was driven primarily by differences in the variance of the distributions (1.97 vs. 3.46, t=4.54, p<.001), although there was a direction difference in kurtosis as well (1.28 vs. 1.63, t=1.92, p=.055).

Some participants (N=282) were instead asked to choose among six histograms. In comparison with the binomial processes, where only 25% of participants were able to identify the most accurate distribution, 43% of participants were able to identify the most accurate distribution of heights. Thus, across the elicitation methods used, people were substantially more accurate at estimating height than estimating the outcomes of repeated probabilistic events. This suggests that the biases in estimating probability distributions documented in this study cannot be solely attributed to the difficulty of responding to distributional questions. Rather, the data suggests a novel and systematic bias in reasoning about the distribution of outcomes that arises even from simple and intuitive events such as coin flips, soccer kicks and survey sampling.

Study 2

The participants in Study 1 constitute a novice population. Only 45% held a Bachelor's degree or higher, and only 5% considered themselves knowledgeable in statistics. Neither variable significantly moderated the accuracy of their results, although those who reported knowing more statistics were directionally more accurate. However, this raises a question about the generality of the findings.

It is important to note that expertise in statistics is not trivial to develop. People invest significant monetary and time resources in order to learn how to conduct statistical inference, and such education typically involves both instruction and practice in working with probability distributions. Would people who have more expertise show less bias (or even no bias) in their estimates?

Method

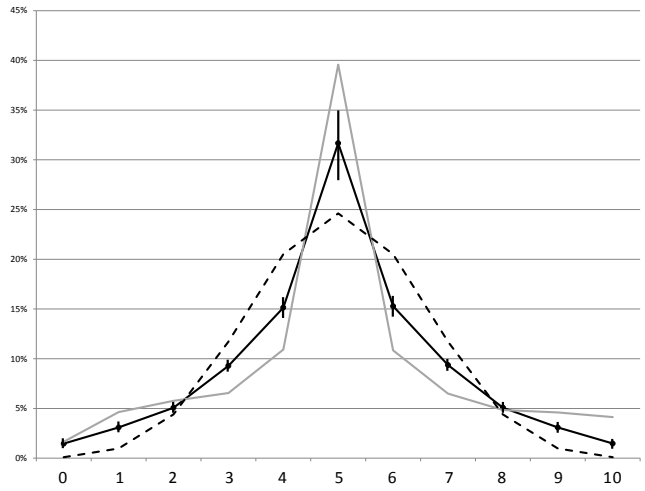
Eighty seven MBA students enrolled in a graduate Marketing Research course, all of whom had completed at least one statistics course, participated in an in-class exercise prior to a discussion of survey sampling. The students completed a one-page pencil-and-paper version of the survey scenario from Study 1 with numeric elicitation. Unlike in Study 1, where participants received computerized feedback to ensure that their estimates summed to 100, 21% of the participants' estimates summed to a different total and were normalized to 100.

Results and Discussion

The MBA student estimates demonstrated significant bias. As shown in Figure 6, the average MBA estimates significantly diverged from the binomial distribution for all the outcomes. While the mean of the distribution was estimated accurately, the MBA students' estimated distribution had

higher variance (3.68 vs. 2.5, t=5.65, p<.001) and more kurtosis (.43 vs. -.20, t=2.89, p=.005) than the actual distribution.

Figure 6: Expert vs. Non-Expert Estimates



MBA students (black line) vs. Study 1 participants (gray line)
Error bars represent 95% confidence intervals.

Nevertheless, the MBA student estimates for most of the outcomes were significantly different from the estimates in Sample 1 made using the same scenario and elicitation procedure (N=61, shown as the gray line Figure 6), and closer to the actual distribution. Overall, the MBA students were significantly more accurate (RMSE=16.0 vs. 33.5, t=11.63, p<.001).

Study 3

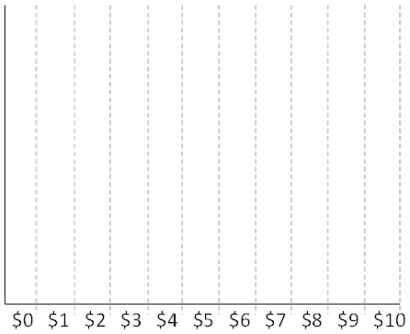
While the MBA students in Study 2 were more accurate than the novice population, they still demonstrated a significant bias in their estimates. Study 3 compared a population with even more expertise, PhDs and PhD students, with a novice population, undergraduate students from commuter colleges in Chicago.

Method

Ninety three attendees at the 2013 Society for Judgment and Decision Making conference (the "expert" sample) completed a paper-and-pencil survey in exchange for a large Toblerone candy bar. In addition, 127 students at a research lab in downtown Chicago (the "novice" sample) completed the same survey in exchange for $2.

In addition to unrelated items, the survey included a description of the probability estimation version of the coin flip scenario in Study 1. There were two elicitation conditions. In the elicitation condition (N=62 expert; N=64 novice), the survey had a pre-printed template for a histogram with an unlabeled y-axis, shown below. Participants were asked to shade in each bar, such that the height of the bar represented the probability of the associated outcome. The bars drawn by participants were measured in millimeters, and the values were rescaled to add up to 100.

Figure 7: Elicitation Task Stimuli Used in Study 3



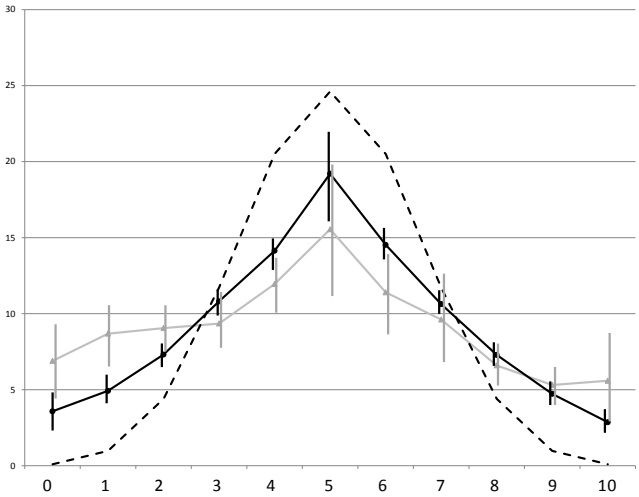
Not shown to scale.

In the choice condition (N=31 expert; N=63 novice), participants read the same scenario, and were shown the same six histograms as in Study 1, and were asked to choose which was the most accurate, second most accurate and least accurate.

Results and Discussion

Both the novice and expert samples provided distributions that diverged significantly from the correct distribution, as shown in Figure 8 below. Both the expert and novice distributions reflected the correct mean (4.96 and 4.73 vs. 5.0). However, the variance was higher for both the expert distribution (5.30 vs. 2.5, t=2.45, p=.017), and the novice distribution (5.82 vs. 2.5, t=2.13, p=.037).

Figure 8: Expert vs. Non-Expert Freehand Histograms



Expert (black line) vs. novice (gray line) samples
Error bars represent 95% confidence intervals.

Once again, the expert distribution, while significantly biased, was more accurate than the novice distribution (RMSE=17.0 vs. 33.3, t=5.18, p<.001). In part, this reflected the greater difficulty the novice sample had with the task. Even among the experts, differences in experience seemed to

matter, as faculty were marginally more accurate than non-faculty (RMSE=13.0 vs. 18.7, t=1.69, p=.096).

The freehand histogram was clearly difficult for participants, particularly the novices. We observed large standard errors and while 82% of experts provided a monotonic distribution, only 39% of novices did. A separate subset of participants were presented with a much easier task, simply selecting which of six graphs represented the correct distribution. The experts directionally outperformed the novices, choosing the correct graph directionally more often (42% vs. 26%, p=.163). However, even the experts chose the correct graph less than half of the time.

General Discussion

These findings demonstrate a robust bias in judgments of outcome distributions from repeated probabilistic events. The observed bias is moderated by but robust to differences in elicitation methods, scenario contexts and level of expertise. These findings have potential implications for several different aspects of decision research.

This bias, demonstrated in simple and realistic settings, poses a challenge for economic and psychological theories which presume that people are able to make near-optimal decisions, because of the ability to efficiently integrate information into accurate probabilistic beliefs. For example, Griffiths and Tenenbaum (2006) argue that everyday cognitive judgments follow optimal statistical principles, and that people have accurate distributional knowledge and then accurately infer conditional probabilities from prior beliefs. Such sophisticated inferences require constructing conceptual distributions from what is known, similar to what the experiments in this paper test explicitly. The difficulty that participants had in constructing the distribution of outcomes from repeated simple probabilistic events presents a reason for caution in assuming that people can do so efficiently.

Beliefs about prospective probabilistic events are an input into an important but understudied kind of decision. Prior work has contrasted decisions among options with explicit probabilistic information from those with probabilities learned from experience (Hertwig and Erev 2009). Rare events are overweighted in decisions when probabilities are explicitly known and underweighted when inferred from experience (Hertwig et al 2004). However, a third type of choice exists, such as betting on the outcome of a sports game. Here, people may have information about the probabilistic inputs into an outcome distribution (i.e likelihood of scoring), but still must prospectively infer the relevant probabilities (without experience) in order to use them in making choices.

The findings in this paper suggest that in these settings, misbeliefs about the resulting probability distribution may provide an additional cause of over-weighting small probabilities, precisely because the probabilities are not known. This is in contrast with popularly-accepted views that unlikely future events, particularly those arising from a confluence of factors (e.g. "black swans", Taleb 2010) are under-estimated. In fact, the task used in this paper is similar to prospective forecasting tools which elicit discrete distributions of beliefs about mutually exclusive events (e.g., Federal Reserve Bank inflation forecasts, Goldstein and Rothschild 2014). The findings suggest that these forecasts may instead over-estimate the likelihood of rare events, arising from low-probability independent joint events.

The tasks used in this paper may also be useful in future research on subjective probabilistic beliefs. Research on perceived probability tends to investigate isolated judgments, which measure absolute rather than relative probability. However, such judgments are prone to a series of biases, such as unpacking and subadditivity (Tversky and Koehler 1994) precisely because they are about isolated events. The distribution-elicitation approach explored here, in contrast, provides a test of subjective relative probability. This allows researchers to test whether external factors (e.g., wishful thinking, Krizan and Windschitl 2007) systematically distort the relative probability of one outcome versus another, and to quantify resulting differences in errors.

References

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3), 519.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1-14.

Gilovich, T., Vallone, R.& Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 295–314.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17(9), 767-773.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8), 534-539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12), 517-523.

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological bulletin*, 133(1), 95.

Oskarsson, A. T., Van Boven, L., McClelland, G. H. & Hastie., R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin*, 135(2), 262–285.

Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable Fragility*. Random House Digital, Inc..

Tune., G. S. (1964). Response preferences: A review of some relevant literature. *Psychology Bulletin*, 61, 286–302.

Tversky A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105-110.

Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, 101(4), 547.