Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Choosing the Devil You Don't Know: Evidence for Limited Sensitivity to Sample Size-Based Uncertainty When It Offers an Advantage

Florian L. Kutzner, Daniel Read, Neil Stewart, Gordon Brown

To cite this article:
Florian L. Kutzner, Daniel Read, Neil Stewart, Gordon Brown (2017) Choosing the Devil You Don't Know: Evidence for Limited Sensitivity to Sample Size-Based Uncertainty When It Offers an Advantage. Management Science 63(5):1519-1528. https://doi.org/10.1287/mnsc.2015.2394

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 5, May 2017, pp. 1519–1528
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Choosing the Devil You Don't Know: Evidence for Limited Sensitivity to Sample Size–Based Uncertainty When It Offers an Advantage

Florian L. Kutzner,[a,b] Daniel Read,[b] Neil Stewart,[c] Gordon Brown[c]

[a] Department of Psychology, Heidelberg University, 69117 Heidelberg, Germany; [b] Behavioural Science Group, Warwick Business School, Coventry CV4 7AL, United Kingdom; [c] Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom

Contact: florian.kutzner@psychologie.uni-heidelberg.de (FLK); daniel.read@wbs.ac.uk (DR); neil.stewart@warwick.ac.uk (NS); g.d.a.brown@warwick.ac.uk (GB)

Received: August 24, 2013
Revised: January 26, 2015; July 24, 2015
Accepted: October 17, 2015
Published Online in Articles in Advance: April 22, 2016

https://doi.org/10.1287/mnsc.2015.2394

Abstract. Many decision makers seek to optimize choices between uncertain options such as strategies, employees, or products. When performance targets must be met, attending to observed past performance is not enough to optimize choices—option uncertainty must also be considered. For example, for *stretch* targets that exceed observed performance, more uncertain options are often better bets. A significant determinant of option uncertainty is sample size: for a given option, the smaller the sample of information we have about it, the greater the uncertainty. In two studies, choices were made between pairs of uncertain options with the goal of exceeding a specified performance target. Information about the options differed in the size of the sample drawn from them, *sample size*, and the observed *performance* of those samples, the proportion of successes or "hits" in the sample. We found people to be sensitive to sample size–based uncertainty only when differences in observed performance were negligible. We conclude that in the presence of performance targets, people largely fail to capitalize on the value advantages of small samples in the presence of stretch targets.

History: Accepted by Yuval Rottenstreich, judgment and decision making.

Funding: This research was supported by the Deutsche Forschungs Gemeinschaft [Grants KU3059/1 and 2], the Research Economic and Social Research Council [Grants ES/K002201/1 and ES/K004948/1], and the Leverhulme Trust [Grant RP2012-V-022].

Keywords: optimal foraging theory • small sample advantage • Bayesian rationality • bounded rationality • less-is-more • sampling approach • convexity • expected utility

Introduction

Imagine a bettor deciding which of two horses to bet on in the next race, a major event attracting the top stables and riders. Both horses are long shots. The bettor knows that Derring Do has come as high as third in his last 20 races but has never come first. So Derring Do is certainly above average but is probably not a race winner. For Dark Knocks, the bettor only knows about one race, where the horse came sixth. It is likely that Derring Do will outpace Dark Knocks. But is Derring Do the best bet, when the "best bet" does not mean choosing the one who will likely outperform the other, but rather the one more likely to do outstandingly well and come first in a highly competitive race? Because there is so little information about Dark Knocks, it remains possible that Dark Knocks is actually a race-winning horse that had just one not-so-great race. Imagine, as well, an analogous hiring decision. Two applicants are interviewing for a position. Both have graduated with honors from a top institution. Candidate A has been in the workplace for some time and has proven to be a sterling employee, performing significantly better than the typical honors graduate from that institution; candidate B is a fresh graduate. Yet the firm in question is looking for the very best, meaning employees who will be in the top 1%. Who should the firm hire if it must choose between these two? The interviewers have lots of information about candidate A and so can be sure she is above average but not absolutely outstanding. Candidate B is an unknown. On average, candidate A will be better, but candidate B is the only candidate who could be a top 1% performer.

These decisions share two key features. First, the target for acceptable performance is a *stretch target*, meaning it exceeds the average or expected level for both options. And, second, the decision maker has *more information* about the performance of one of the options. In both cases it is likely the low-information

option is the one to go with, *because* of and not in spite of that lack of information. The bettor and the employer can be all but certain that the high-information option (Derring Do and candidate A) will not meet the stretch target, but they can reasonably hope the low-information one might.

When seeking merely acceptable performers the tables are turned. If the observed performance of both options is above a target, that is when dealing with a *below-average* target, then greater uncertainty increases chances to fail. In the hiring example, if an average employee is needed, then the well-established employee is almost certain to meet the bill. The untested employee could still prove a disaster. Additionally, for both stretch and below-average targets, option uncertainty can compensate for lower observed performance. For stretch targets, lower observed performance can be compensated for by greater uncertainty, as we suggested for Dark Knocks, whereas for below-average targets, it can be compensated for by lower uncertainty. Appendix A provides an illustration and computational details.

In this paper we investigate peoples' sensitivity to amount of information when making choices under uncertainty in the presence of performance targets. We ask whether, for options having comparable sample performance, people will favor high-information options when facing below-average targets and low-information options when facing stretch targets. We also ask whether variations in uncertainty can be traded off against lower sample performance. We investigate settings in which people receive large or small samples about pairs of options. Each option is an outcome-generating process (analogous to horses or job candidates), and respondents must choose one option to attempt to reach or exceed a performance target. We find that people are sensitive to sample size–based uncertainty, but only when the options do not differ in observed sample performance.

Background

Our analysis is conditioned on the presence of a target for performance. The importance of targets, under names such as criterion values, aspiration levels, goals, or reference points, is widely recognized by researchers in the social and behavioral sciences (Markowitz 1952, Fishburn 1977, Heath et al. 1999, Kahneman and Tversky 1979, Lopes 1981, Payne et al. 1980, Simon 1955). Daniel Bernoulli (Bernoulli 1954, p. 25) used the concept of a target to qualify his assumption that utility was a nonlinear function of wealth, when he proposed that "a rich prisoner who possesses two thousand ducats but needs two thousand ducats more to repurchase his freedom will place a higher value on a gain of two thousand ducats than does another man who has less money than he." Even more, this prisoner will assign essentially zero utility to anything less than 2,000 ducats and relatively little additional utility to anything more. As in this most clear-cut case, targets describe a binary functional relation between the level on a performance dimension such as money and the value of that level.

The idea that in the presence of a stretch target an option's value might be increasing in uncertainty has been considered in many domains. In hiring, uncertainty about an applicant's true performance has been associated with risk premia (Lazear 1995). Young workers might be appealing because there is still little information about performance and so they have a chance of being exceptional performers. Similarly, in human mate selection, a lack of mutual knowledge seems responsible for inflated impressions of mutual attractiveness (Norton et al. 2007). If finding an ideal mate represents a stretch target, when there is little information there is a chance that "this is the one," whereas with more knowledge usually comes the certainty that he or she is not. Finally, in foraging, there is evidence that animals prefer more uncertain options when less uncertain options are unlikely to cover their daily energy targets (Kacelnik and Bateson 1996).

Evidence suggests that choices are at least partially sensitive to these implications of uncertainty. Rode et al. (1999) conducted a study in the context of balls and urns. Respondents chose whether to draw from a "risky" urn having a known number of black and white balls or a maximally "uncertain urn" with no information about its composition. When striving for a stretch target, people preferred to draw from the uncertain urn, but when the target was average or below average, they preferred the risky urn. For example, in one condition the expected proportion was 50% black balls in both urns, and the target was 6 or 7 blacks in 10 draws. Nearly 60% of participants chose the uncertain urn. With a below-average target (3 or 4 blacks from 10 draws), 95% preferred the known option.

Heath et al. (1999) investigated a more concrete context. They gave participants either a stretch target for cost reduction (save \$250,000) or asked them to save "as much as possible." Respondents chose between a cost reduction plan offering a moderate sure result, saving \$80,000 for sure, or a plan that offered a higher but more risky result and a lower average, 20% chance of \$250,000, and \$50,000 otherwise. Only 24% of participants chose the risky plan, trying to save as much as possible. This increased to 47% when they had the stretch target. Hence, there is evidence that people are at least partially sensitive to the value of uncertainty when reaching for stretch targets.

In this paper we seek to generalize this evidence in the context of an additional source of uncertainty, the amount of sample information available about the

options. If we hold the sampling method constant, larger samples produce more certainty about the properties of a population. This principle was formulated as the law of large numbers by Jakob Bernoulli. In 1713, Bernoulli also made the psychological claim that "even the stupidest man [understands the law of large numbers], by some instinct of nature *per se* and by no previous instruction" (as translated by Sung 1966, p. 23, italics in original). Evidence supports his view. For example, when judging group differences, confidence and extremity of judgments increase with sample size (Irwin et al. 1956, Obrecht and Chesney 2013). And, supporting Bernoulli's nativism, this tendency emerges early in development (Jacobs and Narloch 2001, Masnick and Morris 2008).

Yet when presented in combination with sample proportions, people underappreciate the greater precision of large samples. Griffin and Tversky (1992) illustrated this in a number of studies showing that what they called evidence "strength," i.e., the sample proportion, was given much more importance in evaluating the truth of hypotheses than evidence "weight," i.e., sample size. For example, they asked participants to report their confidence that a coin was biased at a ratio of 3 to 2. Confidence increased strongly when a skewed sample proportion changed the posterior probability in favor of the bias but only weakly when the same change in posterior probability was brought about by a change in sample size (see also Antoniou et al. 2014, Obrecht et al. 2007). These and other studies on making inferences from samples of different sizes suggest that although sample size is not completely neglected, it is significantly underweighted (see also Bar-Hillel 1979, Evans and Dusoir 1977, Sedlmeier and Gigerenzer 1997, Peterson et al. 1968).

Combining these strands of research showing that people often choose uncertain options when reaching for stretch targets, and that they underappreciate the importance of sample size relative to sample proportions, we can derive expectations about how sample size–based uncertainty will affect choices. First, since sample size is an important determinant of option uncertainty, we predict that people will, at least under some circumstances, use sample size information. But at the same time, studies have shown that sample size is given little weight when it is pit against sample proportion. Consequently, we expect differences in sample proportions to be an important moderator, with the influence of sample size on choices being greatest when proportions are relatively undiagnostic.

Overview of Experiments

In our paradigm, choice options are presented as "wheels of fortune," each with one blue and one red segment. The premise is that one wheel will be spun 100 times, and a target of N blue outcomes must be reached to win a payment. Participants make many such choices and are paid only if the wheel they choose reaches the target. Prior to each set of 100 spins, the participant is provided with a sample of spins from both wheels. The samples vary in size and performance—here, the proportion of blue outcomes.

In Study 1, we construct the samples to provide a powerful test of sensitivity to sample size–based uncertainty. We test whether people prefer small sample options for stretch targets when observed proportions are equal but large sample options for below-average targets. We also test whether people trade off differences in observed proportions with sample size–based uncertainty and, if so, by how much. In Study 2, we randomly generate samples and use two different target levels to move toward a more representative design (Brunswik 1955).

Study 1: Factorial Design

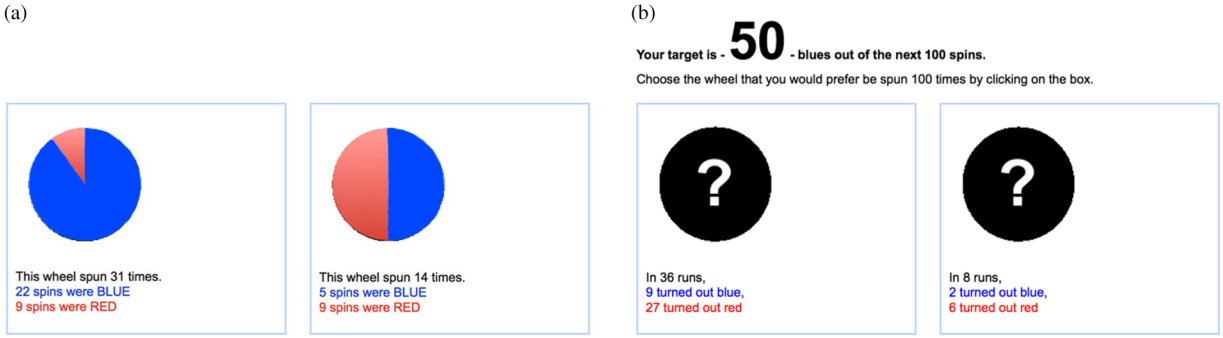
In Study 1, we employed a factorial design in which the probability of reaching the target and the sample proportions of blue outcomes were varied orthogonally and the target was to reach at least 50 blue results in 100 additional spins.

Methodology

Study 1 was conducted in a laboratory using computers and a specially designed web browser–based survey. The participants were 73 students from the University of Warwick (44% female, M_{age} = 21.93, SD_{age} = 2.72) who were paid a £5 show-up fee plus £0.25 each time they chose a wheel that reached the target. Up to 22 participants worked simultaneously on the experiment, each seated in an individual cubicle.

The generating mechanisms were spinning wheels divided into "blue" and "red" segments; the probability of success on each trial was represented by the size of the blue segment. All possible wheels with segments ranging from 0% to 100% blue were graphically presented on one screen, and participants were asked to assume that each possible wheel was equally likely to be chosen for the upcoming task. They were then shown an example of a task in which the wheels were revealed, along with the sample information. The sample information was given as the number of spins and the number of times the wheel came up blue and red (see Figure 1(a)). In this example task, the segment area and the sample drawn from the wheel were chosen to highlight how the sample was not identical to the underlying population: the wheel having a blue segment covering 50% had come up blue in 5 out of 14 spins. The next task was a practice task, identical to the actual choice tasks, in which the wheels were blanked out, and only the sample information was provided (see Figure 1(b)). Participants were told they would win

Figure 1. (Color online) Example of a Task Presented to Illustrate the Spinners (a) and One Actual Choice Task (b)



Notes. In the example of a task, wheels were revealed. In the actual choice tasks, wheels were covered and only the samples were visible.

money if and when the wheel they chose reached a target of at least 50 blues in the next 100 spins.

All participants faced the same nine tasks in random order. In each task a small sample was provided from one wheel and a large sample from the other. The tasks are summarized in Table 1, which classifies tasks according to two differences between the large and small sample option. The first classification is the sample proportions of blue outcomes, which we denote as $\Delta Prop(S)$ for the *difference in sample proportion of successes*. The second classification is in terms of probabilities of reaching the target of $t = 50$ in 100 spins, $\Delta p(t)$, or the *difference in the probability of reaching the target*. Positive values of both $\Delta Prop(S)$ and $\Delta p(t)$ indicate an advantage for the small sample spinner. We varied $\Delta Prop(S)$ and $\Delta p(t)$ orthogonally, producing a 3 ($\Delta Prop(S)$: 10%, 0.0%, or −10%) × 3 ($\Delta p(t)$: 0.1, 0.0, or −0.1) design.

The three choices in the middle columns of Table 1 involve equal observed proportions, $\Delta Prop(S) = 0\%$, for stretch, average, and below-average targets. For the other choices, trade-offs occur between differences in observed proportions and uncertainty. For stretch targets and $\Delta Prop(S) = -10\%$, higher uncertainty compensates for lower observed proportions for the most extreme negative deviation from the target. Conversely, for below-average targets and $\Delta Prop(S) = 10\%$, lower uncertainty compensates for lower observed proportions for the most extreme positive deviation from the target.

Results

Figure 2 displays the proportion of respondents choosing the small sample option in each condition. For $\Delta Prop(S) = 0\%$, there was a tendency to choose the option with the higher $p(t)$. For the stretch target, most (63%) participants chose the small sample spinner; for the below-average target, most (77%) chose the large sample spinner (i.e., 23% chose the small sample spinner). For the average target and when both options had the same $p(t)$, only a minority chose the small sample option (27%). No effect of $\Delta p(t)$ was evident when $\Delta Prop(S)$ was either +10% or −10%. In these cases, participants always favored the option with the higher sample proportion of successes.

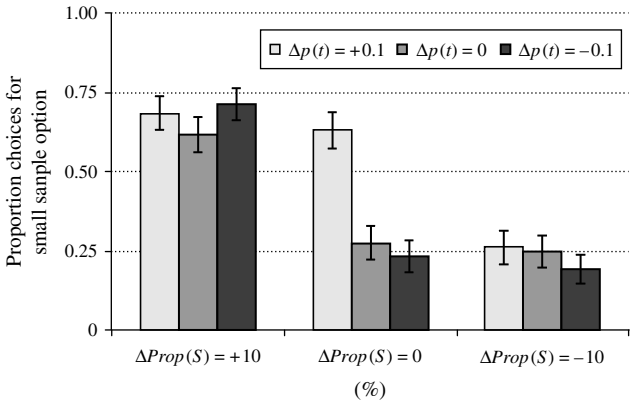
Formal analyses confirmed these findings. Choices for the small sample option were regressed onto the z-standardized scores of $\Delta Prop(S)$ and $\Delta p(t)$ and their interaction using a logistic mixed effects model with correlated random intercepts and slopes for every

Table 1. Design of Study 1

		$\Delta Prop(S)$ = Proportion (S | Small) − Proportion (S | Large)								
		10%			0%			−10%		
$\Delta p(t) = p(t$ | Small$) - p(t$ | Large$)$:		0.1	0	−0.1	0.1	0	−0.1	0.1	0	−0.1
Target type:		Below	Below	Below	Stretch	Average	Below	Stretch	Stretch	Stretch
Small sample	No. of S	4	4	3	2	4	6	1	2	2
	No. of F	2	2	1	6	4	2	3	6	4
	$Prop(S)$	67%	67%	75%	25%	50%	75%	25%	25%	33%
	$p(t)$	0.77	0.77	0.81	0.1	0.49	0.9	0.19	0.1	0.23
Large sample	No. of S	17	30	21	9	18	27	12	10	15
	No. of F	14	23	12	27	18	9	21	18	18
	$Prop(S)$	55%	57%	64%	25%	50%	75%	36%	36%	45%
	$p(t)$	0.67	0.77	0.91	0	0.48	1	0.08	0.09	0.31

Note. S, successes; F, fails.

Figure 2. Proportion of Choices for the Small Sample Option as a Function of the Differences in Observed Proportions of Successes, $\Delta Prop(S)$, and the Differences in the Probabilities of Reaching the Target, $\Delta p(t)$



Notes. Values of $\Delta p(t)$ and $\Delta Prop(S)$ greater than zero indicate an advantage for the small sample option. Error bars = indicate standard errors of the mean.

participant in the lme4 package for the R statistical software. Choices were sensitive to $\Delta p(t)$ ($\beta = 0.36$, $z = 2.90$, $p < 0.01$) and $\Delta Prop(S)$ ($\beta = 1.23$, $z = 5.87$, $p < 0.001$). The interaction was not significant ($z = -1.07$, $p = 0.28$). Directly testing whether $\Delta p(t)$ was relevant only for $\Delta Prop(S) = 0\%$, we contrasted the effect of $\Delta p(t)$ for choices where $\Delta Prop(S)$ was substantial (i.e., +10% or −10%, coded 0) with those where $\Delta Prop(S)$ was zero (coded 1). The regression analysis revealed the expected interaction ($\beta = 0.84$, $z = 3.76$, $p < 0.001$), indicating that the influence of $\Delta p(t)$ was larger in the absence of a difference in $\Delta Prop(S)$ ($\beta = 1.64$, $z = 3.85$, $p < 0.001$) than in the presence of a difference in $\Delta Prop(S)$ of either +0.10 or −0.10 ($\beta = -0.03$, $z = -0.34$, $p = 0.73$).

Discussion

These results support the hypotheses that choices are sensitive to the value created by sample size–based uncertainty under some conditions. We found that only when options did not differ in sample performance, i.e., the proportion of hits, did respondents prefer the small sample option for stretch targets and the large sample option for below-average targets. When the options differed in sample performance, here by 10%, choices favored the option with the higher observed performance, regardless of sample size. Thus, choices did not reflect trade-offs between the value of sample size–based uncertainty and differences in sample performance. Sample size seems to be used as a tie-breaker when sample performance cannot distinguish between options.

One might ask how far these conclusions depend on the assumed priors for the spinners. The $\Delta p(t)$ values in Table 1, on which we based our the factorial design, are derived assuming a uniform distribution over all possible spinners, or a $\text{Beta}_{(1,1)}$-distribution. This is what we instructed our participants to assume. If participants' priors deviate a great deal from this assumption, this could change the sign of $\Delta p(t)$. We conducted simulations to determine the range of priors that (a) maintain the order of $\Delta p(t)$'s for each level of $\Delta Prop(S)$ and (b) maintain $\Delta p(t)$'s larger (smaller) than zero for tasks where $\Delta p(t)$ is supposed to be +0.1 (−0.1). We varied the a and b parameters of the prior $\text{Beta}_{(a,b)}$-distribution orthogonally from 0.001 to 150. The results are illustrated in Figure 3, which gives a simplified summary. The range of admissible prior distributions given our design is highlighted in dark grey shading. As can be seen, over a wide range of possible prior distributions, the predictions of Study 1 remain unchanged.

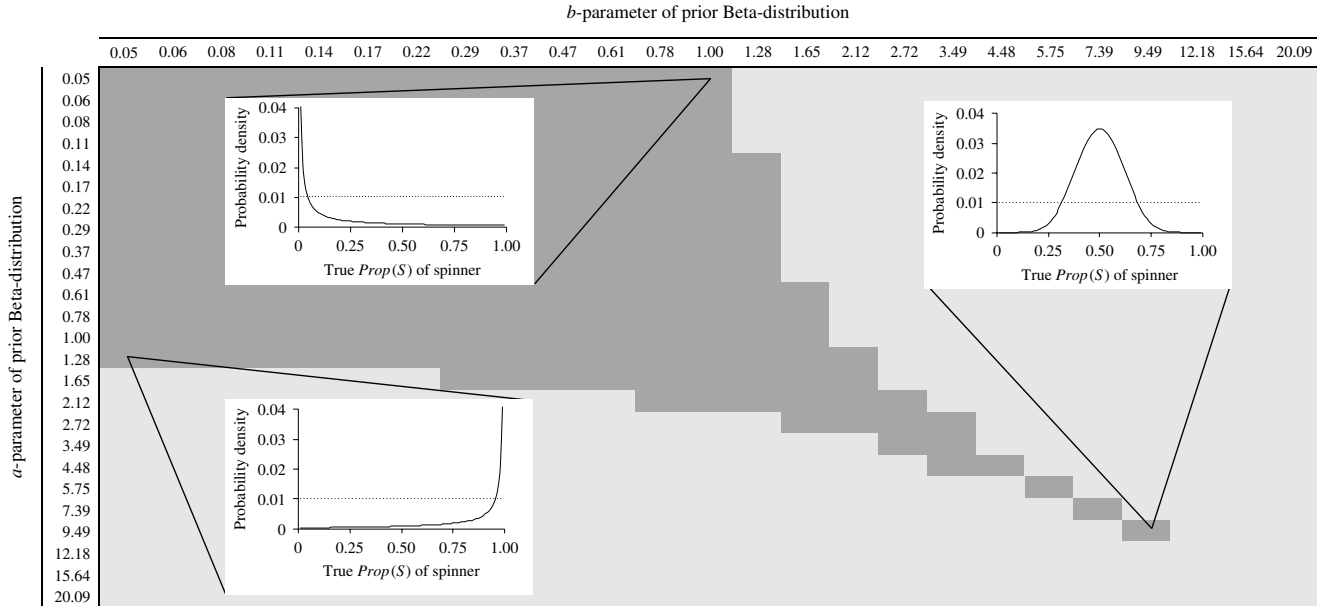
Furthermore, the choice pattern cannot be readily explained by assuming prior beliefs outside the admissible range. Calculating $\Delta p(t)$ values assuming that spinners with extremely small (or large) success segments were highly likely results in the prediction that the large (or small) sample options would have always had the higher chance of reaching the target. Contrary to this, we do not observe an overall preference for large or small sample options. Calculating $\Delta p(t)$ values assuming highly peaked prior distributions, $\Delta p(t)$ values favor the large sample options for below-average targets, i.e., tasks with $\Delta Prop(S) = 10\%$, and the small sample options for stretch targets, i.e., tasks with $Prop(S) = -10\%$. This is opposite to the observed choice pattern. For $\Delta Prop(S) = 0\%$, the sign and order of $\Delta p(t)$ values remain unchanged. Thus, $\Delta p(t)$ would still account for this choice pattern. Importantly, concluding that choices are sensitive to sample size–based uncertainty for $\Delta Prop(S) = 0\%$ and insensitive for $\Delta Prop(S) \neq 0\%$ is valid for a wide range of prior beliefs about the likelihood of possible spinners, and prior beliefs outside this range seem unlikely.

One limitation of Study 1 is that choice patterns might have been due to the particular set of nine tasks. In particular, it remains unclear whether differences in observed proportions smaller than 10% would elicit trade-offs with sample size–based uncertainty. Additionally, we cannot generalize beyond the specific target level of 50 successes in 100 spins, which is arguably special in that it represents the notion of chance. In Study 2 we generalized and replicated our results by randomly generating choice tasks and adding a more extreme target (80 successes in 100 spins).

Study 2: Random Generation of Choice Tasks

In Study 2, we investigated the sensitivity to sample size–based uncertainty across a more varied set of

Figure 3. Range of a and b Parameters of a $\text{Beta}_{(a,b)}$-Distribution



Notes. Dark grey shading indicate prior distributions that maintain the order and sign of $\Delta p(t)$ values compared with the factorial design; cf. Table 1. Plots illustrate the most extreme admissible right-skewed ($\text{Beta}_{(0.05,1)}$), left-skewed ($\text{Beta}_{(1.28,0.05)}$), and peaked ($\text{Beta}_{(9.49,9.49)}$) prior density distributions. Dotted lines indicate assume flat prior distribution.

choice tasks. For every task, we randomly generated a large and a small sample option and randomly chose the target to be either 50 or 80 successes out of 100 spins of the wheel.

Methodology

Study 2 was conducted in the same laboratory as Study 1 using similar methods. One hundred seventeen University of Warwick students (65 female, M_{age} = 21.67, SD_{age} = 3.91) were recruited based on a £5 show-up fee plus £0.25 each time an option they chose met or exceeded a target. Instructions were identical to those in Study 1.

To generate the stimuli for each trial, we randomly drew four random numbers: (1) a number between 1 and 10 that represented the blue outcomes in sample 1, (2) a number between 1 and 10 that represented the red outcomes in sample 1, and (3) and (4) different numbers between 1 and 50 that represented the number of blue and red outcomes in sample 2. Typically, sample 1 (ranging from 2 to 20 observations) was the small sample, and sample 2 (ranging from 2 to 50) was the large one, although only sample size and composition were entered into the analysis. The resulting proportions lay between 5% and 95% for the small samples and between 2% and 98% for the large samples. For each task we also randomly chose a target of either 50 blue or 80 blue outcomes.

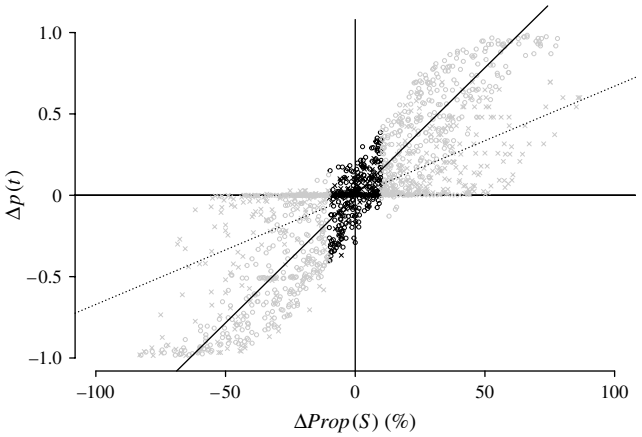
All participants made 13 choices. The first 12 involved choices randomly generated as just described. The final choice ensured enough responses to replicate the finding from Study 1 that for equal sample proportions people preferred the small sample when striving for a stretch target. For this 13th choice, the small sample spinner had produced 2 blue and 2 red outcomes and the large sample spinner had produced 34 blue and 34 red outcomes (i.e., $\Delta Prop(S) = 0$). With a target of 80, this creates 6% advantage for the small sample option (i.e., $\Delta p(t) = 0.06$). After all choices, participants were given feedback and informed about how much money they had won.

Characteristics of Randomly Generated Choice Tasks

For the 1,404 observed choices (117 participants × 12 choices), the average sample sizes were 10.75 (SD = 4.04) for the small and 51.12 (SD = 20.22) for the large sample spinners. The average absolute difference between sample proportions, $|\Delta Prop(S)|$, was 25% (SD = 19). The average $|\Delta p(t)|$ was 0.30 (SD = 0.30) combining both targets, 0.43 (SD = 0.30) for the target of 50, and 0.16 (SD = 0.24) for the target of 80. The average likelihood of reaching the target if the option with the higher $p(t)$ was chosen on each trial was 0.71 when the target was 50 and 0.17 when it was 80.

As visible from the scatterplot in Figure 4, $\Delta Prop(S)$ and $\Delta p(t)$ were strongly and positively correlated across choice tasks. A regression analysis confirmed this relationship ($\beta = 3.01$, $t(1{,}400) = 44.28$, $p < 0.001$) and showed it to be moderated by the target level ($\beta = -0.03$, $t(1{,}400) = -28.66$, $p < 0.001$). The relationship was weaker for the target of 80 ($\beta = 0.67$, $t(682) = 27.88$,

Figure 4. Scatterplot and Regression Lines of the 1,404 Choice Tasks Used in the Representative Design of Study 2 Relating $\Delta Prop(S)$ and $\Delta p(t)$



Notes. Circles and the solid line indicate tasks with the target of 50 blue outcomes. Crosses and the dotted line indicate tasks with the target of 80 blue outcomes. Dark symbols refer to tasks with a $\Delta Prop(S)$ less than 10%.

$p < 0.001$) than for the target of 50 ($\beta = 1.57$, $t(718) = 77.48$, $p < 0.001$).

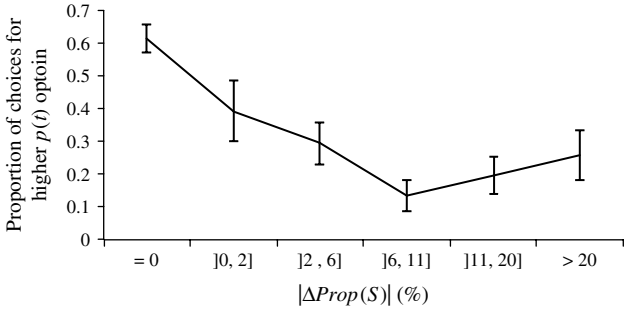
Despite the strong relationship between $\Delta Prop(S)$ and $\Delta p(t)$, in 15% of trials the option with the higher $p(t)$ did not have the higher $Prop(S)$. These critical trials, where sample size information should be decisive, are located in the top left and bottom right quadrants of the scatterplot. When the target was 80, 22% of cases were critical; when it was 50, 9% were. If we focus only on cases where $|\Delta Prop(S)|$ is less than 10%, these critical cases are even more prevalent, making up 34% overall and 42% of cases involving the target of 80.

Experimental Results

We regressed the 12 choices made within the representative design on $\Delta Prop(S)$, $\Delta p(t)$, and the target level t. Choices were coded 1 for choices of the small sample option and 0 otherwise. We conducted a logistic mixed effects regression analysis with correlated random intercepts and slopes for every participant using the lme4 package for the R statistical software. We used z-standardized scores and included all interactions.

As in Study 1, choices were sensitive to $\Delta p(t)$ ($\beta = 1.89$, $z = 2.39$, $p = 0.017$), over and above sensitivity to $\Delta Prop(S)$ ($\beta = 4.52$, $z = 7.15$, $p < 0.001$). An unexpected interaction between $\Delta Prop(S)$ and the target level ($\beta = 1.58$, $z = 2.54$, $p = 0.011$) indicated that $\Delta Prop(S)$ predicted choices better when the target was 80 rather than 50. No other effect was significant; $|z| < 0.83$. Analyzing the 13th choice ($\Delta Prop(S) = 0\%$, target = 80) also revealed sensitivity to sample size–based uncertainty with 60.7% of participants choosing the small sample option ($\chi^2(1) = 5.34$, $p = 0.002$).

Figure 5. Proportion of Choices, of the 333 Critical Choice Tasks, That Favored the Higher $p(t)$ Option as a Function of the Absolute Difference in Observed Proportions, $\Delta Prop(S)$



Note. Error bars represent the standard error of the mean.

We next explored whether the size of differences in observed proportions moderated the impact of sample size–based uncertainty. To this end, we only analyzed choices for trials on which sample size was decisive either because the difference in observed proportions pointed toward the option with the lower likelihood of reaching the target or because the difference was zero. This analysis included 216 tasks from the representative design and 117 from the 13th task. As shown in Figure 5, the proportion of choices in line with $\Delta p(t)$ rather than $\Delta Prop(S)$ decreased as the size of the absolute difference in $\Delta Prop(S)$ increased. In fact, in these critical trials, choices tended to follow $\Delta p(t)$ only when $\Delta Prop(S)$ was zero.

Discussion

These results generalize those of Study 1 to a random selection of choice tasks. Again, choices tended to follow differences in sample proportions while neglecting sample size. Only when the difference in sample proportions was zero were choices sensitive to sample size–based uncertainty and tended toward the options with the higher probability of reaching the target. It appears that trading off even small differences in observed performance with the value of uncertainty is a hard task to master.

Conclusion

In two experiments, we investigated whether people maximize their chances of reaching performance targets by integrating option uncertainty—here, sample size—with sample performance—here, the proportion of positive outcomes in that sample. Our evidence suggests sensitivity to the value of uncertainty only when the differences in sample performance are virtually zero. When those differences are zero, people responded to sample size appropriately: for below-average targets they preferred the large sample option, whereas for stretch targets they preferred the small sample option.

Why are people insensitive to sample sized–based uncertainty, even though they recognize that sample size matters? We tentatively suggest two explanations. First, sample size information is logically subservient to performance information. Knowing that a sample of size N has been obtained from a population is not useful at all unless more is specified about the performance of that sample. On the other hand, knowing that a sample revealed a performance p is useful even if the sample size is unknown. As a consequence, the interpretation of sample performance might take priority over that of sample size that, indeed, may never be used. Second, forgoing a better option for a worse but more uncertain one might be hard to justify. Having chosen an employee *because* we did not know much about her is a difficult argument to make, especially if the employee underperforms. By contrast, if we choose the candidate who looks better "on paper," it is hard to argue that the wrong choice procedure was followed. Without the reasoning provided in this paper, relative uncertainty might remain an intangible basis for choice, which cannot easily be pointed to. It is no accident that we rarely hear people say, "better the devil you *don't* know."

Griffin and Tversky (1992) reached a similar conclusion, studying the advantages large samples should have on making judgments more precise. They showed that what they call evidence strength, which for us would be the sample proportion, was more important than evidence weight, which for us would be sample size. We show that people also underestimate the advantages of small samples, when small samples have a value advantage. The overall message appears to be that the implications of sample size are generally underappreciated and that people focus primarily on the central tendency of samples as a guide to judgment and decision making, and they will use sample size information as a tie-breaker.

Our studies add to the literature on ambiguity aversion. As often portrayed, when faced with a choice between an option having a known probability and an option having a completely unknown or ambiguous probability, people prefer to choose the known option (Ellsberg 1961). This might seem like a bias, since in the settings that were discussed by Ellsberg and that have been the focus of much subsequent research, ambiguous and risky options have the same expected outcome. In line with earlier work (Rode et al. 1999), we show that to achieve a target, ambiguous options are sometimes objectively better than unambiguous ones for stretch targets and sometimes worse for below-average targets. This might explain documented reversals of ambiguity aversion for low probability gains and high probability losses, both situations arguably involving stretch targets (Curley and Yates 1985, 1989; Hogarth and Einhorn 1990; Kahn and Sarin 1988).

Furthermore, the present research on the role of small samples in reaching absolute performance targets complements research on the role of small samples for reaching "relative" performance targets. Faced with a choice between two uncertain options, existing performance differences are systematically inflated in small samples (Kareev 2000). If decision makers have a stretch target for performance differences before making a choice, small samples increase choice quality (Cahan 2010, Fiedler and Kareev 2006). Thus, not only for absolute but also for relative performance targets can less be more, but only when the odds of reaching the target are against the decision maker—when they are not, less is less.

Finally, the two-sided finding that value generated by uncertainty only affects choices under very narrow conditions seems to resonate with organizational behavior and managerial decision making. Pointing toward sensitivity, sample size–based uncertainty seems to play a role in the considerations of individual team members (Kareev and Avrahami 2007). If competitive bonuses are based on large samples of their performance history, motivation to improve will be limited for both, those routinely above average and those routinely below. When based on more uncertain small samples, effort is necessary on every task and all the time, increasing overall performance. Similarly, in their classic review of managerial risk taking, March and Shapira (1987) report that "most managers seem to feel that risk taking is more warranted when faced with failure to meet targets than when targets were secure. In 'bad' situations risks would be taken" (p. 1409). Experiments with managers support this claim (Laughhunn et al. 1980) as do measures of firm performance. Firms below the industry's median show higher variability on returns on investment, presumably reflecting the adoption of more uncertain strategies when below the stretch target of the industry benchmark (Fiegenbaum and Thomas 1988). Yet only sometimes do young workers receive risk premiums when it seems they should (Bollinger and Hotchkiss 2003, Burgess et al. 1998, Hendricks et al. 2003, Lazear 1995).

Our evidence is limited to a paradigm with binary outcomes and a gambling task. Although we expect our findings to be robust, this should be investigated in other settings. It is also important to know whether feedback or statistical literacy improve performance. In our paradigm, participants did not receive feedback about their performance until the end and were thus prevented from optimizing their behavior. Because differences in probabilities of exceeding the target are in the single digits, extensive training might be necessary to optimize choices. Also, in our studies we did not include measures of statistical literacy, which might

influence acknowledgement of sample size implications. In our studies, proxies for statistical literacy, such as the highest degree or the type of degree, did not enter significantly into any of our analyses.

To conclude, it has become a common observation that important decisions involve uncertainty about their outcomes. The result might be a focus on the observed performance of one's options. To use the words of Dixit and Nalebuff (1991, p. 169), "Even though you can't guess right all the time, you can at least recognize the odds." Here, we suggest that an exclusive focus on *observed* odds can be detrimental. Going beyond notions of more-is-better and less-is-more, we show that the appropriate answer depends on where we stand. Knowing more about an option adds value when the odds are favorable. When the odds are against us, we are often better off with the devil we don't know than the devil we do.

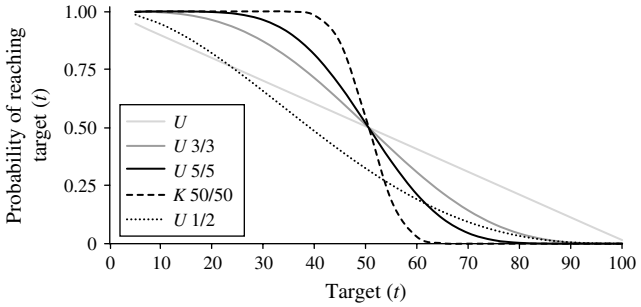
Appendix A

Here, we illustrate how uncertainty combines with observed performance in the presence of performance targets to create value. Imagine a decision maker faced with two urns and a performance target (t) to be met or exceeded in the next 100 draws with replacement. Each urn contains 100 balls of two colors, one corresponding to successes (S) and the other to failures (F). The known (K) urn contains 50 S and 50 F balls; the unknown (U) urn can contain any number of S balls from 0 to 100, where S was drawn from a uniform distribution. The K urn is the urn analogue to Derring Do or the seasoned job applicant from our introduction; the U urn is the analogue of Dark Knocks or the new graduate.

To illustrate the effects of sample size, assume we start with two U urns. A sample of 6 balls is drawn from one urn and 10 from the other, and both samples reveal a 50/50 distribution of S and F balls. We denote the two urns by $U_{3/3}$ and $U_{5/5}$, respectively. For both urns all possible combinations, except 0 and 100 S balls, are still possible. Yet based on those samples, they are no longer equally likely. Both samples indicate that the modal composition is 50 S and 50 F balls, and the expected performance of a single draw is at $p(S) = 0.5$ for both urns. At the same time, the posterior distribution associated with the $U_{3/3}$ urn is more spread out.

Formally stated, the probability of reaching a given target with a $U_{S/F}$ urn can be calculated by the weighted average of the decumulative Binomial distribution for each possible urn, where weights are the urns' posterior densities based on the sample, $\text{Beta}_{(S+1,F+1)}$. Figure A.1 shows the corresponding posterior probability distributions of achieving different targets for five urns: a U urn with no sample, a K urn with a known composition of 50 S balls, and three urns about which there is sample information, a $U_{1/2}$ urn and the $U_{3/3}$ and $U_{5/5}$ urns. For every target above the expected performance of 0.50, with the exception of the $U_{1/2}$ urn, the probability of reaching the target is higher for smaller samples. Figure A.1 also shows how uncertainty resulting from small samples can compensate for a lower observed performance. The $U_{1/2}$ urn with a $p(S) = 0.33$ has a higher likelihood of achieving stretch targets—here, $t > 62$—than the $U_{5/5}$ urn.

Figure A.1. Probability Distributions of Reaching Targets for Urns with an Expected Performance of $p(S) = 0.50$ and $p(S) = 0.33$, and Different Amounts of Prior Information as Illustrated in Sample Sizes of 0 (U), 3 (U 1/2), 6 (U 3/3), 10 (U 5/5), and ∞ (K 50/50)



References

Antoniou C, Harrison GW, Lau MI, Read D (2014) Information characteristics and errors in expectations: Experimental evidence. Working paper, University of Warwick, Coventry, UK.

Bar-Hillel M (1979) The role of sample size in sample evaluation. *Organ. Behav. Human Performance* 24(2):245–257.

Bernoulli D (1954) Exposition of a new theory on the measurement of risk. *Econometrica* 22(1):23–36.

Bollinger CR, Hotchkiss JL (2003) The upside potential of hiring risky workers: Evidence from the baseball industry. *J. Labor Econom.* 21(4):923–944.

Brunswik E (1955) Representative design and probabilistic theory in a functional psychology. *Psych. Rev.* 62(3):193–217.

Burgess S, Lane J, Stevens D (1998) Hiring risky workers: Some evidence. *J. Econom. Management Strategy* 7(4):669–676.

Cahan S (2010) Decision quality (always) increases with the size of information samples-provided that the decision rule is statistically valid: Comment on Fiedler and Kareev (2006). *J. Experiment. Psych. Learn., Memory, Cognition* 36(3):829–41.

Curley SP, Yates JF (1985) The center and range of the probability interval as factors affecting ambiguity preferences. *Organ. Behav. Human Decision Processes* 36(2):273–287.

Curley SP, Yates JF (1989) An empirical evaluation of descriptive models of ambiguity reactions in choice situations. *J. Math. Psych.* 33(4):397–427.

Dixit A, Nalebuff B (1991) *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life* (W. W. Norton & Company, New York).

Ellsberg D (1961) Risk, ambiguity, and the savage axioms. *Quart. J. Econom.* 75(4):643–669.

Evans JSBT, Dusoir D (1977) Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psych.* 41(2–3): 129–137.

Fiedler K, Kareev Y (2006) Does decision quality (always) increase with the size of information samples? Some vicissitudes in applying the law of large numbers. *J. Experiment. Psych.: Learn., Memory, Cognition* 32(4):883–903.

Fiegenbaum A, Thomas H (1988) Attitudes toward risk and the risk–return paradox: Prospect theory explanations. *Acad. Management J.* 31(1):85–106.

Fishburn PC (1977) Mean-risk analysis with risk associated with below-target returns. *Amer. Econom. Rev.* 67(2):116–126.

Griffin D, Tversky A (1992) The weighing of evidence and the determinants of confidence. *Cognitive Psych.* 24(3):411–435.

Heath C, Larrick RP, Wu G (1999) Goals as reference points. *Cognitive Psych.* 38(1):79–109.

Hendricks W, DeBrock L, Koenker R (2003) Uncertainty, hiring, and subsequent performance: The NFL draft. *J. Labor Econom.* 21(4):857–886.

Hogarth RM, Einhorn HJ (1990) Venture theory: A model of decision weights. *Management Sci.* 36(7):780–803.

Irwin FW, Smith WA, Mayfield JF (1956) Tests of two theories of decision in an expanded judgment situation. *J. Experiment. Psych.* 51(4):261–268.

Jacobs JE, Narloch RH (2001) Children's use of sample size and variability to make social inferences. *J. Appl. Developmental Psych.* 22(3):311–331.

Kacelnik A, Bateson M (1996) Risky theories—The effects of variance on foraging decisions. *Amer. Zoologist* 36(4):402–434.

Kahn BE, Sarin RK (1988) Modeling ambiguity in decisions under uncertainty. *J. Consumer Res.* 15(2):265–272.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.

Kareev Y (2000) Seven (indeed, plus or minus two) and the detection of correlations. *Psych. Rev.* 107(2):397–403.

Kareev Y, Avrahami J (2007) Choosing between adaptive agents: Some unexpected implications of level of scrutiny. *Psych. Sci.* 18(7):636–641.

Laughhunn DJ, Payne JW, Crum R (1980) Managerial risk preferences for below-target returns. *Management Sci.* 26(12):1238–1249.

Lazear EP (1995) Hiring risky workers. NBER Working Paper 5334, National Bureau of Economic Research, Cambridge, MA.

Lopes LL (1981) Decision making in the short run. *J. Experiment. Psych.: Human Learn. Memory* 7(5):377–385.

March JG, Shapira Z (1987) Managerial perspectives on risk and risk taking. *Management Sci.* 33(11):1404–1418.

Markowitz H (1952) The utility of wealth. *J. Political Econom.* 60(2):151–158.

Masnick AM, Morris BJ (2008) Investigating the development of data evaluation: The role of data characteristics. *Child Development* 79(4):1032–1048.

Norton MI, Frost JH, Ariely D (2007) Less is more: The lure of ambiguity, or why familiarity breeds contempt. *J. Personality Soc. Psych.* 92(1):97–105.

Obrecht NA, Chesney DL (2013) Sample representativeness affects whether judgments are influenced by base rate or sample size. *Acta Psych.* 142(3):370–382.

Obrecht NA, Chapman GB, Gelman R (2007) Intuitive tests: Lay use of statistical information. *Psychonomic Bull. Rev.* 14(6):1147–1152.

Payne JW, Laughhunn DJ, Crum R (1980) Translation of gambles and aspiration level effects in risky choice behavior. *Management Sci.* 26(10):1039–1060.

Peterson CR, DuCharme WM, Edwards W (1968) Sampling distributions and probability revisions. *J. Experiment. Psych.* 76(2, Part 1):236–243.

Rode C, Cosmides L, Hell W, Tooby J (1999) When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. *Cognition* 72(3):269–304.

Sedlmeier P, Gigerenzer G (1997) Intuitions about sample size: The empirical law of large numbers. *J. Behavioral Decision Making* 10(1):33–51.

Simon HA (1955) A behavioral model of rational choice. *Quart. J. Econom.* 69(1):99–118.

Sung B (1966) Translations from James Bernoulli Technical Report No. 2, Contract Nonr 1866(37), NR-042-097, Harvard University, Cambridge, MA.