Available online at www.sciencedirect.com
ScienceDirect
Organizational Behavior and Human Decision Processes 106 (2008) 168–179
ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
www.elsevier.com/locate/obhdp

Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice ☆

Tim Rakow [a,*], Kali A. Demes [a], Ben R. Newell [b]

[a] *Department of Psychology, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK*
[b] *School of Psychology, University of New South Wales, Sydney 2052, Australia*

Received 14 August 2007; accepted 4 February 2008
Available online 24 March 2008
Accepted by John Schaubroeck

Abstract

Most experimental investigations of risky choice provide participants with a *description* of the probabilities and outcomes for each option, and observe that small probabilities are *overweighted*. However, when payoffs are learned from repeated *experience* of outcomes (as in many real-world decisions), different patterns of choice are observed—consistent with *underweighting* rare events. We re-examined this phenomenon to determine whether biased sampling and recency effects in experience-based choice could account for this description–experience gap. Two hundred and forty paid participants made choices for 12 pairs of simple gambles. In the *objective description* condition, probabilities and outcomes were specified. In the *free sampling* condition, participants observed repeated plays of each gamble before choosing. Participants in four yoked conditions received the same information as the free sampling participants—either described or experienced. Differences between objective description and free sampling were consistent with underweighting rare events in experience-based choice. However, consistent with a biased sampling account, patterns of choice in the yoked conditions barely differed from the free sampling condition: given identical information, presentation mode has no effect. Recency effects in choice occurred only when outcomes were actively sampled, and were unaffected by working memory capacity. The absence of recency for passive observation implies actor–observer differences in forming expectations or testing hypotheses. The results provide no support for the claim that decisions from description and decisions from experience require separate descriptive theories.
© 2008 Elsevier Inc. All rights reserved.

Keywords: Risky choice; Description–experience; Decisions from experience; Sampling; Prospect theory; Rare events; Working memory; Actor–observer differences

Introduction

In much the same way as the humble fruit fly provides biologists with a simple model for understanding genetic transmission in more complex organisms, choices between simple gambles provide decision psychologists with the means to understand processes that should have some bearing upon important real-world decisions under uncertainty. This approach has taught us a great deal. For instance, we know that people seem to overweight small probabilities—they act as if low probability events are more likely to occur than is

☆ The work reported in this article was supported by a grant awarded by the British Academy to the first author. We thank Nicolas Geeraert, Linda Morison, Fred Westbrook, and the members of the University of Essex Psychology Department Staff Seminar, the University of New South Wales Cognition and Learning Group, and the London Judgment and Decision Making Group for comments on this work. We thank Jonathan Baron for suggestions on data analysis, and Nigel Harvey for comments and helpful suggestions on an earlier draft of this paper.
* Corresponding author. Fax: +44 1206 873801.
E-mail address: timrakow@essex.ac.uk (T. Rakow).

0749-5978/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2008.02.001

actually so. Most people choose a 0.1% chance of winning $5000 in preference to a sure win of $5, but are willing to submit themselves to a sure loss of $5 when the alternative is to gamble on a 0.1% chance of losing $5000 (Kahneman & Tversky, 1979). Overweighting the small probability in these choice problems makes lotteries with low chances of big wins unusually attractive, and risky options with a small chance of a large loss particularly unappealing—thereby making insurance especially seductive. Prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), arguably the most successful descriptive theory of decision making (Laibson & Zeckhauser, 1998), incorporates this observation by assuming a non-linear weighting function for probabilities. Ample data from laboratory and field experiments demonstrate that probabilities in the range $[0, \approx .25]$ are generally overweighted (Tversky & Fox, 1995).

However, recent data have challenged the generality of this phenomenon. The controversy rests on the distinction between decisions from *description* and decisions from *experience*. While most laboratory studies of risky choice provide participants with a complete *description* of the options, everyday life rarely provides us with such clear information. Often we must use personal experience to estimate probabilities: nobody tells you the probability of your car starting when the key is turned. The necessity for estimating probabilities and the distribution of possible outcomes may be particularly common for the kinds of social decisions that abound in organizational settings. Whilst medical journals can provide you with the chances of adverse side-effects for a specific drug, no definitive source exists to inform employees of the probability with which their managers respond positively to calls for assistance.

Crucially, decisions from description and decisions from experience yield different patterns of choice. In laboratory experiments where participants learn about the options by observing outcomes for repeated plays of a gamble, their choices reflect the *underweighting* of small probabilities. Barron and Erev (2003) observed this using a "money-machine" paradigm. Participants saw two buttons (representing money machines) on a computer screen. Each time a button was pressed the machine delivered a small payoff according to pre-programmed probabilities. Choices in various problems were consistent with the underweighting of small probabilities using this experience-based paradigm.

Hertwig, Barron, Weber, and Erev (2004) and Weber, Shafir, and Blais (2004) report similar findings using a "free sampling" paradigm, where participants observed payouts from two sources without monetary consequences until they were confident which option they wished to play once for real (i.e., a one-shot decision with financial consequences). For instance, when described, only 36% of participants preferred a 0.8 probability of 4 points over 3 points with certainty—consistent with the 0.2 chance of 0 points in the gamble being particularly unattractive. However, 88% of participants were willing to gamble when they learned from experience—consistent with a comparatively low weight for the unattractive rare event.

Hertwig et al. (2004) provide two explanations for these differences in choice proportions between described and experienced choices (the "description–experience gap"). First, because most participants chose to examine small samples, they often observed biased samples of outcomes. Specifically, because the sampling distribution for the binomial distribution is highly skewed when both p and N are small, participants were more likely than not to observe samples in which rare events are under-represented—sometimes not even occurring at all. Consistent with this, there were extreme differences in patterns of choice between participants who encountered a rare event less often and those who encountered it more often than expected. Second, consistent with reliance on small subsamples of the most recently observed outcomes, recency effects were observed whereby the expected value (EV) of the latter half of the observations for each option predicted choice more effectively than EV of the first half of the observations.

Both explanations can be framed in terms of Kareev's narrow window hypothesis of information search (Kareev, 1995a, 1995b, 2000; Kareev, Lieberman, & Lev, 1997), which assumes that inferences are based on small samples of information held in working memory. The limited capacity of working memory limits the number of exemplars that people can consider at any one time—pushing them to limit the information that they acquire or to consider only the most recent exemplars. Consequently, recency accentuates the effect of biased sampling: "...rare events have less impact than they deserve not only because decision makers have not encountered them, or have encountered them less frequently than expected, but also because they have not encountered them recently." (Hertwig et al., 2004, p. 538). In this paper, we examine both of the suggestions for the description–experience gap that Hertwig et al. (2004) put forward: (1) sampling bias and (2) a reliance on small samples of recent events arising from limitations in memory capacity.

Examining sampling bias. The striking difference between choices made from description and those made from experience, has led for a call for alternative theories for decisions from experience (Hertwig et al., 2004; Weber et al., 2004)—notably because some of the cornerstones of prospect theory, including the overweighting of small probabilities, do not hold for experience-based choice (Barron & Erev, 2003). Fox and Hadar

(2006) take issue with this position. In a re-analysis of Hertwig et al.'s (2004) data they showed that prospect theory can account for the choice patterns if one considers the *observed* probabilities of rare events, which, because samples were biased, differed from the objective probabilities. We provide an independent empirical test of the adequacy of prospect theory to account for experienced-based choices using a yoked design. We present samples that participants have encountered in a free sampling condition to new participants: either describing the samples, or presenting them sequentially (by experience)—thereby controlling for sampling bias. This permits us to explore the extent to which the description–experience (D–E) gap is a function of biased sampling or the mode of information presentation. Do participants choose differently because they acquire biased information and/or because they acquire it in a different manner? If the D–E gap persists even when sample variation is controlled, it implies that the distinction between experienced and described choices is not merely a statistical phenomenon, and the assumption of a real difference between these modes of presentation is valid. However, if choices depend only on the information given, and not upon the mode of presentation, we can reject calls for alternative descriptive theories of choice for described and experienced choices.

Examining recency and the role of limited memory. We examine whether limitations in memory capacity can fully account for a reliance on small samples and recent subsamples. This is an important proposal to investigate, as it implies that experienced decisions will differ fundamentally from described ones even when there is no sampling bias—because experience-based decisions are based on a subset of the available information, and/or because some observations (e.g., recent ones) receive more weight than other ones. We adopt an individual differences approach to see if information acquisition and the degree of recency are predicted by measures of memory capacity. If capacity limitations influence the D–E gap as proposed—sampling should be less extensive, and recency should be more marked among those individuals whose memory capacity is low. If this is so, memory capacity should moderate the size of the D–E gap: accentuating the difference among individuals with low memory capacity, but reducing it among those with high capacity (who can accumulate larger samples, which are less likely to be biased).

Method

Participants

Two hundred and forty volunteers (83 male) were recruited from the University of Essex participant pool, with mean (median) [sd] age 23.3 (21) [6.8] years. Participants received a turn-up fee of UK£3 (US$6) plus payment contingent upon their choices (range UK£0.00 to UK£2.10). Twenty-six participants who participated for course credit received only the contingent payment.

Design and tasks

Choice problems. Table 2 shows the 12 problems that were used, which were divided into two sets of six (Set 1 = problems 1–4, 9–10). All problems had two options of equal or similar EV, each option having at most two outcomes. We designate the option with the greater variance in outcomes as "risky". Participants played all 12 problems using the "money-machine" paradigm (Barron & Erev, 2003; Hertwig et al., 2004): either Set 1 as described problems and Set 2 as experienced problems, or the reverse pattern (counterbalanced). Therefore, while participants made choices for described and experienced problems, analyses of individual problems are between subjects. The outcome of choices was withheld until all problems were completed in order to reduce wealth effects. Points were converted to money at the end of the study if the participant was in profit (UK£0.05 per point).

Six conditions were employed that varied the presentation format of the problems (summarized with an additional example in Table 1). These were arranged in three description–experience pairs to equalize participants' time commitment across conditions. Each participant was allocated to one of the following pairs of conditions (each $N = 80$).

First pair of conditions:

1. *Objective description:* the percentage chance for each outcome was specified (e.g., (1) 80% chance of 4 points, 20% chance of 0 points, vs (2) 100% chance of 3 points).
2. *Free sampling:* participants sampled outcomes from each money machine, in any order they wished, until they were confident about which machine they wished to play for real. Money machines were programmed according to the objective probabilities of the problems.

For each participant, problem order and the allocation of options (risky/safe) to machines (1 on the left/2 on the right) were randomly determined.

Yoked conditions. The remaining four conditions were "yoked" conditions, in which participants saw the same outcomes that participants in the free sampling condition had seen. Specifically, the exact set of outcomes for a given problem that had been seen by a participant in the free sampling condition (prior to his/her final choice) was presented again to new participants. There were two yoked description conditions and two yoked experience conditions—therefore, each set of outcomes that had been encountered by a free sampling participant was presented again to four participants (with the manner of presenta-

Table 1
Summary of experimental conditions with example of the information seen for a participant encountering Problem 2

Condition (Y = yoked to free sampling)	Prior to choosing one machine to play for actual payoffs, the participant sees	Information format
1. Objective description		
Machine 1	20% chance of 16 points, 80% chance of 0 points	On-screen text
Machine 2	100% chance of 3 points	
2. Free sampling		
Machine 1	0 0 16 0 16 0 0 0 0	Values in sequence
Machine 2	3 3 3 3 3 3 3	
3. Frequency sample description (Y)		
Machine 1	16 points on 2 out of 9 occasions, 0 points on 7 out of 9 occasions	On-screen text
Machine 2	3 points on 7 out of 7 occasions	
4. Ordered passive sampling (Y)		
Machine 1	0 0 16 0 16 0 0 0 0	Values in sequence
Machine 2	3 3 3 3 3 3 3	
5. Percentage sample description (Y)		
Machine 1	16 points on 22% of occasions, 0 points on 78% of occasions	On-screen text
Machine 2	3 points on 100% of occasions	
6. Reversed passive sampling (Y)		
Machine 1	0 0 0 0 16 0 16 0 0	Values in sequence
Machine 2	3 3 3 3 3 3 3	

tion varying between these four yoked participants). Each participant in these yoked conditions was uniquely matched with two participants from the free sampling condition (one for each problem set). Problem order and left/right presentation of options were determined by the yoking procedure to ensure equivalence between conditions. Thus, in yoked experience conditions, participants clicked the same buttons that his/her yoke had clicked to reveal the same information in the same position on the screen as his/her yoke had seen. If a participant observed only one of two possible outcomes under free sampling, there was no reference to the unobserved outcome. (For clarity of understanding in what follows, consider the example of being yoked to a participant who observed 4,4,4,0,4,0,4,4,4,0 from Machine 1, followed by 3,3,3,3,3 from Machine 2.)

Second pair of conditions:

3. Frequency sample description: the yoke's observations were described using frequencies (e.g., (1) 4 points on 7 out of 10 occasions, 0 points on 3 out of 10 occasions, vs (2) 3 points on 5 out of 5 occasions).
4. Ordered passive sampling: the participant observed precisely the same sequence of outcomes as the yoke had (e.g., 4,4,4,0,4,0,4,4,4,0 from Machine 1, then 3,3,3,3,3 from Machine 2).

Third pair of conditions:

5. Percentage sample description: the yoke's observations were described using percentages (e.g., (1) 4 points on 70% of occasions, 0 points on 30% of occasions, vs (2) 3 points on 100% of occasions).
6. Reversed passive sampling: the participant observed the outcomes of the yoke's sequence of observations in reverse (e.g., 3,3,3,3,3 from Machine 2, followed by 0,4,4,4,0,4,0,4,4,4 from Machine 1).

This design ensured that 40 participants in each condition completed each problem. It permits several key comparisons of potential influences upon choice that are of theoretical interest. Described decisions can be compared with decisions from experience by comparing conditions 1, 3 and 5, against 2, 4 and 6. In order to test Fox and Hadar's (2006) critique, choices based on objective probabilities can be compared with those for sampled probabilities (1 vs 2–6). By considering the three experience-based conditions, active sampling can be compared with passive sampling (2 vs 4, 6) to allow consideration of actor–observer differences (Koehler, 1994). Some previous research finds that reasoning about uncertainty varies between frequency and probability formats (Gigerenzer & Hoffrage, 1995), which can be examined by comparing the two yoked description conditions (3 vs 5). Finally, by comparing the two yoked experience conditions (4 vs 6), the difference between preserved and non-preserved order information can be examined.

Memory tasks. Two standard measures of memory capacity were obtained. In the forward digit span task (a measure of 'pure' storage capacity), participants must recall sequences of digits. Sequence length increases by one after every other sequence (from 3 to 8), and a point is awarded for each sequence correctly recalled. In the working memory (WM) for words task (a measure of storage capacity plus processing efficiency), participants identify the odd-one-out among a set of 4 words, to be

recalled in the correct order once several sets have been seen. The number of sets increases by one after every third trial (from 2 to 5), and a point is awarded for each target word recalled in the correct position.

Apparatus and materials

All tasks were programmed using RealBasic and run in a "Windows-style" environment on 1.25 GHz Macintosh eMac computers with 17-inch CRT screens. Responses to the memory tasks were recorded on tape to be scored later.

Procedure

Participants were tested individually, always completing the choice problems before the memory tasks. In order to reduce opportunity costs that might lead to minimal engagement in the choice problems, participants in the first pair of conditions were informed that the second task (memory measures) would not start until at least 20 min after the first task commenced. The time taken in subsequent conditions was determined by the yoking procedure (i.e., a yoked participant would take about the same time to complete each problem as the person that they were yoked to).

On-screen instructions outlined the money-machine paradigm, informing participants of the financial consequences of their decisions, and explaining that there were two types of game named "sample play" (our name for experience-based decisions) and "information play" (description-based decisions). Participants were instructed (and reminded) that: "In each game there will ALWAYS be a difference between how the two machines award points." Participants were given two practice games, one for each type of game.

In the free sampling condition, participants used the computer mouse to click on the money machines, which revealed the payout for 1.5 s. Participants had to sample each machine at least once, and could sample each machine up to 100 times (though participants were not informed of this in advance, nor was this maximum reached on any occasion). Participants were instructed to sample the machines as many times as they liked until they were confident that they knew which machine they wanted to play for real, and were free to sample from the machines in any order that they liked. By clicking a third button, participants moved to the "real play" phase of a choice problem to make a single-shot decision.

In the participant instructions in the yoked conditions, we were careful to stress that: "...your task is to decide which of the machines you would prefer to select to receive 'real points' to be added to your final score." To ensure that there was no contradictory implication that participants should be attempting to guess what their yoke chose in the real play phase, no mention was made of a third party. Participants were simply told that they would "view information about how each of the two machines has awarded points on previous occasions (the most recent occasions on which each machine has awarded points)" and "examine 'free samples' from each machine" in order to inform their decisions.

In the yoked experience conditions, participants performed the same number of mouse clicks on the equivalent buttons to reveal the same information that had been seen by his/her yoke. The presentation format and timing also matched the free sampling condition, ensuring that the time taken to view all the information was similar to the free sampling condition. Then the participant clicked the button to enter the real play phase. To ensure that the correct amount of information was viewed in the prescribed order, the computer program automatically disabled all on-screen buttons except the action that had been determined by the yoking procedure. So that this process was not confusing, participants were also instructed that they should click on the button with a bold caption. These conditions were identical to the free sampling in all other respects, and placed no constraints on choice in the real play phase.

In the description conditions, participants were given the relevant information about the outcomes associated with each pair of options, they chose in their own time and then clicked a button to reveal the next problem.

Brief instructions for the memory tasks were given verbally, with full instructions appearing on screen. Stimuli for these tasks were presented on screen at a fixed rate, and participants gave their responses out loud when prompted by a tone. Finally, participants were thanked, debriefed and paid.

Results

Patterns of choice

Table 2 shows the percentage choosing the risky option in each condition. The expected direction of difference between the objective description and free sampling conditions depends upon whether the rare event is desirable or not (i.e., best vs worst possible outcome). When the rare event in the risky option is *undesirable*, underweighting small probabilities (as anticipated for experience-based decisions) increases the attractiveness of the risky option. Therefore, *more* participants are expected to choose risky options under free sampling than for objective description. When the risky event is *desirable*, risky options become less attractive if rare events are underweighted, so fewer participants are expected to choose risky options under free sampling compared with objective description. Table 2 indicates that the difference between objective description and free sampling falls in the expected direction for 11 of the 12

Table 2
Percentage choosing risky option by condition

Problem	Prior use[+]	Option		Desirable rare event	Percentage choosing risky option						Free sampling, % risky[#]	
		Risky	Safe		Objective description (1)	Free sampling (2)	Frequency sample description	Ordered passive sampling	Percentage sample description	Reversed passive sampling	Rare event seen < expected	Rare event seen ⩾ expected
1	B4 H1	4, .8	3, 1.0	No	8^{2b}	60^{1b}	55^{1b}	58^{1b}	50^{1b}	53^{1b}	95 (18/19)	29 (6/21)
2		16, .2	3, 1.0	Yes	38	28	40	58^{2}	35	33	11 (2/19)	43 (9/21)
3	B6 H4	−4, .8	−3, 1.0	Yes	65^{2}	38^{1}	38^{1}	33^{1b}	45	38^{1}	29 (7/24)	50 (6/16)
4	B5 H2	4, .2	3, .25	Yes	45	58	50	43	58	48	29 (6/21)	89 (17/19)
5	B9 H5	32, .1	3, 1.0	Yes	33	25	40	28	28	23	4 (1/25)	60 (9/15)
6		16, .2	3, 1.0	Yes	38^{2}	18^{1}	43^{2}	30	40^{2}	23	9 (2/23)	29 (5/17)
7	B11 H3	−32, .1	−3, 1.0	No	48^{2}	73^{1}	73^{1}	63	60	58	84 (21/25)	53 (8/15)
8	B10 H6	32, .025	3, .25	Yes	55	35	33^{1}	28^{1}	30^{1}	43	19 (6/32)	100 (8/8)
9	W1	10, .1	1, 1.0	Yes	35	18	25	38^{2}	23	28	4 (1/25)	40 (6/15)
10	B7 W2	10, .9	9, 1.0	No	15^{2b}	55^{1b}	53^{1b}	58^{1b}	43^{1}	38^{1}	94 (15/16)	29 (7/24)
11	B8	−10, .9	−9, 1.0	Yes	83^{2b}	33^{1b}	35^{1b}	35^{1b}	33^{1b}	38^{1b}	17 (4/24)	56 (9/16)
12		10, .05	1, .5	Yes	60^{2b}	28^{1b}	28^{1b}	30^{1}	30^{1}	35^{1}	23 (6/26)	36 (5/14)
Mean ϕ in predicted direction[‡] (vs objective description)						.25	.20	.20	.18	.20		
Mean ϕ in predicted direction[‡] (vs free sampling)						—	−.06	−.05	−.07	−.06		

[1]denotes significantly different from objective description by χ^2 ($p < .05$), [b]denotes also significant applying a Bonferroni correction for 11 comparisons ($p < .0045$).
[2]denotes significantly different from free sampling by χ^2 ($p < .05$), [b]denotes also significant applying a Bonferroni correction for 11 comparisons ($p < .0045$).
[+]D–E effect previously examined for this problem: B = Barron and Erev (2003), H = Hertwig et al. (2004), W = Weber et al. (2004); digit denotes problem number assigned in the cited paper.
[‡]Predicted direction is the direction of effect if small probabilities are underweighted in subsequent conditions relative to the baseline condition.
[#]Percentages, with actual frequencies shown in brackets.

problems. Only Problem 4, for which Hertwig et al. (2004) had found no significant effect, falls in the unpredicted direction. Seven of these 11 differences were found to be significant ($p < .05$) by a χ^2 test (denoted with superscript 1 in Table 2). Four of these differences remain significant when a Bonferroni correction is applied (i.e., $p < .0045$), which adjusts for the 11 pair-wise comparisons that we made per problem (all conditions vs 1 and 2, plus 3 vs 5 and 4 vs 6). Phi (ϕ) correlations were calculated as a measure of effect size for each difference and these indicate that the average effect in the predicted direction for this pair of conditions was moderate ($\phi = .25$). A one-sample t-test showed that this mean effect was significantly greater than zero ($t(11) = 4.65$, $p = .001$), confirming that, for the set of 12 problems taken as a whole, we had replicated the D–E gap that has been reported elsewhere (Hertwig et al., 2004; Weber et al., 2004).

Following Hertwig et al. (2004), we performed a *post hoc* analysis of the samples that participants drew. The final two columns of Table 2 show a remarkably clear difference in choices between those who observed the rare event less often than expected and those who saw it at least as often as expected. Nine of these differences in the percentage of participants choosing the risky option were significant ($p < .05$), suggesting that sampling variation does indeed play an important role in the D–E gap.

A priori evidence of this can also be found by comparing the four yoked conditions with the first pair of conditions. First, readers should note that the percentage of participants choosing the risky option is, in general, very similar to that for the free sampling conditions. Therefore, most of the time that there is a significant difference between the first pair of conditions, this significant difference is preserved between objective description and the yoked conditions—even though option information is experienced or described in a variety of ways in these yoked conditions. This similarity also holds at the individual level: 69% of choices made in the four yoked conditions were the same as the choice made by the participant's yoke for that problem in the free sampling condition, whereas only 56% concordance is expected by chance based on the choice proportions obtained. Concordance rates across problems were significantly above chance rates for all four yoked conditions (all $t(11) > 5.0$, all $p < .001$), though were slightly higher for the two sample description conditions (71% each) than for ordered or reversed passive sampling (66% and 67%). Choice proportions are sufficiently similar between free sampling and the four yoked conditions that only 4 of the 48 paired differences are significant ($p < .05$, indicated by superscript 2 in Table 2)—barely above the 2.4 that would be expected if H_0 were true. None were significant when the Bonferroni correction for multiple comparisons was applied. Furthermore, in contrast to recent data from Gottlieb, Weiss, and Chapman (2007), none of the 12 pairs of choice proportions differed significantly between frequency and percentage sample description (all $p > .23$),

and, only one pair of choice proportions (Problem 2) differed significantly ($p < .05$) between ordered and reversed passive sampling (though not when corrected for multiple comparisons, $p > .0045$).

Whilst absence of significance cannot be taken as absence of effect, having a large number of effects means that we can use a meta-analytic approach using *P–P* plots, which examine the concordance of these effects with H_0 as follows. By definition of the *p*-value, if H_0 is true, *p*-values will be uniformly distributed in the range [0, 1] with equal numbers and distribution of effects in each direction. We therefore examined the 48 paired comparisons between the four yoked conditions and free sampling (12 problems × 4 pairs of conditions). We calculated the cumulative probability for each effect from its *p*-value: $p/2$ for negative effects and $(1 - p/2)$ for positive effects (e.g., zero effects have a cumulative probability of 0.5). Fig. 1a and b show two uniform *P–P* plots, which graph these cumulative probabilities against the expected cumulative probabilities for a uniform distribution. For a perfect uniform distribution, points in the *P–P* plot fall on the identity line—representing an equal distribution of positive and negative effects with X% of results significant at the X/100 level as expected under H_0. The greater the deviation from what is expected under H_0, the greater the *P–P* plot will deviate from this pattern. If there are a large number of positive effects, the *P–P* plot will exhibit a positively accelerating function—consistent with having more small *p*-values for positive effects (i.e., large cumulative probabilities) than expected under H_0. Conversely, a disproportionately large number of negative effects results in a negatively accelerating function—reflecting a larger-than-expected number of small *p*-values for negative effects (i.e., small cumulative probabilities).

Fig. 1a shows that when an *increase* in the percentage of risky choices from free sampling is denoted as a *positive* effect, the distribution of *p*-values is remarkably close to uniform. Fig. 1b shows a small deviation from uniformity when a *positive* effect indicates *inflation* of the effect predicted by overweighting rare events under objective description. Slightly more negative effects are observed than expected (points fall above the identity line), representing a small regression back towards the choice proportions of the objective description condition. However, the final line of Table 2 shows that this effect is very small indeed—on average, accounting for no more than 0.5% of the variance in choice proportions.

Patterns of sampling

For the free sampling condition, we computed each participant's average number of observations per problem (total sample size), average number of periods of uninterrupted observation from a single machine per problem (number of sub-samples), and the average number of observations in each of these sub-samples (sub-sample size). The mean (median) values were 16.2 (15) for total sample size, 4.2 (3) for the number of sub-samples, and 5.5 (4) for sub-sample size. Thus, a "typical" approach to exploring the two money machines might be to observe four outcomes for one machine, then four from the other, then to repeat this process again before making the final choice. However, there is considerable variability between participants, and this variability is related to working memory (WM) scores. Participants with higher WM scores collected larger sub-samples ($r = .38$, $p < .001$) and had larger total sample sizes ($r = .36$, $p = .001$), though WM scores and the number of sub-samples were unrelated ($r = -.05$, $p = .674$). Thus a participant with a standardized WM score of −1 is expected to make 3.8 observations per sub-sample and 12.7 observations in total per

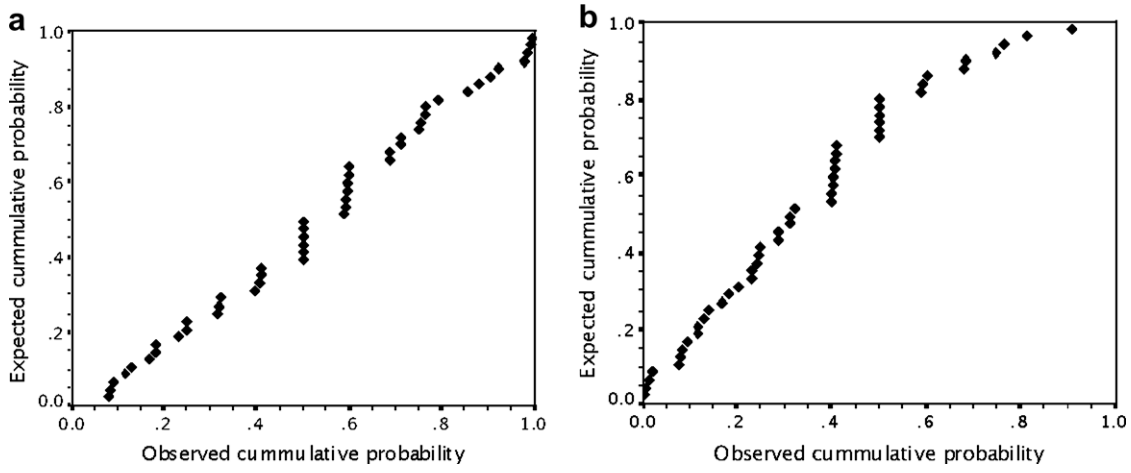

Fig. 1. *P–P* plots of observed cumulative probability (*x*) vs expected cumulative probability (*y*) under the null hypothesis of no difference between the free sampling condition and the yoked conditions. Points falling on the identity line are consistent with H_0. (a) Positive effect denotes higher % of risky options. (b) Positive effect denotes accentuated D–E effect.

problem, whereas with a standardized WM score of +1, sub-samples of 7.5 with 20.2 observations overall are expected. Digit span was not significantly related to any of these three measures of sampling behavior (all $r < .2$), and is not considered further.

Collecting smaller samples should accentuate the D–E effect—therefore, we conducted a median split on WM scores to create low- and high-WM groups to determine whether individual differences in sampling described above were sufficient to increase the D–E effect. The mean D–E effect was greater for the low-span group ($\phi = .27$ vs .21), but this difference was not significant, $t(11) = .91$, $p = .381$. To perform a more direct test of the effect of sampling behaviour on the D–E effect, we performed a median split on total sample size. As expected, the mean D–E effect was greater for those acquiring smaller samples ($\phi = .29$ vs .22), but again this difference was not even close to significant, $t(11) = 1.05$, $p = .316$.

The influence of recent observations

The possibility of a greater influence for recent observations in the three experience conditions was explored using a method similar to Hertwig et al. (2004). The EV of each participant's observations for each option were computed individually for each problem. Then it was determined whether the choice made was consistent with selecting the option with higher EV. (Where options were tied, the prediction favoured the "safe" option.) This procedure was performed separately for the first and second halves of each set of observations, as well as for the entire sequence. In order to provide some control for the wide variation in sample sizes, and to remove trivially small samples where *all* observations are essentially recent, we only examined cases where eight or more observations were made over both options of the problem. The median number of exclusions according to this criterion was 10.5 out of 40 per problem. This analysis was also performed separately for the high- and low-WM groups. The logic of this analysis is that it provides a measure of each participant's sensitivity to the point values that they have observed. For instance, even if someone is generally risk averse and their choices are poorly predicted by EV, it will be possible to infer that he/she seems to pay greater attention to recent samples if recent observations predict choices more effectively than initial observations.

Table 3 shows the outcome of this exercise. Consistent with Hertwig et al. (2004), we found significantly better prediction on the basis of more recent observations under free sampling, consistent with greater impact of more recent observations. This recency effect differed little between the high- and low-span groups. However, as high-span participants had a tendency to make more observations, this may not be a like-with-like comparison. Analysis of subsequent conditions removes this confound between memory group and sample size, because participants were allocated randomly to their yokes. No significant recency effects were observed in either passive sampling condition for either memory group, in fact small (non-significant) primacy effects were more common (better predictions based on *initial* observations).

Discussion

Recent research has observed a striking difference between decisions from experience and decisions from experience. This has led to calls for separate descriptive theories for the two types of decision (Hertwig et al., 2004; Weber et al., 2004), because existing theories of risky choice such as prospect theory seemingly fail to capture some of the key features of experience-based decisions (Barron & Erev, 2003).

Substantially a sampling phenomenon

Our data confirm that biased sampling is in large part responsible for the D–E gap in risky choice. Large discrepancies in patterns of choice were dependably related to whether the rare event was seen more or less often than expected. This translated into differences in choice between participants who received objective descriptions of the problems and those who undertook experience-based sampling of outcomes prior to choice. Crucially, these differences were preserved in the four yoked conditions. Given the same samples of information on which to base their choice, it mattered little whether

Table 3
Percentage of choices predicted by expected value (experience conditions only) by memory group (high vs low) and sequence (first vs second half)

Memory group	Free sampling condition				Ordered passive sampling condition				Reversed passive sampling condition			
	1st half	2nd half	All trials	$t(11)$, p for diff. 1st vs 2nd half	1st half	2nd half	All trials	$t(11)$, p for diff. 1st vs 2nd half	1st half	2nd half	All trials	$t(11)$, p for diff. 1st vs 2nd half
Low	66	76	71	2.51, .029	67	70	72	0.99, .344	67	63	73	−0.50, .631
High	64	76	76	2.78, .018	64	61	66	−0.62, .550	68	63	70	−1.08. .302
Combined	65	76	74	4.10, .002	66	67	70	0.34, .743	67	63	71	−1.84, .093

Positive t-value indicates recency (most recent observations more predictive), negative t-value indicates primacy (most recent observations less predictive).

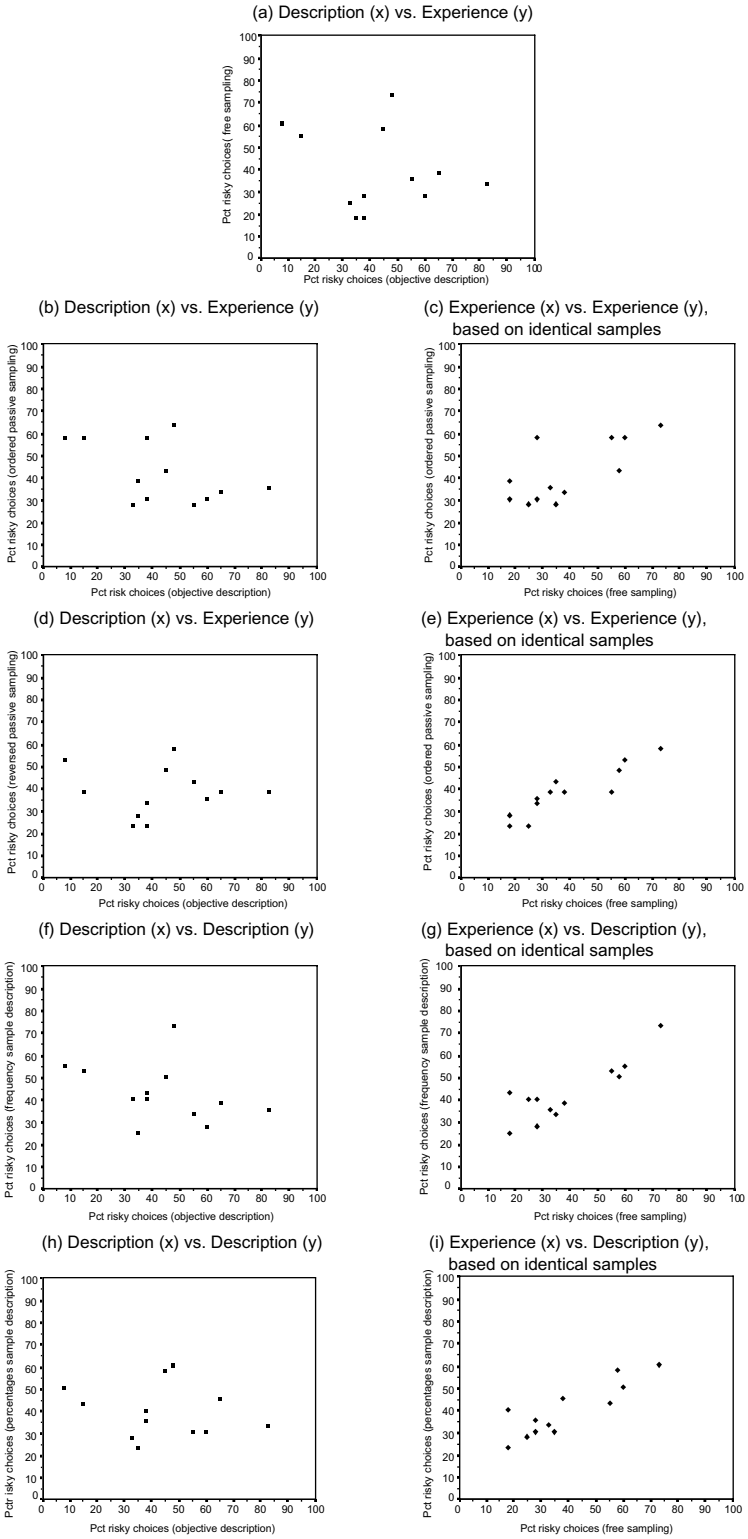

Fig. 2. Scatter-plots showing the match in patterns of choice between conditions. Each axis shows the percentage of risky choices, and each data point represents one of the 12 choice problems. Objective description (left column of plots) and free sampling (right column) are compared against all other conditions.

the information was described or experienced sequentially. When described, it did not matter whether frequencies (which preserve sample size information) or percentages (which do not) are used. When experienced through passive observation, the particular sequence of observations had no impact upon choice.

For the most part, it is the information about the choice problems that matters, not the mode of presentation. Consequently, calls for alternative descriptive theories for decisions from experience to those from description are overstated (Hertwig et al., 2004; Weber et al., 2004). This is confirmed by Fig. 2, which shows a series of scatter-plots that compare the proportion of risky choices between several pairs of conditions. If two conditions exhibit the same patterns of choice, then one would observe a strong positive correlation with points on the scatter-plot falling close to the identity line. This strong positive correlation would be expected if the same underlying mechanism applied to both conditions. The uppermost panel and the first two panels in the left-hand column show negative correlations between choice proportions and therefore imply very different patterns of choice for described and experienced conditions. However, the remaining panels show that two described conditions can also yield different patterns of choice (lower left pair of panels), and that both described and experienced conditions can yield very similar patterns of choice to those observed with free sampling (positive correlation between choice proportions, with points close to the identity line in the right-hand column of panels). Whichever theory describes the data in the free sampling condition, will also provide adequate description of the yoked conditions. Fox and Hadar (2006) demonstrated that prospect theory is one such model.

Some corollaries follow. First, it is misleading to say that small probabilities are underweighted in decisions from experience (Barron & Erev, 2003; Hertwig et al., 2004). This implies a cognitive adjustment. It is more appropriate to simply note that rare events are often observed less frequently than their long-run expected frequency. Second, it seems plausible that the mechanisms for choice based on sequential information acquisition should differ from those for choices made from full descriptions of the options. However, if patterns of choice are the same, once information has been aggregated it may be that the underlying processes for experience-based choice are quite similar to those for decisions from description.

The key thing to note is that when sampling bias is controlled by yoking, we do not observe any meaningful difference between described and experienced choices. In fact, choice concordance with the free sampling condition was slightly higher for the described yoked conditions than for the experienced yoked conditions. Therefore, if anything, choices in two experience conditions were less alike than choices in a described and an experienced condition—presumably because descriptions remove the noise in assessing the sample of observations prior to choice (see Fox & Hadar, 2006). One contributing factor to this similarity between yoked conditions, which can also explain the original difference between objective description and free sampling, is that small samples often create, or amplify, a difference between options which would otherwise be hard to choose between (Hertwig & Pleskac, 2008). For instance, taking a sample of five observations when there is a 10% chance of 32 points (as in Problem 5) will most frequently yield 32 points on 0% or 20% of occasions (i.e., never or once respectively). Therefore 0 points will be observed on 100% or 80% of the occasions. In comparison to the alternative safe option that always returns 3 points, a 100% chance of 0 points is trivially unattractive. On the other hand, a 20% chance of 32 points is quite seductive unless the decision maker is particularly risk averse. However, it cannot be the case that all choices derived from the free sampling condition were reduced to ones that participants regarded as trivial, as, otherwise, concordance rates would have been far higher than the 70% that was observed. Furthermore, it is important to recognize that the creation or amplification of differences between options is a property of small samples, not a property of decisions from experience per se. What our data show is that for any means of presenting information about a small sample—by description or by experience—the patterns of choice are very similar. The apparent description–experience gap in this free sampling paradigm is almost entirely explainable as a population-sample gap.

Rarely a recency phenomenon

We have been careful to conclude that biased sampling accounts for *most* of the description–experience effect. We did replicate the recency effect noted by Hertwig et al. (2004), but only in one of the three experience-based conditions. The penultimate row of Table 2 indicates that the discrepancy from the objective description condition was greater for the free sampling condition than for other conditions where recency was not found (passive sampling) or could not occur (sample description). Based on the ratio of mean effect sizes, up to one fifth of the difference between objective description and free sampling could be attributable to recency. (We say "up to" because some of the reduction in the effect in the yoked conditions could be due to regression effects.) Therefore, our data permit us to add detail to Hertwig et al.'s (2004) summary, that we quoted earlier: "rare events have less impact than they deserve *mainly* because decision makers have not encountered them, or have encountered them less frequently than expected, and, *sometimes, to a lesser extent,* because they have not encountered them recently."

Why is recency only observed under some conditions? One potentially relevant factor is that the information processing demands of free sampling exceed that of passive observation. In all conditions, participants keep

track of observations and form preferences—however, when free sampling, participants must *also* decide whether to continue sampling and which machine to sample next. Memory for instances may degrade more quickly when there is interference from other tasks, which would account for a greater impact of recent events. Consistent with this memory load hypothesis, Ungemach, Chater, and Stewart (2007) failed to find evidence of recency in experience-based choice when participants were relieved of the task of deciding when to stop sampling because the sample size was specified in advance. However, in comparison to the passive conditions, our free sampling condition had more correct predictions based on recent observations and no fewer correct predictions based on earlier observations. The recency effect is therefore more consistent with enhanced influence for recent events than with degraded memory for more distant events.

An intriguing alternative account is suggested by recent data on sequential sampling using the money-machine paradigm (Rakow & Rahim, in preparation). When a fixed number of observations to be made was specified, recency was observed for those participants who chose to alternate frequently between the options, but not for those who examined many payoffs from one machine before examining the other. This accords with the pattern of sequence-order effects in belief updating summarised by Hogarth and Einhorn (1992). Recency effects occur when evaluations are updated after each observation, yet primacy effects prevail when a single summary evaluation is made after a sequence of observations. Therefore, recency in sequential choice problems may be an expectancy-based phenomenon, rather than a memory-based one, which disappeared under passive sampling because, when yoked, participants were unable to form preferences in line with their habitual styles of processing.

Previous research on actor–observer differences in probability judgment suggests a further, possibly related, explanation. Actor–observer effects vary in size and direction with task characteristics (Harvey, Koehler, & Ayton, 1997; Koehler, 1994; Koehler & Harvey, 1997), though seem to occur dependably when the actor has control over the task and the observer does not (Harvey et al., 1997; Koehler & Harvey, 1997). Harvey et al. (1997) note that actors (e.g., active samplers in our experiment) "know their reasons for formulating particular decisions" (p. 268). This could apply to decisions about sampling—specifically, actors know the hypotheses that they are testing when they explore the options, whereas passive observers simply view this information. Thus, it could be that passive observers are constrained by difficulties in making sense of someone else's pattern of observations, rather than (or in addition to) active samplers being constrained by memory limitations. Certainly, this would account for the slightly lower predictability of observers' choices on the basis of EV in the passive sampling conditions (Table 3). This finding could also be because actors are likely to "develop a better understanding" of the task (Harvey et al., 1997, p. 268). For instance, some (active) participants may begin by exploring the two options to get the gist of what the two options are like, before going on to evaluate the options more formally. (For a model of decisions from experience that incorporates exploration, see Erev & Barron, 2005.) If initial observations are treated in this more "casual" manner—for instance, simply to determine whether options are similar enough to warrant formal evaluation—this would have the effect of discounting the impact of initial observations to some degree. This would be somewhat equivalent to considering a small sample of the more recent outcomes. This could be another reason why recency was observed when pre-decisional experience was active. In contrast, passive observers may simply seek to evaluate the options from the first observation. However, because idiosyncratic processes of exploration and evaluation drove the observations that they see, they may not find it so easy to aggregate this information as the (active) participant that they were yoked to. This would account for the lower predictability and the absence of recency among passive observers in the experience conditions.

In keeping with this actor–observer distinction, Newell and Rakow (2007) observed more optimal responding in a simple binary choice task when participants learned about outcomes through active prediction rather than by simple observation of outcomes. Even when outcome feedback was not given, there was some effect of active prediction—though, the effect was more marked when feedback was available. Clearly, all of the mechanisms that we discuss in relation to the transient nature of recency effects require independent testing in further experiments.

Consistent with Kareev's narrow window hypothesis (e.g., Kareev, 2000), participants with lower WM capacity collected less information—though this storage limitation account would have predicted an even stronger relationship for digit span (cf. Gaissmaier, Schooler, & Rieskamp, 2006), which was not the case. However, this effect was not strong enough to translate into a significantly enhanced D–E effect as would be predicted from greater bias for smaller samples. Furthermore, there was no evidence that the effect of recent observations was greater for those with smaller memory capacity—which further suggests that we may need to look beyond memory-based explanations for recency.

Key conclusions and further research

For many of the problems examined, we obtained large D–E differences consistent with previous research. However, the magnitude of the effects varied considerably between problems and we propose that research is now called for to determine what problem features moderate

the size of the D–E gap. However, the most important contribution of our data is that they demonstrate that sampling variability accounts for most, if not all, of the D–E gap in this free sampling paradigm. The difference between biased and unbiased samples was much more important than the difference between description and experience per se. Our participants responded to the information that they encountered—but the mode of presentation for this information (description vs experience) was of little or no consequence.

Recency effects can account for a small proportion of the D–E gap, though not in all situations. Further research is required to determine precisely when and why order effects in sequential choice should be expected. For instance, in real-world settings where observations are more spread out over time than in the laboratory, should we expect recency, primacy, or no effect of order? Furthermore, we have raised the possibility that recency effects, where they do occur, may not be memory-based phenomena. Rather, they may derive from the strategies that participants adopt in order to explore or accumulate evidence or to test hypotheses.

We have demonstrated that subtle changes to the description–experience paradigm (e.g., active vs passive observing) can have some impact upon the results (e.g., presence vs absence of recency effects). This is an important point to consider as research into decisions from experience develops. For instance, an alternative means of controlling sampling bias is to fix samples so that they match the objective probabilities of the problem. However, this requires participants to observe a particular number of samples—because, for instance, you cannot see a rare event 10% of the time if you choose to make six, seven or eight observations. Moreover, there is no reason to suppose that requiring someone to make ten observations is equivalent to someone choosing to make ten observations, let alone that tasks involving small samples of experience (e.g., Hertwig et al., 2004) are equivalent to those that involve large amounts of experience (e.g., Barron & Erev, 2003). The important next steps in research on decisions from experience are to identify the cognitive mechanisms that are associated with different kinds of experience, and to determine whether sampling bias is sufficient to explain the entire description–experience gap in these different situations.

References

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16*, 215–233.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*, 912–931.

Fox, C. R., & Hadar, L. (2006). "Decisions from experience" = sampling theory + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making, 1*, 159–161.

Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful? *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 966–982.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Gottlieb, D. A., Weiss, T., & Chapman, G. B. (2007). The format in which uncertainty information is presented affects decision biases. *Psychological Science, 18*, 240–246.

Harvey, N., Koehler, D. J., & Ayton, P. (1997). Judgments of decision effectiveness: Actor–observer differences in overconfidence. *Organizational Behavior and Human Decision Processes, 70*, 267–282.

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*, 534–539.

Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition*. New York: Oxford University Press.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief adjustment model. *Cognitive Psychology, 24*, 1–55.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econimetrica, 47*, 263–291.

Kareev, Y. (1995a). Positive bias in the perception of covariation. *Psychological Review, 102*, 490–502.

Kareev, Y. (1995b). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56*, 263–269.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107*, 397–402.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*, 278–287.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 461–469.

Koehler, D. J., & Harvey, N. (1997). Confidence judgements by actors and observers. *Journal of Behavioral Decision Making, 10*, 221–242.

Laibson, D., & Zeckhauser, R. (1998). Amos Tversky and the ascent of behavioral economics. *Journal of Risk and Uncertainty, 16*, 7–49.

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin and Review, 14*, 1133–1139.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review, 102*, 269–283.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297–323.

Ungemach, C., Chater, N., & Stewart, N. (2007). Decisions from experience without sampling error. *Paper presented at SPUDM 21 (Subjective Probability, Utility and Decision Making), Warsaw*.

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111*, 430–445.