*Judgment and Decision Making* (2024), **Vol. 19**:e2 1–29
doi:10.1017/jdm.2023.41

EMPIRICAL ARTICLE

# Repeated risky choices become more consistent with themselves but not expected value, with no effect of matched trial order

Jake Spicer[1], Timothy L. Mullett[2], and Adam N. Sanborn[1]

[1]Department of Psychology, University of Warwick, Coventry, UK and [2]Warwick Business School, University of Warwick, Coventry, UK

**Corresponding author:** Jake Spicer; Email: jake.spicer@warwick.ac.uk

**Received:** 27 October 2023; **Accepted:** 2 November 2023

**Keywords:** risky choice; consistency; internal noise; order effects

**Abstract**
Choices made in risky scenarios are considered fundamentally noisy because decisions have often been found to be inconsistent when repeated. Past measures of noise may, however, be confounded by the use of randomized contextual factors that are known to influence choice, in particular, the order of trials. In two experiments, we control trial order to test the extent to which inconsistent choice is attributable to changes in experimental context. Both tasks find strong evidence that trial order has no effect on choice consistency, indicating such experimental factors have little influence on behavior compared with internal noise. Choices also showed an increase in consistency across multiple repetitions, suggesting a fall in noise with experience, but this increase was not associated with any improvement in performance, with choices showing no greater adherence to either expected value or expected utility across repetitions. Instead, choices increasingly adhered to simplistic heuristic decision rules, possibly indicating greater reliance on such strategies as the tasks progressed. These results carry implications for a number of decision-making theories, including true-and-error models, rank-based methods, and strategy shift approaches.

## 1. Introduction

Much work in decision-making has examined decisions made under risk, investigating how people react to scenarios where outcomes are uncertain or unpredictable. A key finding emerging from this research is that such choices can be fairly inconsistent: when the same decisions are repeated, decision makers often make different selections, an effect demonstrated in studies extending back over 70 years (Camerer, 1989; Hey & Orme, 1994; Loomes & Pogrebna, 2014; Loomes & Sugden, 1998; Mosteller & Nogee, 1951; Rieskamp, 2008; Starmer & Sugden, 1989; Wakker et al., 1994). Such inconsistency has been a major factor in developing models of choice in this domain, indicating that deterministic processes do not provide adequate accounts for observed behavior. Many theories have thus assumed some level of noise in the decision-making process; for example, there may be stochasticity in the subjective valuation of potential outcomes that leads to different preferences between repetitions, or choices may be made probabilistically meaning that decisions differ even where the underlying preference is stable. While these are certainly plausible explanations, by attributing any inconsistency observed at the behavioral level to noise in the internal decision-making process, these theories risk potentially overlooking external sources of noise that may be present in such tasks. This is of particular concern given that the data used to support such models often comes from designs which are vulnerable to the influence of contextual factors which themselves have a similarly extensive literature of influencing decision-making. The present study therefore aimed to explore the impact of these external factors on choice consistency, so allowing for a more complete depiction of the influence of internal noise on decision-making. This focuses on empirical measures of behavior rather than the fitting of choice models, though such an examination does of course raise implications for existing theories, discussed further below.

While ‘external factors’ may refer to any element outside of the decision-making process of the individual which may differ between repetitions, we here focus on a particular aspect of risky choice which may introduce noise to decisions: trial order. Decision-making tasks often present trials in randomized orders, creating differences in immediate history between instances (Ballinger & Wilcox, 1997; Camerer, 1989; Hey & Orme, 1994; Loomes & Sugden, 1998; Wakker et al., 1994). Such variation is notable given that changes in the local context have a long history of affecting decision-making: numerous studies have found that decisions are not entirely independent of one another, but can be biased by preceding trials (Fründ et al., 2014; Garner, 1954; Holland & Lockhead, 1968; Lacouture, 1997; Stewart & Brown, 2004; Ward & Lockhead, 1970). This includes both attraction to past cases where stimuli are treated as more similar to their immediate predecessor (Garner, 1954; Hu, 1997), as well as contrast effects where stimuli are considered in comparison to recent items rather than in isolation (Holland & Lockhead, 1968; Lacouture, 1997; Vlaev et al., 2011; Ward & Lockhead, 1970). This is in fact particularly relevant in the case of risky choices as such effects can operate in both described prospects and experienced outcomes: choices may be colored by both recently viewed options as well as the results of prior selections if given. Indeed, existing research in this domain has found that behavior can be influenced by previously observed outcomes from related decisions (Leopard, 1978; Ludvig & Spetch, 2011; Ludvig et al., 2014), meaning any variation in outcomes between repetitions may lead to variations in behavior even where trial order is fixed. Trial-by-trial feedback thus presents another potential contextual source of noise, though such feedback is less commonly used in studies of decisions from description, so is unlikely to fully account for previous findings of inconsistency. Additionally, feedback is more difficult to effectively manipulate given the complications with anticipating actual choices; as such, we will not directly measure the impact of feedback in this study, though we do revisit this aspect in the discussion.

Recent work has suggested that contextual effects are attributable to the use of rank-based decision processes which consider target stimuli according to their relative rank in a given set (Stewart et al., 2006). This extends beyond the previously described sequence effects to the placement of a stimulus within its relevant distribution (Stewart et al., 2015): for example, a reward of £50 could be seen as high if other experienced values tended to fall between £10 and £20, but low if previous rewards tended to be over £100. Such a process has been offered as an explanation for sequence effects in categorization (Stewart et al., 2002), as well as errors in value identification (Stewart et al., 2005) and probability estimation (Ungemach et al., 2011). Choices thus appear to depend on the context provided by recent events, and so may differ where this context is inconsistent. The randomization of trial order (and indeed trial feedback) may then induce inconsistency in decisions through external means by varying the recent context between repetitions, meaning any tasks using such design elements risk confounding internal and external sources of noise.

The current study therefore aimed to test this suggestion by controlling for external factors as much as possible to examine the resulting impact on inconsistency in choice: if these controls lead to more consistent decisions, then behavioral displays of inconsistency may be due in part to variations in context; in contrast, if inconsistency remains, then such behavior may be driven more by internal noise. This involves contrasts between choices made in matched and unmatched contexts, where context is defined by the events of recent trials; in practical terms, this means repeating decisions in both

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

common and differing trial orders to allow for measures of the impact of matching order on consistency. Furthermore, decisions do not receive any feedback to eliminate any potential differences in outcomes between repetitions; while this does not allow measurement of the influence of feedback, this does remove this factor from consideration, and avoids the previously noted complications in attempting to anticipate participant choice. In addition, to further reduce any other sources of variation in this comparison, we focus on within subjects contrasts, asking each participant to repeat choices multiple times in matched and unmatched orders to avoid any individual differences between these conditions. Such comparisons can then offer a behavioral indication of the influence of these factors on decisions, allowing for a distinction between the impact of internal and external factors on behavior independent of the assumptions of any specific model of decision-making.

This approach then provides an opportunity to not only examine choice consistency according to the noted contextual factors, but also how consistency may evolve over time: previous theories have suggested that internal noise may be reduced with repetition, leading to more consistent decisions. This is most evident in so-called 'true and error' models in which preferences are assumed to be stable but perturbed by noise in response selection; repetitions then reduce this error, allowing decisions to better reflect underlying preferences (Birnbaum & Schmidt, 2015; Birnbaum et al., 2017). If, however, behavioral inconsistency also arises from external sources, then such repetition may not substantially reduce this effect so long as context continues to differ between decisions. Tracking the evolution of consistency across choices then offers a further examination of the influence of external noise on decision-making, contrasting these two competing hypotheses. To the best of our knowledge, such an approach has not often been used in previous examinations of noise in decision-making: decisions are rarely repeated more than once (Ballinger & Wilcox, 1997; Birnbaum, 2006, 2008, 2011; Busemeyer et al., 2000; Camerer, 1989; Hey & Orme, 1994; Loomes & Sugden, 1998; Starmer & Sugden, 1989; Wakker et al., 1994), and even where multiple repetitions are used, choice consistency is often measured in aggregate rather than between these repetitions, giving no indication of its evolution (Birnbaum, 2012; Birnbaum & Bahra, 2012; Guo & Regenwetter, 2014; Loomes & Pogrebna, 2014). One exception to this is Hey (2001), which repeated a set of decisions 5 times, though each time in a randomized order, to examine whether consistency increased with experience. This indeed found a general rise in consistency across repetitions, though there were substantial individual differences in this result, while comparisons were limited to immediately subsequent blocks only. The current paper then hopes to build on the results of that study, controlling for contextual factors to provide more detail on the evolution of consistency with experience.

This examination of the evolution of consistency across repetitions also connects to related questions on how the content of choices may change with experience; as noted above, true and error approaches would suggest that repeating choices will allow decision makers to better express their true underlying preference, thus settling on stable decision patterns (Birnbaum & Schmidt, 2015; Birnbaum et al., 2017). The current design can therefore test this prediction by examining whether choices do indeed converge to identifiable preferences across repetitions, and if so, what kinds of preference are displayed; for example, reductions in error may allow decision makers to more reliably select higher-value options, so leading to an increase in the optimality of choices with repetition. Alternatively, if consistency is increasing but optimality is not, then decision makers may be converging on strategies other than the value of the options, for example, focusing on safer prospects. Tracking the evolution of choices across repetitions may then allow for further novel insights into decision-making beyond consistency, potentially helping to reveal underlying strategies used by decision makers, though again, our primary focus is on empirical measures of behavior to avoid limiting results to certain model assumptions.

We therefore continue this paper by presenting two experiments in which participants were asked to make repeated choices between monetary gambles across several repetitions in either matched or unmatched trial orders. We use these tasks to investigate three main aspects of decision-making introduced above: first, the impact of these contextual factors on choice consistency; second, the evolution of consistency across repetitions; and third, the potential decision strategies which participants may use to direct their choices.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

## 2. Experiment 1

### *2.1. Method*

#### 2.1.1. Participants

A total of 166 participants were recruited through the University of Warwick online SONA system in return for financial compensation, composed of a base payment of £4 plus a bonus determined via choices made during the task, detailed below. Sample size was set according to the maximum number permitted by the experimental budget. The experiment was approved by the University of Warwick Humanities and Social Sciences Research Ethics Committee.

#### 2.1.2. Design

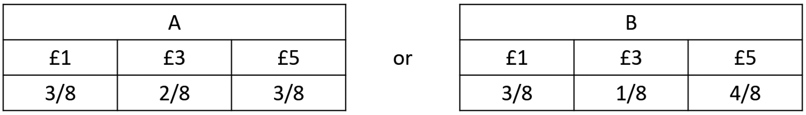
The experiment consisted of a series of choices between monetary gambles: in each trial, participants were presented with two potential gambles, each having 3 matching outcomes with predefined probability. Gamble pairs were based on the set used by Hey (2001), selected due to its previous use in examining the consistency of choice with repetition. These gambles were generated by drawing 3 outcomes from a pool of 4 potential values; for the present experiment, these values were −£1, £1, £3, and £5. Fifty choices were used in the task, with 25 excluding the −£1 outcome and 25 excluding the £5 outcome; half of the choices were therefore purely gains and half were mixed between gains and losses. Of these 50 pairs, 3 were 'dominated' cases in which the value of one gamble was clearly superior to the alternative, illustrated in Figure 1; these dominated pairs acted as catch trials in the subsequent analysis, indicating whether participants were selecting gambles at random rather than according to their value.

As in the original gamble set of Hey (2001), outcome probabilities were limited to increments of 1/8 to aid interpretation and comparison; these probabilities were, however, presented as explicit numerical fractions rather than the segmented circles used by Hey (2001) to avoid perceptual noise in the interpretation of these probabilities. This then reduces any potential differences in perception between subjects, but also prevents the possibility that displayed probability values may be interpreted differently on repetitions of the same decision, which could be one cause for inconsistency in choices. A list of these gambles is provided in Appendix A.

The set of 50 unique gamble pairs comprised 1 block of the experiment, with each session being composed of 4 such blocks, meaning each gamble pair was viewed 4 times by each participant. The key manipulation of the experiment was trial order: each session involved 2 potential orders of the 50 trials, with each order being used in 2 of the 4 experimental blocks. Eight potential trial orders were randomly generated prior to the experiment, organized into 4 block pairs (AB, CD, EF, and GH), with each participant being randomly assigned one of these block pairs at the start of the session to control for any influence of a particular trial order. The order of these blocks was fixed to alternate between the 2 preset trial orders, meaning that trial order in block 1 matched trial order in block 3, and trial order in block 2 matched trial order in block 4 (e.g., ABAB).

#### 2.1.3. Procedure

Upon arriving at the laboratory, participants were first randomly assigned one of the 4 block pairs, determining the trial orders that would be used in the subsequent test blocks. Participants were instructed that they would be asked to choose between monetary gambles, and should pick the gamble which they preferred using a respective key on the keyboard; at the end of the experiment, one of their



| A | | |
|---|---|---|
| £1 | £3 | £5 |
| 3/8 | 2/8 | 3/8 |

or

| B | | |
|---|---|---|
| £1 | £3 | £5 |
| 3/8 | 1/8 | 4/8 |

***Figure 1.*** *Sample slide from Experiment 1, illustrating a dominated gamble pair.*

choices would be randomly selected, and the chosen gamble played out to provide an additional reward payment, thus incentivizing participants to always select their preferred option in each trial rather than attempting to balance risk across their choices. Choices made during the task itself were not, however, followed by any feedback regarding outcomes for the selected gamble, so removing any randomization that may arise between repetitions from such feedback.

Each session began with a set of 5 practice trials common to all treatments to introduce the task before beginning the first block. As stated above, each experimental session involved 4 blocks of 50 choices, for a total of 200 trials.

After completing all trials, the computer then randomly selected one of the 200 trials, ran the gamble chosen by the participant on that trial, and added the resulting outcome to the participant's payment as a bonus. Participants were then paid and debriefed as to the aims and expectations of the study.

#### 2.1.4. Transparency and openness

Data collected in Experiment 1, as well as the files used in the analyses described below, are available on the Open Science Framework (Spicer et al., 2020), while key experimental materials are provided in Appendix A. The design and analyses of Experiment 1 were not preregistered.

### *2.2. Results and discussion*

Data were first filtered to remove any participants who did not appear to be paying attention to the task using choices in the dominated gamble pairs. The exclusion criterion was set to remove participants who were not significantly above chance-level selection across the 4 repetitions of the 3 dominated trials, so requiring a dominant choice rate of 0.83 (10/12); this excluded 11 participants, leaving 155 for further analysis. Following this filtering, the dominated trials were removed from further analysis given that the data were gated for consistency in these choices, meaning all subsequent measures were taken across the remaining 47 gamble pairs. Choices in these dominated gambles are however further examined in Appendix C.

#### 2.2.1. Consistency

Consistency was measured as a simple binary variable indicating whether identical choices were made in each potential comparison between 2 repetitions of the same gamble pair; the 4 repetitions of each pair thus provide 6 potential pairwise comparisons, being either matched or unmatched in trial order. Average consistency scores are summarized in Figure 2, showing an overall mean rate of 80.7% common choices across all comparisons, falling slightly below the mean consistency rate of 89.8% observed by Hey (2001) using an expanded version of this gamble set (though that study only considered comparisons between adjacent blocks).

Consistency was investigated using a mixed model logistic regression with the factors of match in trial order, distance between repetitions, the inclusion of the second, third and fourth blocks in a given comparison (representing position in the experiment, with the first block as a baseline), and absolute difference in expected value between the two gambles of a given trial (representing the 'difficulty' of choice for that pair). Given that no interactions between these factors were assumed *a priori*, no interactions were included in the model. A mixed model was used given the within-subjects nature of the design, with participant ID acting as a random factor to account for potential individual differences in consistency. A Bayesian version of this regression model was also performed to provide 95% credible intervals on the coefficient estimate of each factor, as well as Bayes factors ($BF_{10}$) measuring relative evidence that the true coefficient for each factor is or is not 0, where higher values indicate stronger evidence against this null hypothesis. This can, however, lead to disagreements between frequentist and Bayesian results due to greater conservatism in the Bayesian prior; while such cases likely reflect differences in prior definition rather than the actual findings, we avoid strong conclusions in such cases as a precaution. More detail on the Bayesian regression is provided in Appendix B.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

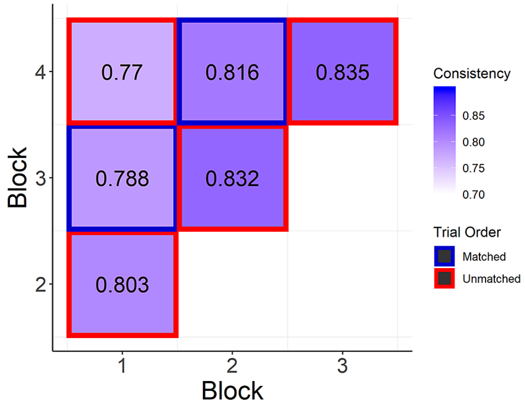


***Figure 2.*** *Average consistency measures between the 4 blocks of Experiment 1. Colored outlines denote match in trial order.*

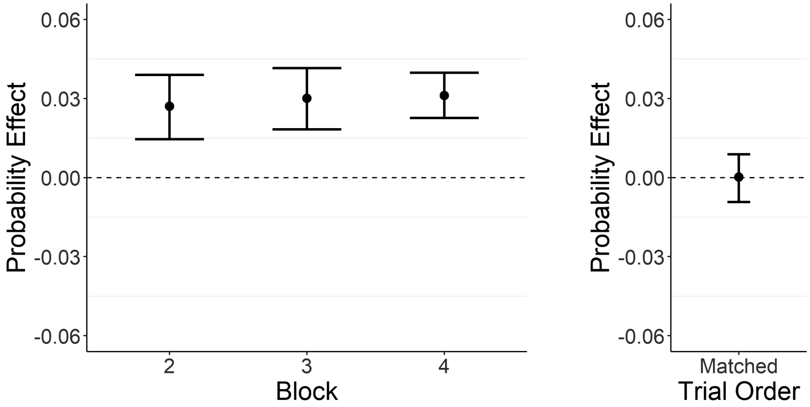


***Figure 3.*** *Predicted effect of block factors (relative to block 1) and match in trial order on the probability of choice consistency in Experiment 1. Error bars indicate 95% credible intervals taken from the Bayesian version of the model.*

Distance was found to be a significant predictor ($\beta = -0.172$, $z = 3.78$, $p < 0.001$, $BF_{10} = 7.90$), with consistency falling as distance between blocks increased. Absolute difference in expected value also had a significant effect ($\beta = 0.202$, $z = 7.97$, $p < 0.001$, $BF_{10} = 28{,}769$), with choices being more consistent with larger differences in expected value. Consistency was also found to be significantly higher when comparisons included the second ($\beta = 0.180$, $z = 4.10$, $p < 0.001$, $BF_{10} = 37.0$), third ($\beta = 0.201$, $z = 4.61$, $p < 0.001$, $BF_{10} = 48.2$) or fourth block ($\beta = 0.208$, $z = 6.82$, $p < 0.001$, $BF_{10} = 4{,}995$) relative to the first block, suggesting a role of experience: as participants progressed through the task, their choices became more consistent, with the first block essentially defining the reward space (summarized in Figure 3). This was further examined with a secondary logistic regression restricting data to only repetitions at one block distance (i.e., 1 vs. 2, 2 vs. 3, and 3 vs. 4), providing a clearer representation of position in the task, though this is at the cost of a substantial reduction of the data, being only 50% of the collected trials. Here, consistency was found to significantly increase with comparison order ($\beta = 0.117$, $z = 5.28$, $p < 0.001$, $BF_{10} = 188$), suggesting choices did become more stable with experience[1]. There was no significant effect of match in trial order, however, ($\beta = 0.001$, $z =$

[1] This increase appears to be primarily driven by a particularly sharp step following the earliest comparison; we examine this in more detail in Appendix D.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

0.03, $p = 0.976$, $BF_{10} = 0.01$), with the Bayesian analysis in fact showing strong evidence that choices were no more consistent between blocks with matching trial sequences (also illustrated in Figure 3).

Consistency does not then appear to depend on trial order, implying that such contextual differences were not a major source of variability in this task. Consistency does, however, seem to increase with repetition, matching with the results of Hey (2001) using the same choice set in a similar design. We therefore next explore the possible reasons for this increase by tracking the content of participant choices across trial blocks.

### 2.2.2. Choice

Choice content was assessed to determine whether the increase in consistency in this task was associated with any increased adherence to particular decision strategies with experience; while this is not necessarily a direct relationship, any shift toward a certain strategy might result in more stable decisions and therefore greater consistency. There are of course many existing models of such choices, creating an extensive set of potential decision rules that could be considered. While a full comparison of such models is outside of the scope of this paper, requiring more targeted methodologies (Jekel et al., 2011), we here seek to provide an initial exploration of this question by examining some basic models based on features of the gamble set rather than the decision maker. The following comparison therefore considers the rational baseline of expected value alongside a set of heuristic decision rules selected from the existing literature. Such a focus also helps to avoid possible complications raised by more complex decision models including inferred internal parameters which may differ either between participants or between repetitions; the fitting of these models requires assumptions on whether variations in behavior are attributable to variations in model parameters across the task (if fitting blocks separately) versus higher-level selection or imperfect execution of a stable rule (if fitting blocks together). In contrast, the rules considered here each make identical predictions across participants and blocks, preventing such differences and further simplifying the comparison. This analysis then measures general adherence to these fixed rules across the task, avoiding restrictions to a particular interpretation of any change in adherence; increased adherence for a given rule could reflect more reliable selection of that precise rule, a reduction in noise for that rule, or greater weight applied to a related feature in a more complex rule. We do, however, include one exception to this in the form of subjective expected utility given the importance of this function in past studies of risky choice (Camerer, 1989; Hey, 2001; Hey & Orme, 1994; Loomes & Pogrebna, 2014; Mosteller & Nogee, 1951; Rieskamp, 2008; Wakker et al., 1994). To be clear, the purpose of this analysis was not to find the best fitting model for behavior in this task, but rather to examine any relation between these strategies and the above findings on consistency; as such, we focus here on the change in adherence to a given rule across repetitions rather than the absolute level of adherence.

Heuristics were taken from Brandstätter et al. (2006), summarizing several approaches to decision-making offered in the existing literature. Not all of these heuristics can be applied to the current data set, however, as several of these rules depend solely on differences in outcomes which do not exist in the matched outcome gamble pairs of Hey (2001); we therefore restrict attention to only those rules able to predict preferences for at least a subset of these pairs. Participant choices were coded into measures of adherence to each rule in each trial, defined below for clarity, being 1 where responses followed predictions and 0 where responses opposed predictions. For trials in which a rule was not able to produce a preference, adherence was coded as 0.5, reflecting an indifferent value; this was included to account for rules which were only able to predict a subset of trials, which are thus drawn toward the middle of the scale. Each rule was then analyzed using a separate mixed model linear regression examining average adherence rates across gamble pairs from each participant for the described rule over the 4 trial blocks, with participant being a random effect to account for individual differences. As with the consistency analysis, Bayesian versions of each regression were also used to provide Bayes factors to allow for assessment of support for the null hypothesis of no effect for each factor, though different prior definitions were used in this case given that these were linear rather than logistic regression models (see Appendix B for further details). In addition, we also report the proportion of participants

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

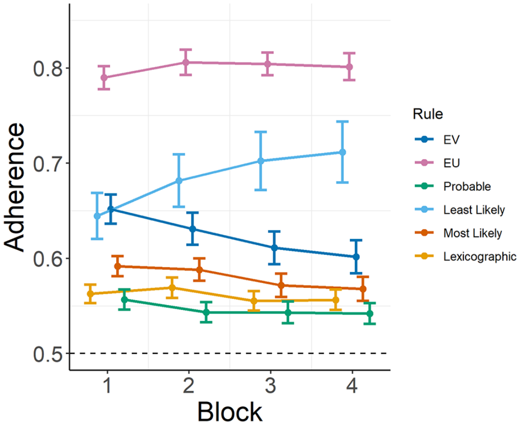


***Figure 4.*** *Mean adherence rates to the considered decision rules across Experiment 1. Error bars show 95% CIs, while the dashed line indicates random selection.*

whose random effect slope for the block factor was positive/negative to help characterize individual differences between subjects in adherence change, though it should be noted that this measure is unfortunately insensitive to the magnitude of individual slopes. Average adherence rates for these rules across participants in Experiment 1 are illustrated in Figure 4.

#### *2.2.2.1. Expected value*

Expected value was defined using the average of outcomes weighted by their stated probabilities:

$$E[x] = \Sigma_i p_i x_i, \tag{1}$$

where $i$ reflects each potential outcome of gamble $x$. Use of this feature is then coded as selection of the gamble with higher expected value for that pair, reflecting idealized rational behavior; this produced a preference for 35 of the 47 gamble pairs where EV was not equal. Participants showed lower adherence to expected value as the task progressed ($\beta = -0.017$, $t(154) = 7.82$, $p < 0.001$, $BF_{10} > 10{,}000$), with 87.7% of participants having a negative slope, suggesting choices were not improving with repetition by this metric.

#### *2.2.2.2. Expected utility*

Expected utility was also examined to account for potential differences between objective and subjective valuations of outcomes that could impact expected value. Unlike the other considered rules, expected utility held free parameters which were able to vary between participants, though as our focus here is on the change in adherence across blocks rather than absolute fit, this difference does not affect comparisons with the other rules. Utility was defined using a power-law transformation of outcome value:

$$E[u(x)] = \Sigma_i sign(x_i) p_i |x_i|^{\alpha_j}, \tag{2}$$

where $\alpha_j$ reflects participant $j$'s inferred sensitivity to changes in value. Best fitting parameters were estimated for each participant via maximum likelihood search using their collected choices across the whole experiment[2]. This provided individual parameter values for all but 4 participants who could not be reliably fit, and hence, were excluded from analysis of this particular rule. The mean fitted $\alpha$

[2] For fitting purposes, a softmax function was applied to the estimated expected utility of each gamble to predict choice probabilities for calculation of likelihoods, though we do not include this function in the adherence measure examined here as this creates a discrepancy with the other considered rules; adherence is thus defined as selection of the option with higher expected utility given the best fitting utility function of that participant.

parameter across participants was 0.54 (±0.11 95% CI), producing preferences for all 47 gambles for all but 1 participant whose $\alpha$ value was especially low (<0.001).

As shown in Figure 4, subjective expected utility notably held the highest absolute adherence level of the considered rules. This did not, however, show any reliable difference over the course of the task, though Bayes factors indicate no strong evidence toward a slope of 0 either ($\beta = 0.003$, $t(150) = 1.51$, $p = 0.134$, $BF_{10} = 0.403$), leaving results for change in adherence to this rule ambiguous. It is, however, notable that 69.5% of participants did show a positive slope at the individual level, potentially implying some shift toward this function.

#### *2.2.2.3. Probable*

The probable rule classifies outcomes as 'probable' or 'improbable' based on whether their probability is above or below the average probability for the total number of outcomes in that gamble (in this case, 1/3). Options are then valued according to the mean of the 'probable' outcomes only, with the higher mean being preferred. In the current set, this generates a preference for 19 of the 47 pairs. Participants showed a significant reduction in adherence to this rule across the task ($\beta = -0.004$, $t(154) = 3.03$, $p = 0.003$, $BF_{10} = 30.6$), with 77.4% of participants showing a negative slope, suggesting use of this heuristic does not explain the observed increase in consistency in the data.

#### *2.2.2.4. Least likely and priority heuristic*

We consider the least likely rule and the priority heuristic jointly here as these two rules make equivalent predictions for the present gamble set. The least likely rule selects the option which has the lower probability between each gamble's respective lowest outcomes. The priority heuristic meanwhile sequentially proceeds through multiple criteria, beginning with minimum outcome, then probability of minimum outcomes, and finally maximum outcome, stopping if the difference between options exceeds a preset threshold (here being either 10% of the maximum outcome for that trial or 0.1 probability as in the original definition by Brandstätter et al., 2006) and selecting the winning option. As the present gamble pairs are matched in outcomes, the priority heuristic is only able to make predictions based on differences in the probability of minimum outcomes, thereby matching the least likely rule. Furthermore, because probabilities in this set are limited to increments of 1/8, any non-zero differences between these probabilities will exceed the standard threshold, meaning the priority heuristic is unaffected by the scale of these differences and so matches the least likely rule in all cases. This means that both rules produce a preference for all non-dominated gamble pairs as both options have a common lower outcome which necessarily differs in probability. Choices in Experiment 1 demonstrated increased adherence to these rules across repetitions ($\beta = 0.022$, $t(154) = 6.51$, $p < 0.001$, $BF_{10} > 10,000$), with 78.1% of participants showing a positive slope, suggesting an apparent increase in safety-seeking behavior across the task as participants move toward options which are more certain to avoid lower outcomes.

#### *2.2.2.5. Most likely*

The most likely rule compares the outcomes with highest probability from each gamble, suggesting a preference for the higher value. This produces a prediction for 29 of the 47 gamble pairs due to the common outcome structure of this set. Adherence to this rule was found to significantly fall across blocks ($\beta = -0.009$, $t(154) = 5.14$, $p < 0.001$, $BF_{10} > 10,000$), with 82.6% of participants showing a negative slope, implying participants in fact shifted away from this strategy.

#### *2.2.2.6. Lexicographic*

The lexicographic rule extends the most likely rule above to account for possible equalities between most likely outcomes: if this is unable to identify a preference, comparisons are then made between the second most likely outcome from each gamble, continuing until a prediction is made. This then allows for predictions in a wider range of trials than the most likely rule, though this still does not account for all pairs in the set (37 of 47 pairs). Participants showed no significant change in adherence to the

***Table 1.*** *Choice rule analysis results from Experiment 1.*

| Rule | Prediction rate | $\beta$ | $t$ | $p$ | $BF_{10}$ | Positive rate |
|---|---|---|---|---|---|---|
| Expected value | 74.5% | −0.017 | 7.82 | <.001 | >10000 | 12.3% |
| Expected utility | 99.7% | 0.003 | 1.51 | .134 | 0.403 | 69.5% |
| Probable | 40.4% | −0.004 | 3.03 | .003 | 30.6 | 22.6% |
| Least likely and priority heuristic | 100% | 0.022 | 6.51 | <.001 | >10000 | 78.1% |
| Most likely | 61.7% | −0.009 | 5.14 | <.001 | >10000 | 17.4% |
| Lexicographic | 78.7% | −0.003 | 1.94 | .054 | 0.974 | 32.3% |

*Note*: Prediction rate gives the proportion of gamble pairs for which that rule provides a preference, $\beta$, $t$, $p$, and $BF_{10}$ give regression results for adherence to that rule over blocks, and positive rate gives the proportion of participants showing a positive slope at the individual level.

lexicographic rule across the task according to frequentist tests, while Bayesian tests suggest ambiguous evidence for this factor ($\beta = -0.003$, $t(154) = 1.94$, $p = 0.054$, $BF_{10} = 0.974$), with 67.7% of participants showing a negative slope, though both results would suggest that reliance on this heuristic is unable to explain the observed consistency effects.

#### *2.2.2.7. Collected choice results*

Comparisons of these choice functions with behavior thus find that the best explanation for the observed increase in consistency in this task is via increased adherence to risk averse decision rules which avoid prospects with higher chances of lower outcomes. All other candidate rules meanwhile show either decreasing or flat adherence functions, meaning use of these rules in isolation cannot account for increased consistency. This is of course far from an exhaustive comparison of potential decision strategies, but does offer some insight into the driving forces of decision-making in this task, with aversion to risk emerging as a key aspect motivating choices. Furthermore, it is especially notable that the traditional 'rational' choice functions of expected value and expected utility did not demonstrate reliable positive slopes here, as this suggests decisions were not improving according to conventional measures.

### *2.3. Summary*

Data from this experiment suggest decisions under risk are more consistent when separated by less time and with greater experience with the choice set, but not when choices are presented in a consistent order. There is, however, a confound in this design: because the block sequence always alternated between the two potential orders, matching trial orders were always separated by a distance of two blocks, meaning any potential effect of trial order may have been masked by this distance. We therefore sought to address this issue in a second experiment which separated these two effects from one another to provide more confidence in these findings.

## 3. Experiment 2

Experiment 2 extended the design of Experiment 1 to allow for different block sequences in which matched trial orders vary in distance from one another to provide greater confidence that this has no effect on choice consistency. This also provided an opportunity to expand on the role of experience observed in Experiment 1: given that choices are seemingly more consistent in later blocks, extending the design to include additional blocks may further increase this effect, or reveal a cut-off point. In addition, this also allows for generalization beyond the specific gamble types of Hey (2001) used in the first experiment: these gambles were selected given their previous use in examining choice consistency in risky decisions, but do differ from gambles more commonly used in related research. This second experiment then replaced these gambles with an alternate set to provide a greater variety of outcomes and probabilities.

## *3.1. Method*

### 3.1.1. Participants

A total of 29 participants were recruited through the University of Warwick online SONA system in return for financial compensation, composed of a base payment of £15 plus a bonus determined via choices made during the task, detailed below. Sample size was again set according to the maximum permitted by the experimental budget. The experiment was again approved by the University of Warwick Humanities and Social Sciences Research Ethics Committee.

### 3.1.2. Design

As in Experiment 1, Experiment 2 consisted of a series of choices between monetary gambles, though here these gambles came from a different source with a wider range of both outcomes and probabilities. A total of 75 gamble pairs were taken from Glöckner and Pachur (2012), themselves taken from multiple studies on decisions under risk; this set was selected to provide a representative sample of the choice problems commonly used in the risky choice literature to assess whether the consistency effects observed in Experiment 1 would generalize outside of the specific gamble pairs used by Hey (2001). Unlike Experiment 1, all gambles in Experiment 2 involved only two outcomes, and probabilities were expressed as decimal numbers rather than fractions due to the higher variety of values. Gambles were again either purely gains or mixed between gains and losses, though gamble pairs no longer shared common outcomes. The set also included 2 dominated gamble pairs again to detect for random responses during analysis. These gambles are listed in Appendix A.

Participants again completed multiple blocks of the gamble set, being in one of 2 orders. Eight potential orders of the set were randomly generated and organized into 4 block pairs, with each participant viewing only 1 block pair throughout the experiment. In contrast to Experiment 1, however, Experiment 2 involved 3 sessions per participant, with each session being composed of 4 blocks, again including 2 of each assigned trial order. This allowed for 3 different potential block sequences for each participant, with each being used in one of their 3 sessions. The position of blocks with matched trial orders could therefore vary between sessions: for example, a participant may have viewed the sequence ABAB in their first session, AABB in their second, and ABBA in their third. While participants each viewed all 3 potential sequences of their particular block pairs across their 3 sessions, the order of these sequences across sessions was randomized.

### 3.1.3. Procedure

Experiment 2 followed a similar procedure to Experiment 1 with the addition of multiple sessions for each participant: participants were required to attend 3 separate sessions of the experiment on 3 different days across 1 week, with each individual session being similar to those of Experiment 1.

The first session began with participants being assigned one of the 4 block pairs, determining the trial orders that would be viewed in each block, and setting the sequence of these blocks in each session. Participants then completed a set of practice trials to introduce the task before beginning the main trial set. These practice trials were included in all 3 sessions to maintain consistency in procedure between sessions. Participants then ran through 4 blocks of the 75 choices, for a total of 300 trials per session. As with the previous experiment, choices were not followed by any feedback regarding gamble outcomes to eliminate any variations between repetitions.

After completing their third session, one of the participant's choices from any of their 3 sessions was again randomly selected and the chosen gamble played out to provide a bonus payment. Because outcome values were substantially higher for this gamble set compared to the previous experiment, bonuses were in this case rescaled from the listed outcomes for that gamble to be 1.5% of the stated value; participants were informed of this at the start of their sessions. Participants were then debriefed as to the aims and expectations of the study.

### 3.1.4. Transparency and openness

As with Experiment 1, both data from Experiment 2 and the files used in the subsequent analyses are available on the Open Science Framework (Spicer et al., 2020), with key experimental materials again provided in Appendix A. The design and analyses of Experiment 2 were also not preregistered.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

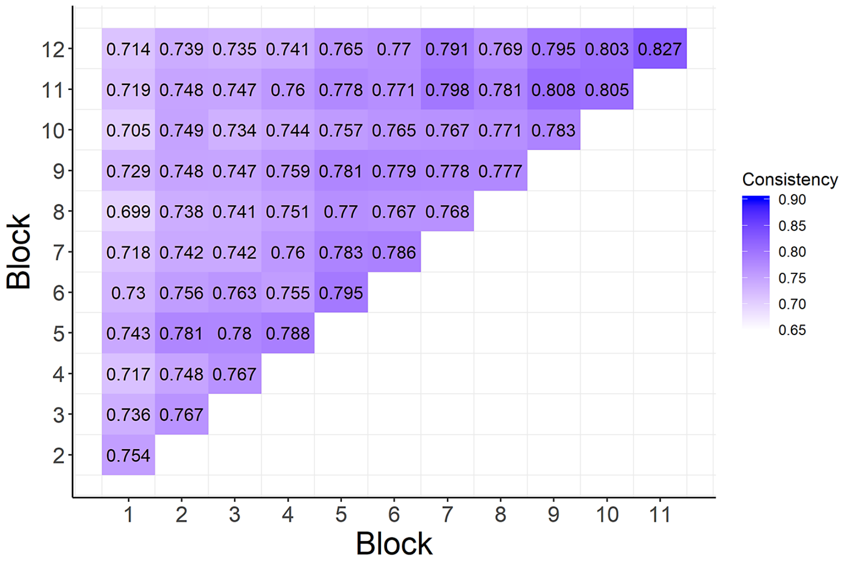


***Figure 5.*** *Average consistency measures between the 12 blocks of Experiment 2.*

### *3.2. Results and discussion*

Data were again filtered according to responses on the dominated gamble pairs to remove any participants who appeared to be responding randomly. The criterion for exclusion was set to be equivalent to the proportion used in Experiment 1, being a dominant choice rate of 0.83 (here 20 of 24). This excluded only 1 participant, leaving 28 in the main analyses. Dominated choices were again removed following this filtering, though these choices are also examined further in Appendix C. In addition, all responses for gamble pair 74 were also excluded due to an error in listed probabilities, leaving 72 gambles in subsequent tests.

#### 3.2.1. Consistency

The 3 sessions of 4 blocks in Experiment 2 produces 12 repetitions of each choice, and therefore, 66 potential pairwise comparisons of response consistency; these values are summarized in Figure 5. Consistency levels were reasonably similar to those observed in Experiment 1, with a mean of 76.0% across all comparisons, with the slight reduction between tasks likely being due to the higher number of comparisons at longer distances. This also places consistency at a similar level to the 79% found by Glöckner and Pachur (2012) using a subset of the same choice set across a single repetition.

Experimental data were again analyzed using a mixed model logistic regression analysis examining consistency according to match in trial order, distance between blocks, inclusion of each block in a comparison as a proxy for experience (with the first block acting as the baseline), whether the comparison was within or between sessions, and absolute difference in expected value of each choice (representing choice difficulty). Participant ID was again included as a random factor to account for potential individual differences in consistency between subjects. To aid regression, expected values were rescaled to correspond with payment values rather than display values (being 1.5% of their listed value) to place coefficient estimates at a similar magnitude to other factors. As with Experiment 1, a Bayesian version of this model was also used to provide credible intervals and Bayes factors for each coefficient estimate.

As in the previous experiment, choices were found to be less consistent with greater distances between blocks ($\beta = -0.037$, $z = 9.02$, $p < 0.001$, $BF_{10} = 86{,}558$) and with less experience in the task (summarized in Figure 6). As with Experiment 1, we further analyzed this experience effect with an additional logistic regression using only comparisons between adjacent blocks; this again found a

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

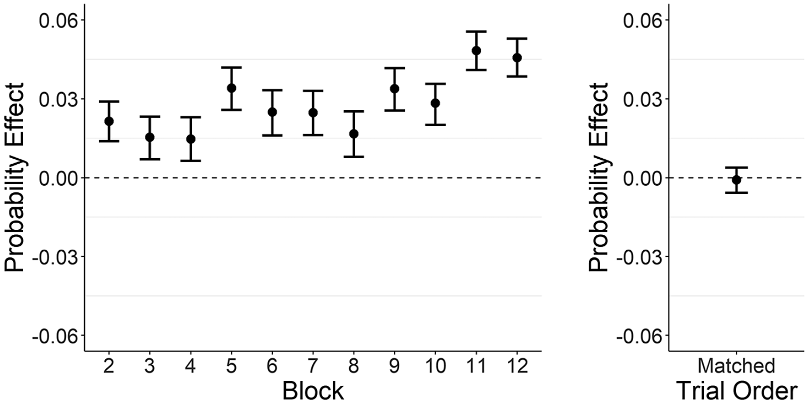


***Figure 6.*** *Predicted effect of block factors (relative to block 1) and match in trial order on the probability of choice consistency in Experiment 2. Error bars indicate 95% credible intervals taken from the Bayesian version of the model.*

significant increase in consistency over the course of the task ($\beta = 0.029$, $z = 5.46$, $p < 0.001$, $BF_{10} = 223$). Choices were also again more consistent for gambles with larger differences in expected value according to frequentist measures, though the Bayesian model suggests greater support for no effect of this factor ($\beta = 0.013$, $z = 2.34$, $p = 0.019$, $BF_{10} = 0.082$). Consistency was not, however, found to significantly vary between comparisons within or between sessions ($\beta = -0.001$, $z = 0.41$, $p = 0.685$, $BF_{10} = 0.008$), with Bayesian results in fact suggesting strong evidence against any distinction between sessions. Crucially, match in trial order was again found to have no significant effect on consistency ($\beta = 0.005$, $z = 0.38$, $p = 0.706$, $BF_{10} = 0.005$), again showing strong evidence for a null result according to Bayes factors.

### 3.2.2. Choice

Choices in Experiment 2 were coded according the same decision rules used in Experiment 1, defined in the same manner as the previous experiment, and were again analyzed using separate mixed linear regression models tracking adherence to each decision rule described above over the 12 trial blocks. In addition to the 6 rules listed above, however, we also examine a further set of heuristics based upon gamble outcomes that could not be applied to the previous experiment due to the structure of that gamble set. These rules were again taken from Brandstätter et al. (2006), collecting several existing approaches to risky decision-making. Average adherence rates from Experiment 2 are illustrated in Figure 7, separated between those applied to Experiment 1 and those novel to Experiment 2.

#### *3.2.2.1. Expected value*

Expected value again provides a rational baseline expectation for choice, making predictions for 65 of the 72 gamble pairs in this set. In contrast to Experiment 1, choices in Experiment 2 showed no significant change in use of expected value across the task ($\beta = -0.001$, $t(27) = 0.96$, $p = 0.345$, $BF_{10} = 0.258$), implying choices were not becoming less optimal in this case, though performance was also notably not improving either, and 71.4% of participants did show negative slopes.

#### *3.2.2.2. Expected utility*

Expected utility was again approximated using a power-law transformation of outcome value as described above, fitted to each participants collected choices across the experiment. The mean fitted $\alpha$ value was 0.51 (±0.08), with preferences for all but 1 gamble for all participants. Similar to the previous experiment, while expected utility again held the highest absolute adherence level of the considered set, evidence for a change in adherence across blocks was ambiguous ($\beta = -0.002$, $t(27) = 1.15$, $p = 0.261$, $BF_{10} = 1.67$), though in this case results in fact point toward a fall in adherence, with 67.9% of

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

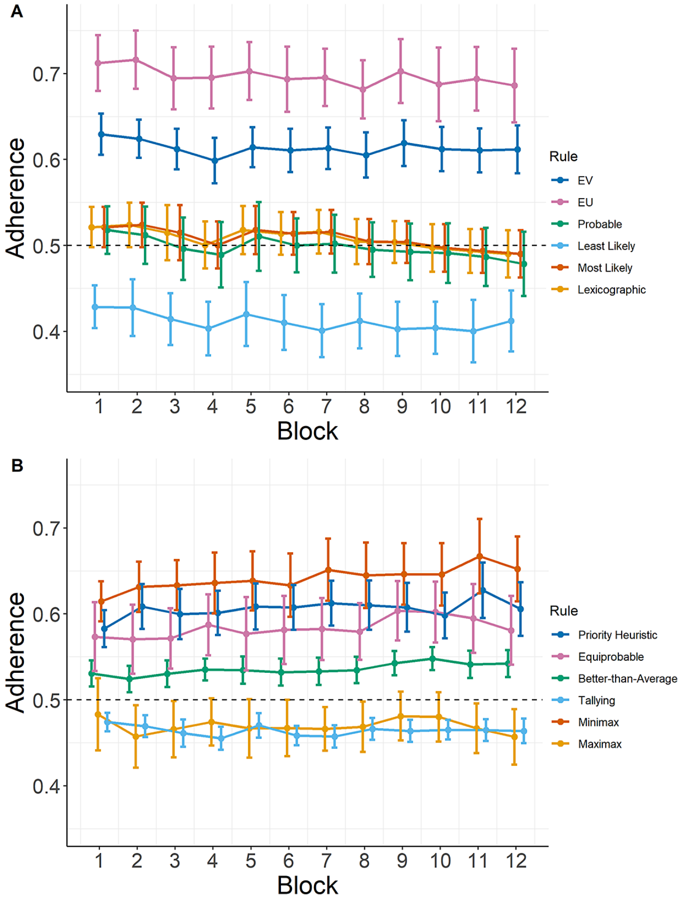


***Figure 7.*** *Mean adherence rates to the decision rules in Experiment 2. Panel A shows the rules previously evaluated in Experiment 1, whereas panel B shows the rules new to Experiment 2. Error bars show 95% CIs, while the dashed line indicates random selection.*

participants showing a negative slope. While these results are not conclusive, this does imply that the observed increase in consistency cannot be attributed to increased adherence to expected utility.

### *3.2.2.3. Probable*

In contrast to Experiment 1, the wider range of outcomes and probabilities in this gamble set allows the probable rule to generate a preference for 71 of the 72 pairs. Similar to expected utility above, participants showed a near-significant fall in adherence to this rule across the task by frequentist measures, with strong evidence of an effect by Bayesian measures ($\beta$ = –0.003, $t(27)$ = 1.90, $p$ = 0.069, $BF_{10}$ = 14.0), with 71.4% of participants showing a negative slope, meaning use of this heuristic is unable to explain the observed consistency effects.

### *3.2.2.4. Least likely*

The least likely rule predicted preferences for 58 of the 72 pairs in Experiment 2, being unable to account for cases in which minimum outcomes from each gamble held equal probability. Despite being one of the 2 heuristics which showed increased adherence in Experiment 1, this rule showed no significant change across Experiment 2 ($\beta$ = –0.002, $t(27)$ = 1.23, $p$ = 0.229, $BF_{10}$ = 1.87), with 67.9% of participants showing a negative slope, suggesting increased reliance on this rule does not explain the rise in consistency in this case.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

### *3.2.2.5. Most likely and lexicographic*

The most likely rule predicted responses for 54 of the gamble pairs, in this case being unable to account for trials where both outcomes from a given gamble have equal probability. This also means that the most likely rule is identical to the lexicographic rule for this set, as second most likely outcomes will also be undefined in such cases. Adherence to these rules showed a significant fall across the task ($\beta = -0.003$, $t(27) = 2.33$, $p = 0.028$, $BF_{10} = 135$), with 75.0% of participants showing a negative slope, in keeping with the finding from Experiment 1 that participants move away from this strategy.

### *3.2.2.6. Priority heuristic*

The greater range of probabilities and outcomes in this experiment means that the priority heuristic is now separable from the least likely rule, able to generate preferences for all 72 gamble pairs. This rule did not, however, shows any substantial change in adherence across the task either ($\beta = 0.002$, $t(27) = 1.86$, $p = 0.074$, $BF_{10} = 1.52$), suggesting use of this heuristic does not explain the observed increase in consistency, though adherence slopes were notably positive for 78.6% of participants.

### *3.2.2.7. Equiprobable and equal weight*

The equiprobable rule suggests a preference for the option with the highest mean outcome ignoring associated probabilities, while the equal weight rule suggests a preference for the option with the highest sum of its respective outcomes. As all pairs in Experiment 2 have 2 possible outcomes, these rules are functionally equivalent within this set, and so are considered together here. These rules can be applied to the gamble set of Experiment 2 as the pairs are no longer matched in outcomes, so producing a preference for all but one of the gamble pairs. No significant change in adherence to this rule was found across the task, however ($\beta = 0.002$, $t(27) = 1.22$, $p = 0.233$, $BF_{10} = 2.12$), with 60.7% of participants showing a positive slope, again suggesting no connection to the observed consistency effects.

### *3.2.2.8. Better-than-average*

The better-than-average heuristic compares the number of outcomes from each gamble that are higher than the grand average of all outcomes from both gambles, selecting the option with the higher count. While this rule is only able to predict a preference in 17 of the gamble pairs, adherence to this rule was found to significantly increase across trial blocks ($\beta = 0.002$, $t(27) = 2.99$, $p = 0.006$, $BF_{10} = 5{,}852$), with 75.0% of participants showing a positive slope, potentially indicating increased use of this strategy. As noted, however, the limited scope of this rule in this set suggests use of this heuristic is unlikely to explain all behavior in the experiment.

### *3.2.2.9. Tallying*

The tallying rule compares options across 4 criteria, selecting the option which wins more comparisons: higher minimum outcome, higher maximum outcome, lower probability of minimum outcome, and higher probability of maximum outcome. This rule predicts preferences in only 28 of the 72 pairs, and showed no significant change in adherence across the experiment, ($\beta = -0.0003$, $t(27) = 0.60$, $p = 0.552$, $BF_{10} = 0.163$), with 60.7% of participants showing a negative slope.

### *3.2.2.10. Minimax*

The minimax rule selects the option with the higher respective minimum outcome, leading to preferences in all 72 pairs. Adherence to this rule showed a significant increase across the task ($\beta = 0.003$, $t(27) = 2.59$, $p = 0.015$, $BF_{10} = 210$), with 78.6% of participants showing a positive slope, suggesting an increased reliance on options which avoid more negative outcomes. In addition, this is based upon all trials, so potentially offering a more complete explanation than the better-than-average rule above. To assess this suggestion, an additional regression model was performed for the minimax rule restricting the data to only those trials where the better-than-average rule is unable to make a prediction; this again demonstrates a significant increase in adherence across blocks ($\beta = 0.003$, $t(27)$

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

***Table 2.*** *Choice rule analysis results from Experiment 2.*

| Rule | Prediction rate | $\beta$ | $t$ | $p$ | $BF_{10}$ | Positive rate |
|---|---|---|---|---|---|---|
| Expected value | 90.3% | −0.001 | 0.96 | 0.345 | 0.258 | 28.6% |
| Expected utility | 98.6% | −0.002 | 1.15 | 0.261 | 1.67 | 32.1% |
| Probable | 98.6% | −0.003 | 1.90 | 0.069 | 14.0 | 28.6% |
| Least likely | 80.6% | −0.002 | 1.23 | 0.229 | 1.87 | 32.1% |
| Most likely and lexicographic | 75.0% | −0.003 | 2.33 | 0.028 | 135 | 25.0% |
| Priority heuristic | 100% | 0.002 | 1.86 | 0.074 | 1.52 | 78.6% |
| Equiprobable and equal weight | 98.6% | 0.002 | 1.22 | 0.233 | 2.12 | 60.7% |
| Better-than-average | 23.6% | 0.002 | 2.99 | 0.006 | 5,852 | 75.0% |
| Tallying | 38.9% | −0.0003 | 0.60 | 0.552 | 0.163 | 39.3% |
| Minimax | 100% | 0.003 | 2.59 | 0.015 | 210 | 78.6% |
| Maximax | 98.6% | −0.0002 | 0.15 | 0.880 | 0.121 | 50.0% |

*Note*: Prediction rate gives the proportion of gamble pairs for which that rule provides a preference, $\beta$, $t$, $p$, and $BF_{10}$ give regression results for adherence to that rule over blocks, and positive rate gives the proportion of participants showing a positive slope at the individual level.

= 2.61, $p$ = 0.015, $BF_{10}$ = 136), suggesting the minimax rule is able to account for the increase in consistency beyond the limited set explained by the better-than-average rule.

### *3.2.2.11. Maximax*

The maximax rule offers a counterpoint to the minimax rule, preferring options with the highest outcome. This produces a preference in all but one of the 72 gamble pairs, but showed no change in adherence in the experiment ($\beta$ = −0.0002, $t$(27) = 0.15, $p$ = 0.880, $BF_{10}$ = 0.121), and an equal 50% split between positive and negative slopes at the individual level.

### *3.2.2.12. Comparing choice between tasks*

Adherence patterns in Experiment 2 offer an interesting contrast with the those observed in Experiment 1, where participants seemingly showed increased preference for options with lower minimum outcome probabilities based on increased adherence to both the least likely rule and the priority heuristic: these rules show no reliable change in adherence in this data set, with the best explanation for the increase in consistency here being an increased aversion to the lowest outcome in a given pair. It is notable, however, that such behavior does in fact correspond with underlying principles of the priority heuristic, even if this is not displayed quantitatively for the specific implementation used here: as previously noted, the priority heuristic suggests decision makers begin by contrasting minimum outcomes, but may move to the probability of those minima where this contrast is indecisive. Such sequential consideration may then account for the apparent disconnect between the results of these 2 tasks: contrasts of minimum outcomes are unhelpful in the matched outcome gamble pairs of Experiment 1, forcing participants to proceed to comparisons of probability, but offer greater guidance in the more varied gambles of Experiment 2. The adherence patterns for the priority heuristic above meanwhile are dependent on the thresholds used to determine a meaningful difference in outcomes; indeed, reducing these thresholds to 0 leads to agreement with the minimax rule, and so equivalent performance. Such an equivalence is notable given that the thresholds used in the present implementation were taken from the original definition by Brandstätter et al. (2006), but are not inherent aspects of the rule, and so could be varied, though again we avoid such parameter fitting in the present study.

Behavior in these experiments may then be more consistent than it initially appears, expressed differently due to the different structures of the 2 gamble sets used in these tasks: where outcomes are not matched between pairs, participants may specifically avoid the lowest outcome across options, but where outcomes are matched, participants may instead rely on the probability of the shared minimum to guide decisions, so avoiding the option which is more likely to produce this minimum. Both cases could

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

then reflect an increase in forms of safety-seeking, indicating an aversion to either the worst available outcome in a given trial or a greater likelihood of such an outcome. Such a suggestion is, however, only speculative, and will require further verification in more targeted studies. Both experiments do, however, show agreement that choices do not appear to become any more optimal over the respective tasks whether defined by value or utility, suggesting repetition of choice does not lead to improved performance.

### *3.3. Summary*

The results of Experiment 2 broadly correspond with those of Experiment 1: choices were more consistent with fewer intervening trials and greater experience with the choice set, but were seemingly unaffected by match in trial order. There is, however, more confidence in this finding in this case as match in trial order is separated from distance, so providing stronger evidence that matching trial order has little impact on consistency. In addition, the increased block count of this experiment gives more detail on the role of experience, suggesting a continued increase in consistency as the experiment progressed, though this does not appear to be a strictly monotonic pattern.

## 4. General discussion

Across 2 experiments, we find that inconsistency in risky decisions does not appear to be driven solely by a common contextual factor: inconsistency remains without random variations in trial order, with strong evidence that this in fact has no impact on the consistency of choices. This then provides further clarity on the role of internal noise in decision-making on the consistency of choices: while it is likely impossible to completely control for all sources of external noise, by accounting for a key factor raised by previous research, the present data do indicate that fluctuations in choice extend beyond variations in context, being seemingly inherent in the internal decision process.

Such findings carry major implications for existing models of decision-making, providing assurance that previous estimates of internal noise are accurate and not confounded by the use of randomized trial orders. This is particularly notable given the large literature finding sequence effects across varied estimation tasks (Garner, 1954; Holland & Lockhead, 1968; Lacouture, 1997; Stewart & Brown, 2004; Ward & Lockhead, 1970), potentially indicating a difference in sequential effects between direct identification or valuation tasks and the binary choice tasks commonly used in studies on decisions under risk. These results could then offer confidence that randomized trial orders can be used in repeated risky choice tasks without worry of introducing confounding sequence effects, though this may require further verification before being accepted.

More broadly, these results demonstrate the importance of considerations of internal noise in models of decision-making: in order to provide accurate depictions of behavior, some element of cognitive variability should be included in these models, either at the level of individual decisions or perhaps in the higher-level selection of decision strategies. At the same time, however, our other observed consistency effects suggest such noise is not simply randomly distributed across decisions, but evolves over time: consistency is higher both where choices are closer together and following further experience with the task. Such results carry important implications for decision models, indicating behavior cannot be adequately captured by the basic addition of independent and identically distributed noise to decisions or valuations; instead, noise appears to gradually decrease over time, leading to more stable decisions. This in fact corresponds with suggestions from some existing decision models: as noted previously, true and error models have proposed that repetition helps the decision maker discover their preferences, allowing more consistent choices with experience (Birnbaum & Schmidt, 2015), while related models have suggested that the parameters of a decision maker's choice rule may gradually drift over time, leading consistency to fall with distance (Birnbaum & Wan, 2020). While these are distinct explanations, these could be fairly easily combined to offer a joint account for the current

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

results. This being said, such an explanation does carry certain assumptions regarding the origins of internal noise: for example, by attributing distance effects to drift in parameters, this assumes a common decision rule is used throughout the task as opposed to variation in higher-level selection of choice rules. As previously stated, we have sought to avoid such strong assumptions in this paper: the current experiments assess the role of internal noise on behavior where certain external sources of noise are addressed, but do not offer any specific suggestion of how such noise actually arises in the mind. As such, we restrict the implications of the current findings for theories of decision-making to the broad conceptual level rather than support for any specific mechanism. In this regard, the suggestion offered by the present data is that contextual factors introduced by randomized trial orders are unlikely to offer major concerns when identifying the reasons for inconsistency, allowing for greater focus on such internal sources in future research.

The present findings also offer an interesting comparison with rank-based models of decision-making which argue the importance of context on choices: the collected data oppose such suggestions, instead indicating that contextual factors have little influence on choice consistency, and so offers some challenge to these theories. This is, however, dependent on the focus on local context as defined by trial order rather than the wider context provided by all trials in the decision set: if decisions are based on the position of an option in the distribution of observed outcomes as rank-based theories would suggest (Stewart et al., 2006, 2015), then this may depend on all trials in a block rather than simply the most recent events. Any sequence effects may then be masked by the influence of this wider distribution: recent trials may be less impactful when options are already considered against all values observed in the task. This then provides an alternate explanation for the experience effects seen in these tasks: consistency may increase with repetition as participants learn the distribution of values across trials and so are better able to place an option relative to its competitor. If this is the case, however, it is unclear why experience effects should continue with further exposure, as was observed in Experiment 2: if decisions were purely dependent on the position of target values within the wider distribution, then consistency would be fully achieved after the first block once all values had been viewed. This could be attributable to memory effects: with further repetitions, the decision maker is able to more fully acquire the distribution, and so becomes more stable in their choices. Alternatively, participants may be better able to direct attention toward crucial features as trials are repeated, better identifying their chosen criteria with practice. These are, however, merely speculative answers, meaning the present data still present some challenges to rank-based approaches that will need to be answered.

Moving beyond the consistency of risky decisions, our analyses of the content of these choices also raise questions on the preferences reflected by participants in these tasks: despite the increase in consistency in both experiments, choices do not show increased adherence to either expected value or expected utility with repetition, in fact showing a decrease in optimality in some cases. This would then imply that, if choices are indeed converging on stable preferences, these are seemingly not dependent on the value of the considered gambles, or even correlated value-based models such as expected utility. Conversely, the increase in adherence to simplistic decision rules in the current analyses suggests that choices may instead be converging on heuristic strategies, with participants in both tasks potentially making choices according to only minimum values from each gamble rather than all possible outcomes. This further constrains models of decision-making, suggesting that while experience does impact choice, this does not necessarily lead to more rational decisions. Further research is thus required to determine whether such a shift toward simple decision rules is reliable, and if so, why this might occur; for example, decision makers may move toward such strategies for reasons of cognitive economy, or these rules may be generally preferred and simply better executed with practice.

It should of course be noted that the present choice analyses are far from comprehensive, examining adherence to only a set of simplistic choice rules; there are a number of alternate possible strategies which could be used in these tasks, including much more complex decision models, making identification of the true strategy difficult. This is further complicated by the uncertainty in the present results as to the level at which variation in behavior originates: changes in adherence could arise from a number of sources, such as increased selection of a given rule, a reduction in noise in a particular

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

strategy, or higher weighting of a feature in a more complex system. Furthermore, there is no assurance that all participants rely on a common strategy, leaving open the possibility of individual differences in rule use. This provides an extraordinarily large set of potential explanations to be investigated, and thus an infeasible model fitting challenge. While such an examination may be informative, the current tasks were not intended for such extensive model comparison, instead offering an empirical exploration of consistency both in reaction to external factors and across time. In addition, as noted above, the fitting of more complex models with multiple free parameters to these tasks carries its own issues regarding the in-built assumptions of such models, most notably on the origins of variation in behavior between choice blocks: models using consistent parameters across blocks inherently assume that a given choice rule is fixed and any variation in fit arises in the selection or execution of that rule, whereas models using distinct parameters instead assume that variations are directly attributable to fluctuations in model parameters. The present focus on empirical observations of consistency and base adherence to simple fixed decision rules was deliberately chosen to avoid such hard assumptions, with these findings reflecting general behavioral patterns without being constrained by any specific theory. In addition, as the above text should indicate, this does not limit the implications of these results for existing models of decision-making, as these patterns offer key challenges to be addressed when composing such theories. This being said, such an approach does restrict what can be concluded on the methods used by participants to make their choices. We therefore leave such modeling as an open direction for future work on this subject using more targeted experimental designs. Fortunately, there are methods available in the existing literature to assist with such contrasts: for example, eye- or mouse-tracking can be used to infer attention patterns indicative of particular decision strategies (Pachur et al., 2018; Payne et al., 1993), while model comparisons can specifically contrast consistent but noisy choice models with 'toolbox' models collecting an assortment of potential strategies that could be used at any time (Olschewski & Rieskamp, 2021; Scheibehenne et al., 2013), offering a number of viable paths to build on this work.

A similar point may be made on the form of consistency targeted by the present study: we here focus on the most basic form of consistency, that is whether identical choices are made when a decision is exactly repeated. This is in contrast to methods used in previous studies which have examined consistency across different but related decisions, assessing aspects such as violations of transitivity or risk aversion (Birnbaum, 2004, 2011; Regenwetter et al., 2011), or stability in inferred model parameters (Glöckner & Pachur, 2012). The current definition is therefore a more fundamental reading of consistency, providing a universal measure more easily applied to alternate tasks and scenarios. Even so, future work may wish to examine whether similar contextual patterns appear in more complex forms of consistency.

There is, however, one important caveat to this study which should be noted before closing: the current definition of context focuses on the values presented by nearby trials (as dictated by trial order) given their apparent influence on decisions in related literature, but gives little consideration of other factors which could be considered contextual. Such a focus is warranted given that trial order is an inherent aspect of risky choice paradigms: any study examining multiple choices will have to consider the order of these decisions, with randomized order being a common default, meaning evaluation of this factor is crucial. This being said, this naturally limits the scope of any conclusions regarding the importance of context on decisions to this particular definition. This is most notable in the absence of trial-by-trial feedback, which as previously noted in the Introduction to this paper is another key source of variability in risky choice: any feedback provided for such decisions is naturally variable where outcomes are inherently stochastic (Levin & Hart, 2003; Ludvig & Spetch, 2011; Ludvig et al., 2014; Mosteller & Nogee, 1951; Studer et al., 2016). While the present data do indicate that variations in feedback are not the sole driver of inconsistency in choice, without direct manipulation and measurement of this factor, we cannot conclude that this has *no* influence on decisions in the same manner as trial order. Such a distinction is important as feedback may impact not only the consistency of choice but also its content, as experience with actual outcomes might be a key determinant of changes in strategy between instances: receiving unexpected or aversive outcomes may push people

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

toward alternate decision functions by showing their current strategy is flawed. This is particularly apparent in so-called ‘win-stay/lose-shift’ strategies in which decision makers alter selections where a previously favored prospect provides a poorer outcome (Otto et al., 2011; Worthy et al., 2013), though such experience might also offer signals toward the expected value or utility of a prospect, and so allow for increased adherence to these rules that was not observed here. Further contrasts are thus required to more directly estimate the influence of feedback on consistency using more direct manipulations of this factor, though as previously noted, such control is challenging where choices are made freely. Similarly, future work may wish to explore alternative contextual factors such as salient environmental values or past experience using more specialized designs, with the current study answering one of the most fundamental concerns.

Finally, another concern that may be raised with this study is the role of fatigue: while the multiple repetitions used in these experiments are necessary to examine the evolution of consistency, this could also lead to exhaustion with the task. The observed increases in consistency could then be attributable to greater reliance on more simplistic decision rules which expend less cognitive effort, following the previously noted findings on heuristic use above, rather than greater experience with the task. Such a suggestion would, however, require comparisons of the present findings on rule adherence with tasks in which participants perform a similar amount of decisions but without repetition of the same trial set; this falls outside of the focus on consistency of the present study, but could be an avenue for future work. Alternatively, fatigue could be alleviated with longer breaks between repetitions, though this could also lead to confounds if task experience is corrupted through memory decay or intervening events.

### *4.1. Conclusion*

The present work offers assurance that inconsistency observed at the behavioral level is unlikely to arise purely from external factors, so helping to confirm assumptions of internal noise in decision-making. Furthermore, such noise is apparently alleviated as decisions are repeated, with choices becoming more consistent with experience. This does not, however, appear to be associated with any improvement in performance, instead seemingly reflecting greater reliance on more simplistic decision rules to guide selection. Such findings carry important implications for existing theories of decision-making regarding not just noise, but the factors which determine choices themselves, requiring further attention in such research. We therefore hope that this study indicates the various elements that influence consistency in risky choices, and offers some direction toward the reasons for any change in decisions to be further investigated in future work.

**Data availability statement.** All empirical data collected for this study as well as the files used in the described analysis are available on the Open Science Framework at https://osf.io/p34gj/.

**Acknowledgments.** We thank our MSc students Anna Sali, Phutthaphon Suvanrak, and Xueying Hu for coding and running of experiments.

**Funding statement.** J.S. and A.N.S. were supported by the European Research Council consolidator grant (817492-SAMPLING). T.L.M. was supported by the Economic and Social Research Council (Grant No. ES/N018192/1 and ES/K002201/1). A.N.S. was supported by the Economic and Social Research Council (Grant No. ES/K004948/1).

**Competing interests.** We declare no competing interests in this work.

## References

Ballinger, T. P., & Wilcox, N. T. (1997). Decisions, error, and heterogeneity. *Economic Journal*, *107*(443), 1090–1105. https://doi.org/10.1111/j.1468-0297.1997.tb00009.x

Birnbaum, M. H. (2004). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: Effects of format, event framing, and branch splitting. *Organizational Behavior and Human Decision Processes*, *95*(1), 40–65. https://doi.org/10.1016/j.obhdp.2004.05.004

Birnbaum, M. H. (2006). Evidence against prospect theories in gambles with positive, negative, and mixed consequences. *Journal of Economic Psychology*, *27*(6), 737–761. https://doi.org/10.1016/j.joep.2006.04.001

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463–501. https://doi.org/10.1037/0033-295X.115.2.463

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*(4), 675–683. https://doi.org/10.1037/a0023852

Birnbaum, M. H. (2012). A statistical test of independence in choice data with small samples. *Judgment & Decision Making*, *7*(1), 97–109.

Birnbaum, M. H., & Bahra, J. P. (2012). Separating response variability from structural inconsistency to test models of risky decision making. *Judgment & Decision Making*, *7*(4), 402–426.

Birnbaum, M. H., & Schmidt, U. (2015). The impact of learning by thought on violations of independence and coalescing. *Decision Analysis*, *12*(3), 144–152. https://doi.org/10.1287/deca.2015.0316

Birnbaum, M. H., Schmidt, U., & Schneider, M. D. (2017). Testing independence conditions in the presence of errors and splitting effects. *Journal of Risk and Uncertainty*, *54*(1), 61–85. https://doi.org/10.1007/s11166-017-9251-5

Birnbaum, M. H., & Wan, L. (2020). MARTER: Markov true and error model of drifting parameters. *Judgment & Decision Making*, *15*(1), 47–73.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*(2), 409. https://doi.org/10.1037/2F0033-295X.113.2.409

Busemeyer, J. R., Weg, E., Barkan, R., Li, X., & Ma, Z. (2000). Dynamic and consequential consistency of choices between paths of decision trees. *Journal of Experimental Psychology: General*, *129*(4), 530–545. https://doi.org/10.1037/0096-3445.129.4.530

Camerer, C. F. (1989). An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty*, *2*(1), 61–104. https://doi.org/10.1007/BF00055711

Fründ, I., Wichmann, F., & Macke, J. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*, 1–16. https://doi.org/10.1167/14.7.9

Garner, W. R. (1954). Context effects and the validity of loudness scales. *Journal of Experimental Psychology*, *48*(3), 218. https://doi.org/10.1037/h0061514

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383. https://doi.org/10.1214/17-BA1051

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*, 21–32. https://doi.org/j.cognition.2011.12.002

Guo, Y., & Regenwetter, M. (2014). Quantitative tests of the perceived relative argument model: Comment on Loomes (2010). *Psychological Review*, *121*(4), 696—705. https://doi.org/10.1037/a0036095

Hey, J. D. (2001). Does repetition improve consistency? *Experimental Economics*, *4*(1), 5–54. https://doi.org/10.1007/BF01669272

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, 1291–1326. https://doi.org/10.2307/2951750

Holland, M. K., & Lockhead, G. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, *3*(6), 409–414. https://doi.org/10.3758/BF03205747

Hu, G. (1997). Why is it difficult to learn absolute judgment tasks? *Perceptual and Motor Skills*, *84*(1), 323–335. https://doi.org/10.2466/pms.1997.84.1.323

Jekel, M., Fiedler, S., & Glöckner, A. (2011). Diagnostic task selection for strategy classification in judgment and decision making: Theory, validation, and implementation in R. *Judgment & Decision Making*, *6*(8), 782–799.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research*, *60*(3), 121–133. https://doi.org/10.1007/BF00419760

Leopard, A. (1978). Risk preference in consecutive gambling. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 521–528. https://doi.org/10.1037/0096-1523.4.3.521

Levin, I. P., & Hart, S. S. (2003). Risk preferences in young children: Early evidence of individual differences in reaction to potential gains and losses. *Journal of Behavioral Decision Making*, *16*(5), 397–413. https://doi.org/10.1002/bdm.453

Loomes, G., & Pogrebna, G. (2014). Testing for independence while allowing for probabilistic choice. *Journal of Risk and Uncertainty*, *49*(3), 189–211. https://doi.org/10.1007/s11166-014-9205-0

Loomes, G., & Sugden, R. (1998). Testing different stochastic specifications of risky choice. *Economica*, *65*(260), 581–598. https://doi.org/10.1111/1468-0335.00147

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, *27*(2), 146–156. https://doi.org/10 .1002/bdm.1792

Ludvig, E. A., & Spetch, M. L. (2011). Of black swans and tossed coins: Is the description-experience gap in risky choice limited to rare events? *PloS One*, *6*(6), e20262. https://doi.org/10.1371/journal.pone.0020262

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*, 371–404. https://doi.org/10.1086/257106

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

Olschewski, S., & Rieskamp, J. (2021). Distinguishing three effects of time pressure on risk taking: Choice consistency, risk preference, and strategy selection. *Journal of Behavioral Decision Making*, *34*(4), 541–554. https://doi.org/10.1002/bdm.2228

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*(2), 274–279. https://doi.org/10.1016/j.cognition.2010.11.009

Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of Experimental Psychology: General*, *147*(2), 147. https://psycnet.apa.org/doi/10.1037/xge0000406

Payne, J. W., Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56. https://doi.org/10.1037/a0021150

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1446–1465. https://doi.org/10.1037/a0013646

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903. https://doi.org/10.1080/00273171.2012.734737

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*(1), 39. https://doi.org/10.1037/a0030777

Spicer, J., Mullett, T. L., & Sanborn, A. N. (2020). *Consistency in risky choice*. (Open Science Framework [Data Set]. https://osf.io/p34gj/)

Starmer, C., & Sugden, R. (1989). Probability and juxtaposition effects: An experimental investigation of the common ratio effect. *Journal of Risk and Uncertainty*, *2*(2), 159–178. https://doi.org/10.1007/BF00056135

Stewart, N., & Brown, G. D. (2004). Sequence effects in the categorization of tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 416–430. https://doi.org/10.1037/0278-7393.30.2.416

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Cognition*, *28*(1), 3–11. https://doi.org/10.1037/0278-7393.28.1.3

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881–911. https://doi.org/10.1037/0033-295X.112.4.881

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Stewart, N., Reimers, S., & Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, *61*(3), 687–705. https://doi.org/10.1287/mnsc.2013.1853

Studer, B., Scheibehenne, B., & Clark, L. (2016). Psychophysiological arousal and inter- and intraindividual differences in risk-sensitive decision making. *Psychophysiology*, *53*(6), 940–950. https://doi.org/10.1111/psyp.12627

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science*, *22*, 253–260. https://doi.org/10.1177/0956797610396225

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, *15*(11), 546–554. https://doi.org/10.1016/j.tics.2011.09.008

Wakker, P., Erev, I., & Weber, E. U. (1994). Comonotonic independence: The critical test between classical and rank-dependent utility theories. *Journal of Risk and Uncertainty*, *9*(3), 195–230. https://doi.org/10.1007/BF01064200

Ward, L. M., & Lockhead, G. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, *84*(1), 27–34. https://doi.org/10.1037/h0028949

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*(2), 364–371. https://doi.org/10.3758/s13423-012-0324-9

## Appendices

### *Appendix A. Gamble sets*

**Experiment 1.**

Table A1 lists the gamble pairs used in Experiment 1, adapted from Hey (2001). As these gambles all drew from a common pool of 4 outcomes, each row reports only probability values for these outcomes, though note that only 3 outcomes were presented on each trial.

**Experiment 2.**

Table A2 lists the gamble pairs used in Experiment 2, adapted from Glöckner and Pachur (2012). Unlike the set used in Experiment 1, these gambles did not share outcomes, meaning the table lists both probabilities and outcomes, where o1a reflects outcome 1 from gamble A, and p1a is the probability of that outcome. Note that gamble pair 74 was excluded from analysis due to errors in listed probability.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

***Table A1.*** *Gamble pairs from Experiment 1.*

| | Gamble A | | | | Gamble B | | | |
|---|---|---|---|---|---|---|---|---|
| Pair ID | −£1 | £1 | £3 | £5 | −£1 | £1 | £3 | £5 |
| 1 | – | 0 | 0.875 | 0.125 | – | 0.125 | 0 | 0.875 |
| 2 | – | 0.125 | 0 | 0.875 | – | 0 | 0.875 | 0.125 |
| 3 | – | 0.125 | 0.5 | 0.375 | – | 0 | 0.875 | 0.125 |
| 4 | – | 0.375 | 0 | 0.625 | – | 0 | 0.875 | 0.125 |
| 5 | – | 0.375 | 0.125 | 0.5 | – | 0 | 0.875 | 0.125 |
| 6 | – | 0 | 0.875 | 0.125 | – | 0.375 | 0.25 | 0.375 |
| 7 | – | 0.625 | 0 | 0.375 | – | 0 | 0.875 | 0.125 |
| 8 | – | 0.125 | 0.5 | 0.375 | – | 0.375 | 0 | 0.625 |
| 9 | – | 0.375 | 0.125 | 0.5 | – | 0.125 | 0.5 | 0.375 |
| 10 | – | 0.375 | 0 | 0.625 | – | 0.125 | 0.875 | 0 |
| 11 | – | 0.125 | 0.875 | 0 | – | 0.375 | 0.125 | 0.5 |
| 12 | – | 0.375 | 0.25 | 0.375 | – | 0.125 | 0.875 | 0 |
| 13 | – | 0.125 | 0.875 | 0 | – | 0.375 | 0.5 | 0.125 |
| 14 | – | 0.625 | 0 | 0.375 | – | 0.125 | 0.875 | 0 |
| 15 | – | 0.875 | 0 | 0.125 | – | 0.125 | 0.875 | 0 |
| 16 | – | 0.25 | 0.75 | 0 | – | 0.375 | 0 | 0.625 |
| 17 | – | 0.375 | 0.125 | 0.5 | – | 0.25 | 0.75 | 0 |
| 18 | – | 0.375 | 0.25 | 0.375 | – | 0.25 | 0.75 | 0 |
| 19 | – | 0.375 | 0.5 | 0.125 | – | 0.25 | 0.75 | 0 |
| 20 | – | 0.25 | 0.75 | 0 | – | 0.375 | 0.5 | 0.125 |
| 21 | – | 0.625 | 0 | 0.375 | – | 0.25 | 0.75 | 0 |
| 22 | – | 0.875 | 0 | 0.125 | – | 0.25 | 0.75 | 0 |
| 23 | – | 0.625 | 0 | 0.375 | – | 0.375 | 0.5 | 0.125 |
| 24* | – | 0.25 | 0.75 | 0 | – | 0.125 | 0.875 | 0 |
| 25* | – | 0.375 | 0.25 | 0.375 | – | 0.375 | 0.125 | 0.5 |
| 26 | 0 | 0.75 | 0.25 | – | 0.125 | 0 | 0.875 | – |
| 27 | 0 | 0.75 | 0.25 | – | 0.125 | 0.375 | 0.5 | – |
| 28 | 0 | 0.75 | 0.25 | – | 0.375 | 0.125 | 0.5 | – |
| 29 | 0 | 0.75 | 0.25 | – | 0.375 | 0.25 | 0.375 | – |
| 30 | 0.5 | 0 | 0.5 | – | 0 | 0.75 | 0.25 | – |
| 31 | 0 | 0.75 | 0.25 | – | 0.5 | 0.125 | 0.375 | – |
| 32 | 0.125 | 0 | 0.875 | – | 0 | 1 | 0 | – |
| 33 | 0 | 1 | 0 | – | 0.125 | 0.375 | 0.5 | – |
| 34 | 0.25 | 0.625 | 0.125 | – | 0 | 1 | 0 | – |
| 35 | 0.375 | 0.125 | 0.5 | – | 0 | 1 | 0 | – |
| 36 | 0.375 | 0.25 | 0.375 | – | 0 | 1 | 0 | – |
| 37 | 0 | 1 | 0 | – | 0.5 | 0 | 0.5 | – |
| 38 | 0 | 1 | 0 | – | 0.5 | 0 | 0.5 | – |
| 39 | 0 | 1 | 0 | – | 0.5 | 0.125 | 0.375 | – |
| 40 | 0 | 1 | 0 | – | 0.75 | 0.125 | 0.125 | – |
| 41 | 0.25 | 0.625 | 0.125 | – | 0.375 | 0.125 | 0.5 | – |
| 42 | 0.25 | 0.625 | 0.125 | – | 0.375 | 0.25 | 0.375 | – |
| 43 | 0.25 | 0.625 | 0.125 | – | 0.5 | 0 | 0.5 | – |
| 44 | 0.25 | 0.625 | 0.125 | – | 0.5 | 0.125 | 0.375 | – |
| 45 | 0.5 | 0 | 0.5 | – | 0.375 | 0.25 | 0.375 | – |

*Table A1.* *Continued.*

| | Gamble A | | | | Gamble B | | | |
|---|---|---|---|---|---|---|---|---|
| Pair ID | −£1 | £1 | £3 | £5 | −£1 | £1 | £3 | £5 |
| 46 | 0.375 | 0.25 | 0.375 | – | 0.5 | 0 | 0.5 | – |
| 47 | 0.375 | 0.625 | 0 | – | 0.5 | 0 | 0.5 | – |
| 48 | 0.375 | 0.625 | 0 | – | 0.5 | 0.125 | 0.375 | – |
| 49 | 0.75 | 0.125 | 0.125 | – | 0.375 | 0.625 | 0 | – |
| 50* | 0.375 | 0.125 | 0.5 | – | 0.5 | 0.125 | 0.375 | – |

*Note*: Asterisks denote dominated pairs.

*Table A2.* *Gamble pairs from Experiment 2.*

| | Gamble A | | | | Gamble B | | | |
|---|---|---|---|---|---|---|---|---|
| Pair ID | p1a | o1a | p2a | o2a | p1b | o1b | p2b | o2b |
| 1 | 0.1 | £400 | 0.9 | £320 | 0.1 | £770 | 0.9 | £20 |
| 2 | 0.3 | £400 | 0.7 | £320 | 0.3 | £770 | 0.7 | £20 |
| 3 | 0.4 | £400 | 0.6 | £320 | 0.4 | £770 | 0.6 | £20 |
| 4 | 0.5 | £400 | 0.5 | £320 | 0.5 | £770 | 0.5 | £20 |
| 5 | 0.9 | £400 | 0.1 | £320 | 0.9 | £770 | 0.1 | £20 |
| 6* | 1 | £400 | 0 | £320 | 1 | £770 | 0 | £20 |
| 7 | 0.5 | −£100 | 0.5 | £220 | 0.5 | £0 | 0.5 | £0 |
| 8 | 0.5 | −£100 | 0.5 | £50 | 0.5 | £0 | 0.5 | £0 |
| 9 | 0.5 | −£100 | 0.5 | £200 | 0.5 | £0 | 0.5 | £0 |
| 10 | 0.5 | −£100 | 0.5 | £400 | 0.5 | £0 | 0.5 | £0 |
| 11 | 0.5 | −£100 | 0.5 | £300 | 0.5 | £0 | 0.5 | £0 |
| 12 | 0.5 | −£100 | 0.5 | £100 | 0.5 | £0 | 0.5 | £0 |
| 13 | 0.5 | −£100 | 0.5 | £150 | 0.5 | £0 | 0.5 | £0 |
| 14 | 0.5 | −£100 | 0.5 | £240 | 0.5 | £0 | 0.5 | £0 |
| 15 | 0.25 | £0 | 0.75 | £705 | 0.5 | £60 | 0.5 | £1,000 |
| 16 | 0.6 | £225 | 0.4 | £330 | 0.33 | £0 | 0.67 | £400 |
| 17 | 0.4 | £500 | 0.6 | £333 | 0.6 | £0 | 0.4 | £1,000 |
| 18 | 0.45 | £100 | 0.55 | £400 | 0.4 | £0 | 0.6 | £440 |
| 19 | 0.7 | £300 | 0.3 | £400 | 0.6 | £200 | 0.4 | £525 |
| 20 | 0.6 | 3,250 | 0.4 | £300 | 0.5 | £200 | 0.5 | £340 |
| 21 | 0.3 | £0 | 0.7 | £450 | 0.6 | £40 | 0.4 | £730 |
| 22 | 0.55 | £840 | 0.45 | £820 | 0.4 | £800 | 0.6 | £850 |
| 23 | 0.4 | £510 | 0.6 | £800 | 0.3 | £0 | 0.7 | £980 |
| 24 | 0.3 | £0 | 0.7 | £450 | 0.5 | £30 | 0.5 | £600 |
| 25 | 0.65 | £200 | 0.35 | £700 | 0.5 | £0 | 0.5 | £750 |
| 26 | 0.5 | £300 | 0.5 | £400 | 0.3 | £0 | 0.7 | £500 |
| 27 | 0.5 | £0 | 0.5 | £400 | 0.65 | £39 | 0.35 | £500 |
| 28 | 0.4 | £0 | 0.6 | £500 | 0.55 | £40 | 0.45 | £620 |
| 29 | 0.5 | £100 | 0.5 | £345 | 0.6 | £50 | 0.4 | £480 |
| 30 | 0.5 | £535 | 0.5 | £650 | 0.65 | £400 | 0.35 | £950 |
| 31 | 0.85 | £100 | 0.15 | £900 | 0.02 | £10 | 0.98 | £225 |
| 32 | 0.9 | £590 | 0.1 | £1,000 | 0.02 | £0 | 0.98 | £650 |

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

***Table A2.*** *Continued.*

| | Gamble A | | | | Gamble B | | | |
|---|---|---|---|---|---|---|---|---|
| Pair ID | p1a | o1a | p2a | o2a | p1b | o1b | p2b | o2b |
| 33 | 0.99 | £160 | 0.01 | £170 | 0.85 | £30 | 0.15 | £880 |
| 34 | 0.9 | £160 | 0.1 | £1,000 | 0.02 | £0 | 0.98 | £250 |
| 35 | 0.99 | £470 | 0.01 | £800 | 0.8 | £350 | 0.2 | £1,000 |
| 36 | 0.61 | £680 | 0.39 | £390 | 0.6 | £770 | 0.4 | £480 |
| 37 | 0.32 | −£830 | 0.68 | −£190 | 0.15 | −£150 | 0.85 | −£520 |
| 38 | 0.08 | £820 | 0.92 | £720 | 0.27 | £470 | 0.73 | £770 |
| 39 | 0.73 | −£380 | 0.27 | £400 | 0.54 | −£300 | 0.46 | £70 |
| 40 | 0.21 | −£600 | 0.79 | −£160 | 0.45 | −£60 | 0.55 | −£510 |
| 41 | 0.43 | −£550 | 0.57 | −£530 | 0.08 | −£770 | 0.92 | −£510 |
| 42 | 0.13 | −£290 | 0.87 | −£760 | 0.55 | −£1,000 | 0.45 | −£280 |
| 43 | 0.68 | £300 | 0.32 | £670 | 0.59 | £70 | 0.41 | £590 |
| 44 | 0.76 | £370 | 0.24 | £50 | 0.18 | £120 | 0.82 | £280 |
| 45 | 0.31 | −£440 | 0.69 | £260 | 0.38 | £840 | 0.62 | −£410 |
| 46 | 0.27 | £250 | 0.73 | −£610 | 0.34 | −£820 | 0.66 | £30 |
| 47 | 0.51 | −£260 | 0.49 | −£760 | 0.77 | −£480 | 0.23 | −£340 |
| 48 | 0.74 | £620 | 0.26 | £0 | 0.44 | £230 | 0.56 | £310 |
| 49 | 0.56 | −£60 | 0.44 | −£590 | 0.64 | −£270 | 0.36 | −£830 |
| 50 | 0.73 | £750 | 0.27 | £340 | 0.9 | £560 | 0.1 | £820 |
| 51 | 0.28 | £70 | 0.72 | £740 | 0.71 | £550 | 0.29 | £630 |
| 52 | 0.31 | £1,000 | 0.69 | £480 | 0.58 | £1,000 | 0.42 | £10 |
| 53 | 0.07 | −£550 | 0.93 | £950 | 0.48 | −£130 | 0.52 | £990 |
| 54 | 0.45 | −£370 | 0.55 | −£340 | 0.41 | −£110 | 0.59 | −£540 |
| 55 | 0.19 | £430 | 0.81 | £50 | 0.14 | £910 | 0.86 | £60 |
| 56 | 0.04 | −£980 | 0.96 | −£250 | 0.4 | −£780 | 0.6 | −£220 |
| 57 | 0.29 | −£670 | 0.71 | −£280 | 0.05 | −£460 | 0.95 | −£440 |
| 58 | 0.39 | £760 | 0.61 | −£70 | 0.06 | −£650 | 0.94 | £370 |
| 59 | 0.76 | −£30 | 0.24 | −£920 | 0.62 | −£140 | 0.38 | −£830 |
| 60 | 0.25 | −£940 | 0.75 | −£370 | 0.83 | −£490 | 0.17 | −£110 |
| 61 | 0.28 | −£590 | 0.72 | £960 | 0.04 | −£40 | 0.96 | £630 |
| 62 | 0.43 | −£470 | 0.57 | £630 | 0.02 | −£690 | 0.98 | £140 |
| 63 | 0.24 | −£880 | 0.76 | £790 | 0.01 | −£1,000 | 0.99 | £320 |
| 64 | 0.04 | £390 | 0.96 | £280 | 0.39 | £530 | 0.61 | £200 |
| 65 | 0.58 | −£440 | 0.42 | £430 | 0.83 | £10 | 0.17 | −£700 |
| 66 | 0.61 | £960 | 0.39 | −£670 | 0.5 | £710 | 0.5 | −£260 |
| 67 | 0.84 | −£570 | 0.16 | −£900 | 0.25 | −£630 | 0.75 | −£300 |
| 68 | 0.34 | £240 | 0.66 | £590 | 0.42 | £470 | 0.58 | £640 |
| 69 | 0.93 | £750 | 0.07 | −£900 | 0.87 | £960 | 0.13 | −£890 |
| 70 | 0.63 | £410 | 0.37 | £180 | 0.98 | £560 | 0.02 | £80 |
| 71 | 0.55 | £890 | 0.45 | −£520 | 0.76 | £860 | 0.24 | −£750 |
| 72 | 0.59 | −£410 | 0.41 | −£570 | 0.42 | −£510 | 0.58 | −£300 |
| 73 | 0.24 | −£250 | 0.76 | £360 | 0.48 | −£280 | 0.52 | £670 |
| 74 | 0.3 | £50 | 0.7 | £100 | 0.8 | £100 | 0.2 | £150 |
| 75* | 0.45 | −£200 | 0.55 | £200 | 0.35 | £600 | 0.65 | £200 |

*Note*: Asterisks denote dominated pairs. Note that gamble pair 74 was excluded from all analyses due to an error in listed probabilities.

## *Appendix B. Bayesian regression details*

Both experiments used Bayesian logistic regression models when analyzing choice consistency to supplement the listed frequentist tests with credible intervals on coefficient estimates and Bayes factors for each predictor weighing strength of evidence for the null and alternative hypotheses (calculated using the Savage–Dickey method to compare prior and posterior distributions of coefficient estimates). These Bayesian models were defined in the same manner as the standard regression models, with the addition of distributions representing prior expectations of the coefficients of each predictor. Priors were defined following the procedure suggested by Gelman, Jakulin, Pittau, and Su (2008) and Ghosh, Li, and Mitra (2018), using a Student's *t*-distribution with mean 0, standard deviation 2.5, and degrees of freedom 4. To suit these prior distributions, each predictor was rescaled to have mean 0 and standard deviation 0.5, though note that all values listed in the main text are prior to this rescaling. One effect of the use of such priors is a possible disagreement between frequentist and Bayesian results, as this prior definition is more conservative than those used by the frequentist tests: this is most notable for particularly small effect sizes, where a coefficient is considered reliably different from 0 by frequentist standards, but still within range of the prior in the Bayesian analysis. The appearance of such disagreements in the above analysis is therefore likely reflective of this prior definition rather than any disagreement in results, though as a precaution we avoid making any strong conclusions where such discrepancies are found.

Bayesian regression models were also used in the analysis of choice content to provide credible intervals and Bayes factors on change in adherence to each of the considered heuristic rules, though as these models were linear rather than logistic, a different method was used to calculate these measures, comparing versions of the model including the factor of trial block against a null model with only participant as a randomized factor. This also used alternate priors to those noted above for the consistency analysis, using the default prior of the R package BayesFactor, as defined by Rouder and Morey (2012).

## *Appendix C. Dominated choice rates*

While dominated gamble pairs were excluded from the main analyses in both experiments, we here examine choices in these trials to look for any change in attention over the course of each task. Average dominant choice rates from both experiments are illustrated in Figure C1, and were examined using separate logistic regression models for each experiment. This observed a negative but non-significant

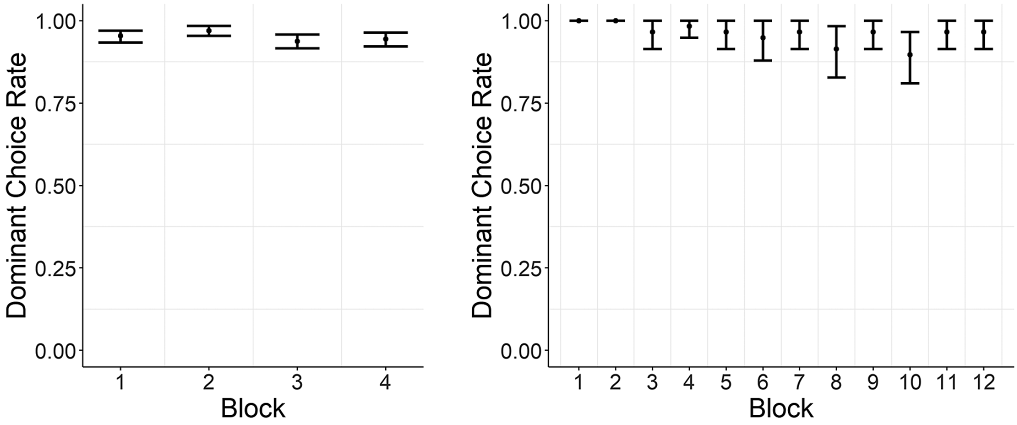


***Figure C1.*** *Mean choice rate for dominated gamble pairs across Experiment 1 (left) and Experiment 2 (right). Error bars show bootstrapped 95% CIs.*

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

effect of trial block on the rate of dominant choices in Experiment 1 ($\beta = -0.16$, $z = 1.58$, $p = 0.115$), and a significant fall in Experiment 2 ($\beta = -0.15$, $z = 2.37$, $p = 0.018$). While such results could suggest that attention fell across the tasks due to fatigue, thereby leading to more random responding, it is notable that this opposes the increase in consistency observed in the main trial set for both experiments. As such, choices in these trials may not be becoming more random, but instead converging to a decision strategy which leads to increased favoring of dominated options. It is unclear, however, what this rule may be, as the key decision strategies targeted here should all favor dominant options.

## *Appendix D. Additional choice analyses*

As noted in the main text, the results of our consistency analyses suggest an increase in consistency across the tasks, with higher rates for comparisons involving later blocks and positive slopes in the cut-down regressions. This may not, however, be an entirely linear effect: such increases might be more substantial between earlier blocks as participants learn the gamble sets, after which choices, and therefore consistency rates, stabilize. This particularly applies to Experiment 1 as consistency rates show larger differences when including block 1 versus the other blocks, but are otherwise reasonably similar: in fact, when considering only adjacent comparisons, if the initial comparison between blocks 1 and 2 is removed, the order effects reported previously become non-significant ($\beta = 0.028$, $z = 0.619$, $p = 0.536$, $BF_{10} = 0.023$), though the same does not apply to Experiment 2 ($\beta = 0.026$, $z = 4.31$, $p = < 0.001$, $BF_{10} = 13.0$). Such effects might then suggest that the above choice analyses using linear predictors across all blocks are inappropriate, as the only crucial difference in adherence would be between blocks 1 and 2.

We thus performed additional regression models for each rule considered in Experiment 1 restricting focus to only the first two blocks. Due to this restriction on the data, these models did not include random slopes for each participant, meaning individual level results are unavailable. Results from these models are summarized in Table D1: as in the main analysis, this finds a significant increase in adherence for the least likely/priority heuristic rules, but also shows a positive effect for the subjective expected utility functions, which previously only held ambiguous evidence. All other rules meanwhile show negative or flat slopes, meaning these continue to mismatch the observed consistency results. Expected utility may then match this suggested pattern, showing an initial increase as participants learn the choice set, but subsequently leveling off. This does, however, ignore a substantial portion of the collected data, including the continued increase in adherence for the least likely/priority heuristic rules in the later blocks demonstrated in Figure 4, meaning such a restriction may overlook other aspects of behavior. Moreover, this pattern is not reflected in the data of Experiment 2, where evidence for a continued increase in consistency across the task is more assured, and equivalent restricted analyses find no meaningful change in adherence to expected utility between the initial blocks ($\beta = 0.004$, $t(27) = 0.32$, $p = 0.753$, $BF_{10} = 0.280$). We therefore focus on the results featured in the main text as these

***Table D1.*** *Alternative choice rule analysis results from Experiment 1 considering adherence in only blocks 1 and 2.*

| Rule | $\beta$ | $t$ | $p$ | $BF_{10}$ |
|---|---|---|---|---|
| Expected value | −0.021 | 4.29 | < 0.001 | 524 |
| Expected utility | 0.016 | 3.09 | 0.002 | 10.3 |
| Probable | −0.013 | 3.66 | < 0.001 | 58.3 |
| Least likely and priority heuristic | 0.037 | 5.26 | < 0.001 | 24686 |
| Most likely | −0.004 | 0.88 | 0.379 | 0.179 |
| Lexicographic | 0.006 | 1.48 | 0.142 | 0.348 |

*Note*: $\beta$, $t$, $p$, and $BF_{10}$ give regression results for adherence to that rule.

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

consider the full range of behavior in the tasks, though such non-linearity may deserve more attention in future work.

## *Appendix E. Reaction time analyses*

In addition to the analyses of consistency featured in the main text, we also performed supplementary analyses on reaction times from both experiments to examine whether these also showed an evolution across each task.

### Experiment 1.

Reaction times from Experiment 1 were examined using a linear regression with trial block as a predictor (see Figure E1). This found reaction times fell across the experiment ($\beta = -1.96$, $z = 59.7$, $p < 0.001$, $BF_{10} > 10{,}000$), potentially suggesting that participants were indeed increasingly relying on simple decision rules as the task progressed, so allowing for faster decisions. This is, however, a fairly speculative suggestion, as the current task was not intended to model reaction times; as such, this data are unable to distinguish between a greater reliance on simpler rules and more efficient use of the same rule across the task Figure E1.

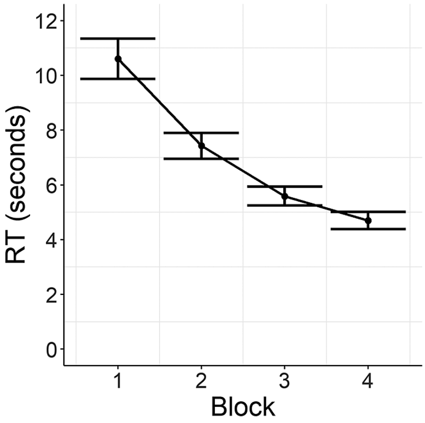


***Figure E1.*** *Mean reaction times across Experiment 1. Error bars show 95% CIs.*

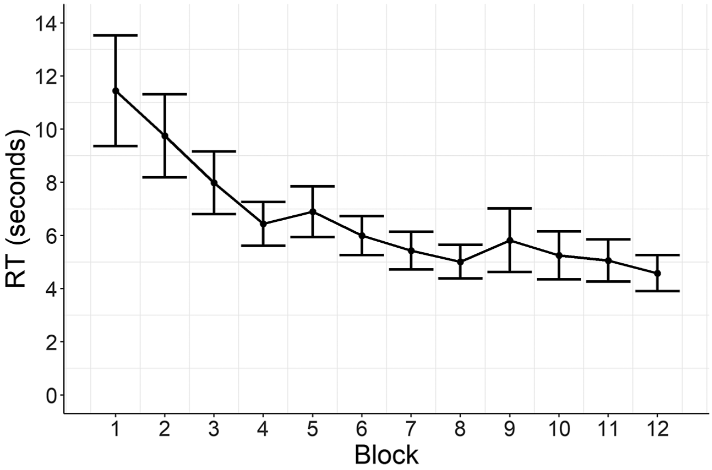


***Figure E2.*** *Mean reaction times across Experiment 2. Error bars show 95% CIs.*

https://doi.org/10.1017/jdm.2023.41 Published online by Cambridge University Press

**Experiment 2.**

As with Experiment 1, reaction times in Experiment 2 were examined using a linear regression across the trial blocks (see Figure E2). This again found a decrease in reaction times across the task ($\beta = -0.51$, $z = 42.0$, $p < 0.001$, $BF_{10} > 10{,}000$), providing a further indication that participants may have increased their reliance on simpler decision rules for cognitive economy with further repetitions. Once again, however, this is speculation based on the patterns observed in reaction times and consistency, and will require more targeted experiments and modeling before such an explanation can be accepted Figure E2.

**Cite this article:** Spicer, J., Mullett, T. L., and Sanborn, A. N. (2024). Repeated risky choices become more consistent with themselves but not expected value, with no effect of matched trial order. *Judgment and Decision Making*, e2. https://doi.org/10.1017/jdm.2023.41