Initial Reply to Banki, Simonsohn, Walatka and Wu (2025) *

Ryan Oprea[†]

March 13, 2025

Abstract

Banki et al. (2025) ("BSWW") show that by simultaneously (i) dropping subjects who made even one error in Oprea (2024)'s training questions (75% of the sample) and (ii) focusing attention on the median, Oprea (2024)'s finding of prospect theoretic behavior in "deterministic mirror" tasks seemingly disappears. I show that it is BSWW's focus on the median within this subsample that drives this result: Oprea (2024)'s findings continue to hold in this subsample at the mean, and non-parametric tests continue to affirm prospect theoretic behavior in mirrors. I present the full distribution of behavior and show that the median is a poor summary statistic for these data: the distributions in BSWW's subsample look quite similar to the full sample, with a clear and significant skew toward prospect-theoretic errors. I also present evidence that BSWW's analysis substantially overestimates the degree of confusion among subjects in Oprea (2024)'s experiment, and show that most of BSWW's further lines of criticism falsify claims that Oprea (2024) does not make and that are not part of his empirical argument that lottery valuations are influenced by complexity.

[*] This note responds to Banki et al. (2025) (draft dated February 7, 2025). I am grateful to the authors of this comment for a thoughtful correspondence and for a number of people who generously read earlier drafts of this response and gave me helpful comments.

[†] Oprea: Economics Department, University of California, Santa Barbara, roprea@gmail.com.

Keywords: complexity, prospect theory
JEL codes: C91, D91, G0

Introduction. In this note I provide an initial response to Banki et al. (2025) (hereafter BSWW) who offer a reassessment and critique of Oprea (2024). I will write a fuller reply in response to a published version but here will only address what I take to be their main critiques. I will also attempt to clarify some misunderstandings about the nature of Oprea (2024)'s claims that I believe are reflected in their comment. I will close with some appreciatory comments that express areas where I agree with BSWW. I make three points.

First, I show that BSWW's key claim (that prospect-theoretic behavior does not appear in deterministic "mirror" tasks) is based on their decision to focus attention on the median, a statistic that poorly represents the distribution of errors in their subsample. It is *not* driven by their decision to remove subjects who made errors in Oprea (2024)'s training questions. Even removing all subjects who made errors in these questions, mirror behavior is still, on average, sharply and highly statistically significantly different from expected value in a prospect-theoretic direction, and mirror and lottery errors remain similarly and strongly correlated across subjects. I show that the median is a knife's edge statistic for these data that does not represent the overall systematic tendencies of errors: over most of the distribution, behavior remains qualitatively (and often quantitatively) similarly prospect theoretic in both lotteries and mirrors, even after removing subjects who made training errors. These analyses suggest that Oprea (2024)'s findings are robust to BSWW's concerns about payoff confusion as measured by errors in the training questions.

Second, I show that BSWW's proposed measure of confusion substantially overestimates the rate of residual payoff confusion in Oprea (2024)'s subjects. I emphasize that the questions in Oprea's experiment that BSWW use to support their claim that his subjects were highly confused were designed to *train* subjects out of initial payoff confusion, not to measure it. I show that these training questions worked as intended in that over the course of the questions subjects' error rate dropped sharply in response to corrective feedback. Given this strong evidence of learning over the course of the questions (and the fact that many of the errors subjects make are inconsistent with payoff confusion), I argue that BSWW's proposal to drop subjects who made *even one* mistake in these questions (leading them to drop 75% of the sample) is based on a substantial overestimate of the level of payoff confusion subjects retained when entering the experiment.

Finally, I argue that BSWW misinterpret the basis of Oprea (2024)'s empirical argument that lotteries are complex. I read BSWW as believing that this conclusion is rooted in an assumption that mirrors and lotteries are "the same" (e.g., that the overall rate of mistakes in mirrors serves as some kind of direct measure of the rate of mistakes that must also be occurring in lotteries), and BSWW spend much of their comment showing counterexamples to this assumption. However, Oprea (2024) does not claim that lotteries and mirrors are identical objects or assume that mirror mistakes rates serve as a direct measure of lottery mistakes rates – to the contrary, although he highlights treatments and tasks in which lottery and mirror behaviors are similar, he also highlights several key ways they are different in the data and explicitly cautions against interpreting the two as being reliably behaviorally identical. Instead, Oprea (2024) bases his conclusion that lotteries, like mirrors, are influenced by complexity on the strong correlation between (i) individual subjects' average propensities to make prospect theoretic mistakes in mirror valuations and (ii) the same subjects' average propensities to express prospect-theoretic deviations from expected value in lottery valuations, which he interprets as evidence that the same error-generating heuristics used in mirrors are likely operating to some extent in lotteries as well. This correlation (unaddressed in BSWW) remains similarly strong when we restrict to subjects who made no training errors, suggesting this, too, is robust to BSWW's concerns.

Summary of Oprea (2024). Oprea (2024) compares the way people value (i) lotteries and (ii) "deterministic mirrors" (hereafter "mirrors"), deterministic payments that are superficially similar to lotteries but that contain no risk.[1] The paper documents two main findings that together form the basis of his main conclusions:[2]

[1] A lottery in the experiment is described as a set of 100 boxes, each containing some amount of money, one of which will be opened randomly to determine payment. A mirror of that lottery is described as the same set of boxes containing the same payments, but subjects are told that the boxes will *all* be opened, summed and divided by the number of boxes to determine payment.

[2] What I will call Finding 1 here is reported in Oprea (2024) in Results 1 and 2. Finding 2 is

- **Finding 1**: Errors in mirrors tend to be biased systematically in the direction of the classical anomalies that inspired prospect theory.
- **Finding 2**: Measures of subjects' average propensity to make prospect theoretic errors in mirrors are highly predictive of (i.e., are highly correlated with) the same measures in lotteries.

Oprea (2024) interprets this as evidence (i) that the patterns described by prospect theory can arise as systematic mistakes in the valuation of lottery-like objects, even in the absence of risk (supported by Finding 1) and (ii) that to a significant extent deviations from expected value observed in lotteries are likely instances of the same sorts of mistakes (supported by Finding 2). He concludes from this that lotteries are not only risky but also tend to be complex (difficult or costly for decision makers to properly value), and that the patterns of prospect theory are to some degree a response to this valuation difficulty rather than to risk, per se.[3]

Summary of Banki et al. (2025). BSWW report that many subjects in Oprea (2024) make at least one error in the training questions included in the experiment's instructions and show that if one drops every subject who made one or more mistakes on these questions (75% of the subjects) from the dataset, and simultaneously switches the statistic used to characterize behavior to the median, Finding 1 seemingly disappears (BSWW do not address Finding 2). BSWW interpret this (and related evidence) as evidence that the results from Oprea (2024) are a consequence of subjects falsely believing that mirrors had lottery incentives (i.e., were "payoff-confused")

reported in Result 4. Oprea (2024) also compares the relative *levels* of prospect-theoretic errors in lotteries and mirrors in his Result 3, highlighting strong similarities in the fourfold pattern in lotteries and mirrors and differences in loss aversion. However, he cautions in his Interpretation section (III) that we should expect such similarities in levels to be fragile. We will discuss the role of similarities in levels, and the minimal role it plays in Oprea (2024)'s conclusions, in more detail in point 3, below.

[3] In his Interpretation Section (III) where he summarizes the conclusions he draws from the experiment, Oprea (2024) emphasizes first that in mirrors "subjects make systematic valuation errors that take the distinctive shape of the classic fourfold pattern of risk and loss aversion, two key empirical regularities in the literature that have inspired a number of behavioral theories of risk preferences" (Finding 1) and then that these "systematic mistakes in mirrors strongly predict the same distinctive behaviors in lotteries, suggesting that the key empirical regularities typically used to measure putative components of preferences like probability weighting, reference dependence and loss aversion in lotteries are likely to a great extent driven by heuristic mistakes as well." (Finding 2) (p. 3805).

rather than a consequence of the fact that mirrors and lotteries are "complex." They also highlight differences between lottery and mirror valuations under several statistics and in several data subsets, and take this as evidence against Oprea (2024)'s empirical argument.

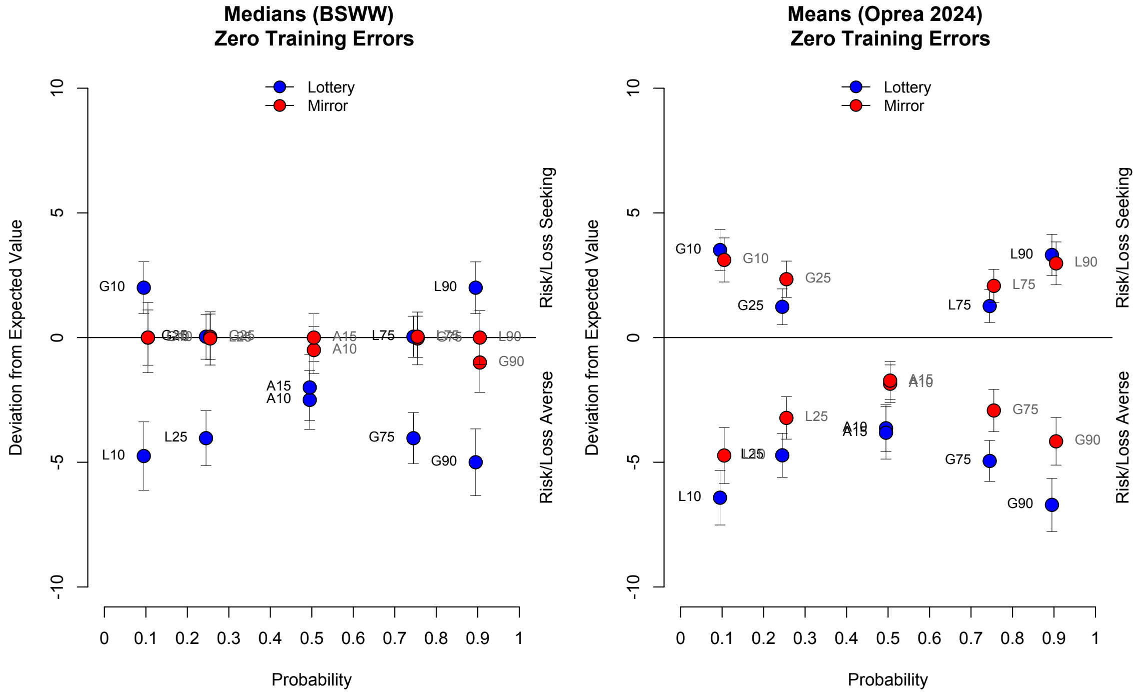
I make three initial replies to this critique.

1. Oprea (2024)'s Findings are Robust to Removing Training Errors. In the left hand panel of Figure 1, I reproduce BSWW's key finding: when they (i) remove subjects who made any training errors (75% of the data) and (ii) *simultaneously* switch the analysis from the mean (as in Oprea 2024) to the median, Finding 1 seems to disappear: median deviations in mirrors (but not lotteries) fall to or near zero in all tasks.[4]

Importantly, however, this apparent reversal of Oprea (2024)'s result is primarily driven not by (i), the removal of subjects who made training errors (the core of BSWW's critique), but rather by (ii), BSWW's decision to simultaneously switch the statistic used to summarize the data from the mean to the median. To see this, in the right panel of Figure 1, I show the same plot using the same sample (zero training error subjects) but using the mean (the statistic used in Oprea (2024)'s original analysis). The plot shows that Oprea (2024)'s key evidence in support of Finding 1 is unaffected by removing subjects who made training errors: even the zero error subjects continue to display systematic evidence of both the fourfold pattern and loss aversion in both mirrors and lotteries. Indeed, in both mirrors and lotteries, in every one of these tasks (e.g., L10, L25, etc.), these deviations from expected value remain highly significant ($p < 0.001$ in each case) using the same Wilcoxon test used in Oprea (2024).

Thus, BSWW's headline result is not that subject confusion (as measured by training

[4] I follow BSWW by pooling data from all treatments included in Oprea (2024)'s dataset. However, I want to emphasize that I have reservations about pooling these data given that they come from experiments that were designed to be different in important ways. I especially have concerns about the inclusion of data from the "error" sample, which comes from experiments with an error in the instructions that lead to their being dropped from Oprea (2024) and were expressly included in the dataset only for transparency. In this initial reply I have included this treatment in my pooled data only to facilitate comparison with BSWW. Throughout my analysis, I omit data from the G50 and L50 tasks which are not part of the classical pattern that are the subject of Oprea (2024)'s claims, and were not included in his primary analysis.



Medians (BSWW)
Zero Training Errors

Means (Oprea 2024)
Zero Training Errors

Figure 1: Median (left panel) and mean (right panel) deviations from expected value in lotteries (blue dots) and mirrors (red dots) for subjects who made zero training errors. *Notes: For fourfold lotteries (G10, G25, G75, G90, L10, L25, L75, L90), the y-axis measures the difference between subjects' valuation and expected value, as stated in the axis label. The x-axis is the probability of the non-zero payoff. For loss aversion tasks (A10, A15), the y-axis measures instead the difference between the certain/simple payoff and the expected value of the mixed lottery/mirror. Two-standard-error bars are included for every task.*

errors) is driving Oprea (2024)'s findings, as the text seems to imply, but instead that switching the analysis to the median within this subset can generate an apparently different conclusion. The key question, then, is whether the median does an accurate job of assessing Oprea (2024)'s claim that deviations from expected value tend to be systematically prospect-theoretic in mirrors.

To help resolve this question and cut through debate about summary statistics, Figure 2 simply plots all of the raw data used in Figure 1 – all deviations from expected value, normalized to be positive if they run in the direction of prospect theory. The left panel plots the pooled dataset and the right panel plots the same dataset restricted to subjects that made zero training errors (following BSWW's proposal). I shade regions green where lottery and mirror distributions are qualitatively both prospect-theoretic (PT), expected value maximizing (EV), or anti-prospect theoretic (Anti-PT); I shade areas red where lottery and mirror distributions are qualitatively

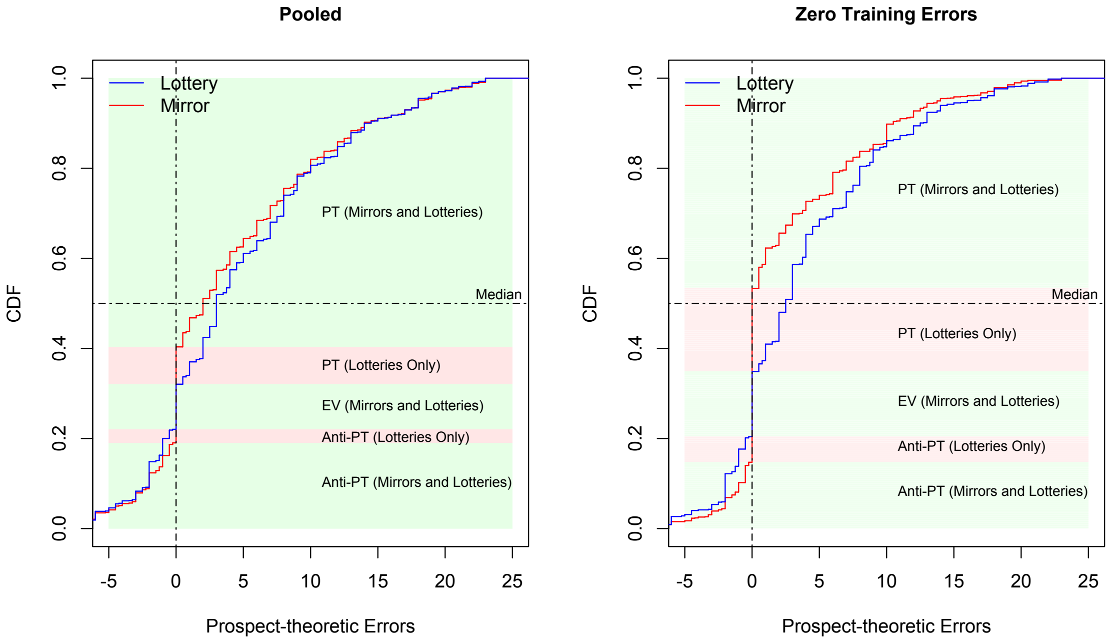


Figure 2: Empirical CDFs of deviations from expected value, normalized to be positive if in the direction of prospect-theoretic predictions. *Notes: Separate CDFs are plotted for lottery and mirror deviations. Regions are shaded green where lottery and mirror distributions are qualitatively both prospect-theoretic (PT), expected value maximizing (EV), or anti-prospect theoretic (Anti-PT); areas are shaded red where lottery and mirror distributions are qualitatively different. The left panel plots the full pooled dataset and the right panel restricts to subjects who made zero training errors.*

different. (In Appendix Figures 9 and 10, I pair each of these pooled plots with a panel of similar plots broken down by task so the reader can verify that the following observations extend to individual tasks.) I make three observations:

1. Over most of the distribution (e.g., at most percentiles), deviations from expected value in mirrors are qualitatively (and often quantitatively) similarly prospect-theoretic in lotteries and mirrors (i.e., the green regions are far larger than the red regions), *whether or not we restrict the sample.*

2. Deviations from expected value (when they occur) are systematically and similarly biased in a prospect theoretic direction (i.e., the green region at the top of each plot showing mirror prospect-theoretic errors is far larger than the green region at the bottom showing errors in the opposite direction) in both lotteries and mirrors (about 75% of deviations in each case are prospect-theoretic rather than the reverse), *whether or not we restrict the sample.*

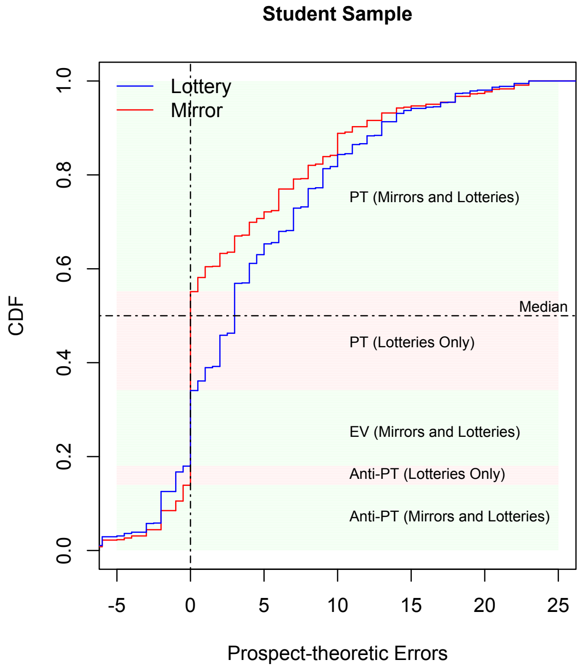


Figure 3: Empirical CDFs of deviations from expected value, normalized to be positive if in the direction of prospect-theoretic predictions, **student sample**. *Notes: Separate CDFs are plotted for lottery and mirror deviations. Regions are shaded green where lottery and mirror distributions are qualitatively both prospect-theoretic (PT), expected value maximizing (EV) or anti-prospect theoretic (Anti-PT); areas are shaded red where lottery and mirror distributions are qualitatively different.*

3. The primary difference between lottery and mirror valuations (and the key change that occurs when restricting the sample) is the relative rate at which subjects value mirrors/lotteries at EV. (As I show in Appendix Figure 8, if we remove differences in rates of EV valuations, the distributions of errors are identical.) But as I emphasize in point 3 below, the similarity in the rates of deviations from EV has little to do with Oprea (2024)'s key claim: that deviations from expected value tend to be systematically prospect theoretic in mirrors (Finding 1) and predict the same tendencies in lotteries (Finding 2).

In Figure 3, I show that the same observations hold also in Oprea (2024)'s student sample: errors remain systematically and generally similarly prospect-theoretic in both lotteries and mirrors and the median does a poor job of showing this.

Figure 2 shows that the median (plotted as a horizontal line at 0.5) is a knife's edge statistic in these data (a small change in the mirror CDF would disrupt their finding that the CDF intersects the median at 0) and does a poor job of representing the overall systematic tendencies of prospect theoretic errors (the rate of expected valuation, another statistic favored by BSWW, is similarly poor at representing the systematic prospect-theoretic nature of deviations). Over *most* of the distribution, errors clearly tend to be systematically prospect-theoretic in both mirrors and lotteries and statistics like the median and rates of expected valuation simply do a poor job of representing this systematicity — the mean, studied in Oprea (2024), by comparison, does a more accurate job.

To better understand how BSWW's choice of (i) a rule for discarding data and (ii) a statistic for summarizing what remains are jointly responsible for their stark conclusions, Figure 4 plots the percentiles of the distribution of errors (normalized positive, again, if in the direction of prospect theory) on the x axis, and (approximate) quartiles of training errors on the y-axis (following BSWW's own binning proposal) using this same data. The green area represents parts of the distribution in which mirror and lottery errors are *both* prospect-theoretic and the red area represents parts of the distribution in which lotteries but not mirrors are prospect-theoretic; the white area represents parts of the distribution in which neither lottery nor mirror errors are prospect-theoretic. Again, I make three observations.

First, over most of the distribution of valuations and training question errors, lottery and mirror choices agree: when *either* lottery or mirror errors are prospect theoretic, *both* tend to be prospect-theoretic (or not). The green area is (and even more so the green and white areas together are) far larger than the red area on this graph. The area where mirrors and lotteries are qualitatively different (in red) is a small portion of the distribution. Second, at zero training errors, the median (plotted as a dashed vertical line) happens to run through the small part of the distribution at which one can observe prospect theoretic errors in lotteries but not in mirrors (i.e., it runs through a red region that is far smaller than the green or even white regions). Third, this finding is itself dependent on BSWW having restricted themselves to the 1/4 of the data (subjects who made no training errors at all) at which the median coincides with prospect-theoretic errors in lotteries but not in mirrors: for the remaining 75% of the data, at the median errors are qualitatively prospect-theoretic in both lotteries and mirrors (i.e., the dashed line runs through a green region except at

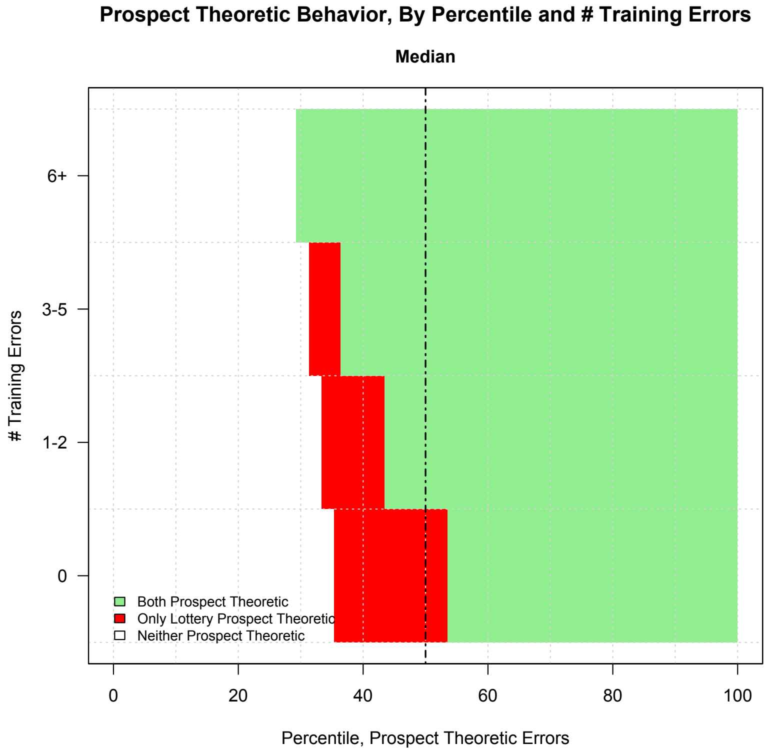


Figure 4: Percentiles of prospect theoretic errors (x axis) against number of training errors (y axis). *Notes: In green is plotted regions in which both mirror and lottery errors are prospect-theoretic, in red in which only lottery errors are prospect-theoretic, in white in which neither is prospect-theoretic.*

zero training errors). Looking even at subjects who made only 1 or 2 errors (instead of the maximally strict criterion of zero errors selected by BSWW), for instance, the median shows qualitatively similar behavior in mirrors and lotteries.

This analysis suggests that BSWW's findings are unrepresentative of the overall character of the mirror data, and do not well-represent the qualitative agreement of prospect-theoretic errors in lotteries and mirrors. It is BSWW's joint choice of a maximally severe rule for discarding data (i.e., the decision to discard subjects who make *any* training questions error at all) and of a statistic (i.e., the median) that happens to overlap a non-representative region of the distribution that is responsible for their striking conclusion – a conclusion that is at odds with the data as a whole.

Finally, not only is Finding 1 robust to the elimination of training errors, Finding 2 (unaddressed in BSWW) is as well. As in Oprea (2024), I calculate a measure of

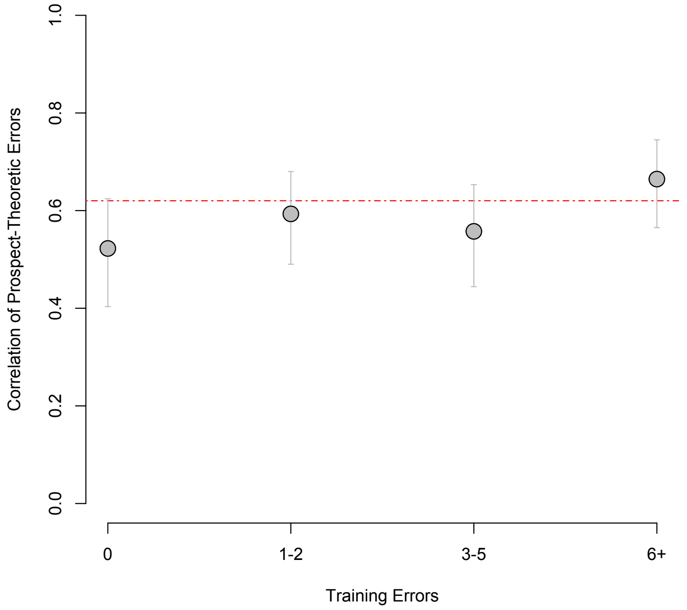


Figure 5: Correlation coefficients between lottery and mirror average prospect-theoretic errors as a function of training question errors. *Notes: Pearson's correlation coefficients between subject-wise mean lottery and mirror errors (normalized to be positive if in the direction of prospect-theoretic errors). Binning of training questions errors follows BSWW's binning.*

each individual subject's overall propensities to make prospect theoretic errors (their mean deviation from expected value, normalized to be positive if in the direction of prospect theory, calculated over tasks) in both lotteries and mirrors. Figure 5 plots the (Pearson's) correlation coefficient between this measure in lotteries and mirrors as a function of training errors (a dashed line shows the correlation coefficient for the sample as a whole). The correlation coefficient remains similarly strong regardless of the number of training errors and restricting to subjects who make zero errors has little impact on this correlation. Prospect theoretic tendencies in mirrors (where they are mistakes) thus strongly predict the same tendencies in lotteries *whether or not we restrict the sample.* As I emphasize in more depth in point 3 below, this subject-to-subject correlation (and not the overall similarity in the rate of deviations from expected value in mirrors and lotteries) is a key basis for Oprea (2024)'s conclusion that the heuristic mistakes that occur in mirrors likely occur in lotteries as well.

To conclude, deviations from expected value are systematically prospect theoretic in mirrors (Finding 1) and subject level propensities to make these prospect-theoretic errors are highly statistically correlated with the same propensities in lotteries (Finding 2), even when truncating the sample in the severe way proposed by BSWW. Even if we were to interpret training errors as clear evidence of payoff confusion (an interpretation I will argue against below), there is little reason to conclude based on these data that Oprea (2024)'s primary findings are artifacts of payoff confusion.

2. Training Errors Largely Measure Cognitive Effort, Not Payoff Confusion. At the center of BSWW's critique is a claim that the total number of errors subjects make in Oprea (2024)'s training questions serves as a measure of subjects' payoff confusion in the experiment - of subjects mistakenly believing they are facing lottery incentives when in mirror tasks. Based on these data, they conclude that the vast majority of subjects were payoff confused during Oprea (2024)'s experiment. However, it is important to emphasize that these training questions weren't designed to measure beliefs (e.g., payoff confusion), and because of this they are poorly suited to the task BSWW repurpose it for, ex post. Indeed, evidence from the patterns of mistakes made in these questions suggests that overall training errors largely serve as a measure of the cognitive effort (an important ingredient in Oprea (2024)'s account) subjects apply to answering these questions, and that BSWW therefore substantially overestimate the level of payoff confusion with which subjects entered the experiment.

These questions were included in Oprea (2024)'s instructions in order to train subjects by correcting errors they make when attempting to apply the mirror or lottery payoff rule (a purpose that is clearly expressed in the paper).[5] During both the lottery and the mirror instructions, subjects were given four different questions[6] in sequence, one at a time, and only moved onto the next question after successfully answering the previous one.[7] Because the questions were designed to train away payoff confusion,

[5] In Online Appendix B.2, Oprea (2024) emphasizes "After subjects have completed the first treatment (Mirror or Lottery) and have read instructions for the next treatment, they are given the same four comprehension questions, now with different correct answers. This makes the difference between the payment schemes especially salient to subjects and is designed to prevent subjects from confusing payoffs in the two treatments."

[6] In the student sample, subjects were given 16 questions. I follow BSWW by focusing on the four questions shared in all of the experiments. However, in Appendix Figure 11 I plot error rates from all 16 questions in the student quiz.

[7] All four questions concerned a choice between a lottery/mirror A that consist of 50 boxes con-

not measure it, subjects (i) were allowed to submit as many wrong answers as they liked on each question, (ii) were given feedback on their mistakes (which they were required to correct before proceeding to the next question) and (iii) were given no motivation (or even pressure) to get answers to the questions correct the first time. Importantly, because answering each of these questions correctly the first time requires cognitive effort (particularly in the deliberately harder mirror training questions),[8] subjects have motivation to avoid these costs by submitting low effort (e.g., noisy, inattentive or unreflective) initial answers and learn by correction. Because of this motivational structure, when repurposed as a measurement tool, overall errors in the training questions to a great extent measure subjects' willingness to expend the cognitive effort required to answer these questions correctly the first time, rather than payoff confusion.

By treating training errors (including those from the beginning of training) as evidence of the payoff confusion subjects carry with them into the experiment, BSWW effectively assume that the questions failed in the purpose they were designed for – to train away payoff confusion.[9] But, in fact there is very clear (and very direct) evidence from within the training question data itself that it was highly successful in solidifying subjects' understanding of the payoff rule: over the course of the four sequential questions, error rates drop substantially, suggesting that subjects learned a great deal about the structure of the payoff rule from the correction they received in response to errors made in previous questions.

taining \$16 and 50 boxes containing \$0, and a lottery/mirror B that consists of 100 boxes containing \$4. Subjects are asked four questions about lottery/mirror A – what is the chance that: #1 \$16 is added to your earnings, #2 \$8 is added to your earnings, #3 \$4 is added to your earnings \$4 and #4 \$0 is added to your earnings. In each case subjects can answer: (a) 0% chance, (b) 50% chance or (c) 100% chance. Subjects receive the same four questions under lotteries and mirrors; in the former the correct sequence of answers is b, a, a, b while in the latter it is a, c, a, a.

[8] Importantly, the training questions were structured in such a way that answering mirror questions correctly requires somewhat more cognitive effort than answering lottery questions correctly. Indeed, the obvious, unreflective answer in the questions included in the training questions will tend to produce evident payoff confusion in mirrors but not in lotteries.

[9] BSWW acknowledge this possibility, but propose to test whether the training questions failed by studying whether training errors predict evidence of mistakes in lotteries and mirrors. But this is not a test of the training efficacy of these questions for the simple reason that to the extent training errors measure low cognitive effort (rather than residual payoff confusion), they should continue to predict the same mistakes under Oprea (2024)'s interpretation as well *even if subjects suffer no payoff confusion at all.*

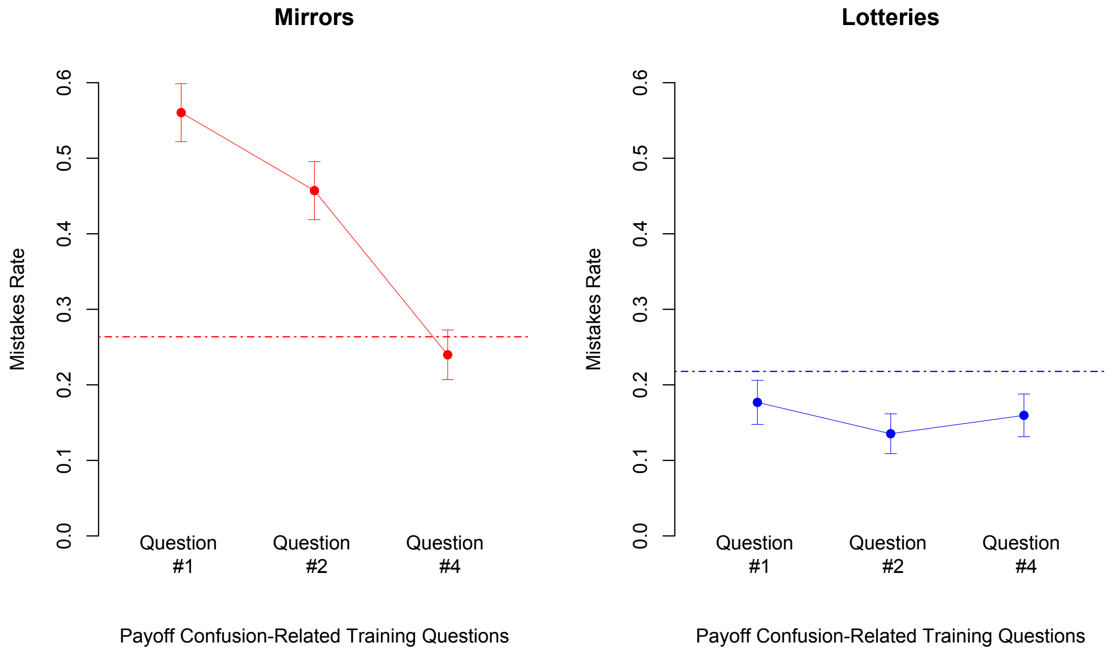


Figure 6: Rates at which subjects make at least one mistake in each of the three training questions concerning payoff confusion. *Notes: The solid line plots mistakes rates for each of the three training questions that relate to payoff confusion. The dashed horizontal line shows the mistakes rate from the question (question #3) that does not measure payoff confusion and therefore serves as a benchmark for mere low effort behavior. The panel on the left shows data from the mirror training questions and on the right from the lottery training questions.*

Figure 6 shows a time series of error rates[10] (for both the mirror and lottery questions) in each of the three separate, sequentially presented questions (#1, #2 and #4) that test and correct payoff confusion (BSWW show similar data with the same patterns in their Appendix figures A.5 and A.6). The dashed horizontal line shows the error rate from an additional question (#3) that has nothing to do with payoff confusion and therefore serves as a benchmark of the error rate we should expect based on the use of low effort (e.g., noisy or inattentive) answering strategies alone (BSWW interpret this question similarly in their appendix). If subjects learned nothing from question to question during the training, we would expect this series to be flat (for error rates to be similar across questions). But (focusing on the mirror questions on the left), error rates clearly fall substantially from question to question over the course of the training, dropping by more than half, suggesting that subjects in fact learned a great

[10] Specifically, the rate at which subjects made at least one error.

deal from the correction they received over the course of these questions. By the final question,[11] due to this learning, mirror errors are similar to lottery errors and most importantly are slightly *below* the benchmark error rate from a question that measures low effort but not payoff confusion. In Appendix Figure 11, I show that this learning over the course of these four questions was lasting: in the student sample in which subjects were given an additional twelve training questions to follow the first four, error rates remain very low, at or below the low effort benchmark. Given this evidence of durable learning, the training data itself gives us little reason to believe that subjects *entered the experiment* with much remaining payoff confusion at all.

In light of this, it is important to emphasize that in repurposing Oprea (2024)'s training questions as a measurement tool, BSWW had significant latitude in how to use subjects' answers to measure payoff confusion. By using *all* training errors as their measure, they include many errors that have little to do with payoff confusion, errors that were likely random or inattentive, and errors that the training data itself suggests were later learned away, greatly exaggerating the degree to which the training data itself actually indicates payoff confusion that survives in the Oprea (2024) behavioral data. For instance, if BSWW wanted to precisely target payoff confusion, they might have used only errors that are actually consistent with payoff confusion, which would have reduced the number of errors treated as evidence of payoff confusion *by more than half.*[12] Or if they wanted to focus on the payoff confusion subjects plausibly retained upon entering the experiment (after the training had had its effect), they might have restricted to those plausibly payoff-confused errors subjects continued to make at the end of the training questions (i.e., in question #4), which would have cut the number of errors treated as evidence of payoff confusion *by an order of magnitude.*[13] Or they might have restricted to subjects who actually *consistently* made payoff confused errors (i.e., in each of the relevant training questions) – subjects whose answers are consistent with the sort of deep-seated confusion about the payoff function BSWW seem concerned about – which also would have cut down the number of subjects

[11] Importantly, the final questions (#4) is extremely similar to question #1 but concerns a different payoff state making comparison of the two ideal to measure learning.

[12] Only 46% of errors in the mirror training questions and 42% of errors in the lottery training questions are actually consistent with payoff confusion.

[13] Only 12% of errors made in the training questions are actually made in the final question (question #4) and are consistent with payoff confusion.

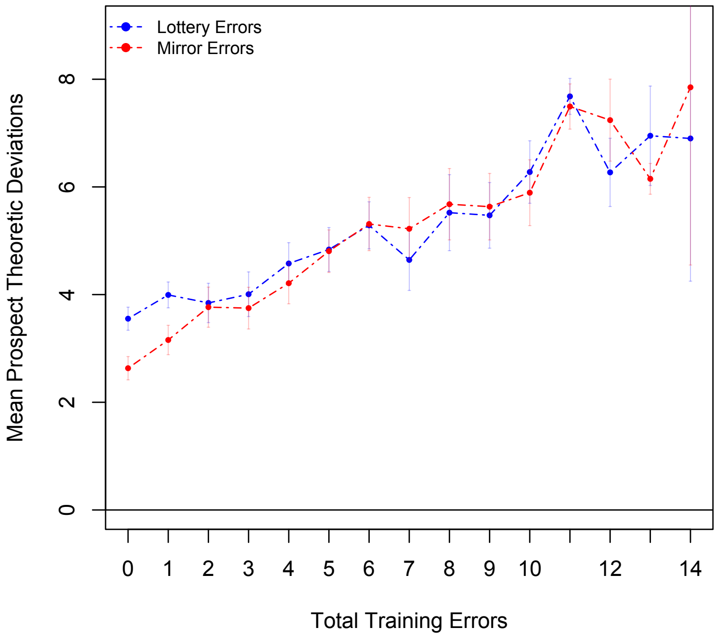


Figure 7: Mean deviations from expected value (normalized positive if in the direction of prospect-theoretic predictions) as a function of training question errors. *Notes: Means of lotteries and mirrors are plotted in blue and red dots, respectively.*

measured as payoff confused *by nearly an order of magnitude.*[14] The decision to instead treat all errors as evidence of payoff confusion is consequential for BSWW's conclusions: as we show in Appendix Figure 12, if we were to instead discard subjects based on any of these more sharply targeted criteria, lottery and mirror behaviors become yet more similar, and even the median mirror decision becomes prospect theoretic in each case.

By choosing instead to use *all* training errors to measure payoff confusion, BSWW greatly overestimate the level of payoff confusion actually suggested by the training data, and likely misattribute behaviors to residual payoff confusion that are due instead to the exact kind of low effort behavior underlying Oprea (2024)'s complexity-based interpretation of his data. Indeed, the data seems to support this hypothesis.

[14] Only 10% of subjects under mirrors and 3% of subjects under lotteries make training errors that are consistent with payoff confusion in all three of the training questions questions (#1, #2 and #3) capable of measuring it.

In Figure 7, I plot mean prospect theoretic errors (deviations from expected value normalized positive if in the direction of prospect theory) as a function of total training errors. The plot shows that prospect-theoretic errors strongly increase with training errors in mirrors, but crucially that they increase to a remarkably similar degree in lotteries too. This similar relationship is consistent with this variable measuring subjects' propensity for low cognitive effort (e.g., noisy or inattentive behavior) under Oprea (2024)'s interpretation of prospect-theoretic errors, but is starkly inconsistent with the idea that it measures mirror subjects' propensity to falsely believe they are in lotteries (i.e., payoff confusion).[15] In Appendix Figure 13, I include this plot alongside similar plots of a number of other measures of cognitive effort and ability studied in Oprea (2024) (Online Appendix A.5.), which show similar evidence that low cognitive effort/ability predicts prospect-theoretic errors similarly in lotteries and mirrors. Looking at training errors in the context of these many other cognitive measures, it is difficult not to conclude that they largely measures the same sort of low cognitive effort and ability Oprea (2024) highlights as evidence *for* his interpretation, rather than evidence against it.[16]

3. Oprea (2024)'s Claim is Not that Lotteries and Mirrors are "The Same."

The remainder of BSWW's critique consists mainly of a series of challenges to the

[15] If training errors measured payoff confusion leading to prospect-theoretic behavior in mirrors (BSWW's primary concern), we would expect to see a relationship between training errors and prospect theoretic behavior in mirrors but not in lotteries. If it measured payoff confusion in both mirrors and lotteries, we would expect the relationship to be *decreasing* in lotteries and increasing in mirrors. Instead, we find a pattern of similarly increasing prospect-theoretic behavior in *both* lotteries and mirrors, which seems inconsistent with *any* payoff confusion story.

[16] Interestingly BSWW conduct a related analysis in their Section 2.5 in which they attempt to *decompose* the fraction of prospect theoretic error that is driven by tastes for risk versus mistakes. This analysis seemingly abandons their earlier concern with payoff confusion (i.e., the claim that mirror behavior represents confused subjects mis-expressing their true tastes for risk), and instead interprets mirror behavior as a measure of mistakes that to some extent occur also in lotteries – an interpretation that is very similar to the one offered in Oprea (2024). BSWW produce a variation on Figure 7 plotting the *difference* between mirror and lottery prospect theoretic errors. Unsurprisingly given how similar mirror and lottery behavior is in Figure 7, this analysis suggests that the behavior that survives this decomposition and that can therefore be treated as possible measures of tastes for risk is very small relative to the raw prospect-theoretic deviations in lotteries pictured in this graph, suggesting that the majority of these errors are due to heuristic mistakes rather than rational expressions of tastes. BSWW treat this evidence as a critique of Oprea (2024), but it seems to accept most of Oprea (2024)'s interpretation of mirror behavior and largely reinforces his primary conclusions about lottery behavior.

claim (which they attribute to Oprea 2024) that “lotteries and mirrors are the same” (p. 25) and “mirror and lottery valuations should largely move together” (p. 20). However, Oprea (2024) does not make this claim. Although Oprea (2024) certainly highlights striking examples in his data in which lottery and mirror behavior are very similar, he also highlights a number of counterexamples in which they differ and explicitly cautions readers that we should not expect such similarities to be general.[17] Most importantly, the empirical argument underlying Oprea (2024) does not depend on an assumption that mirrors and lotteries are the same. To make this clear, I will recapitulate Oprea (2024)’s empirical argument and explain why it is perfectly consistent with BSWW’s (and Oprea (2024)’s) finding of some average differences in mirror and lottery valuations.

The purpose of a mirror in Oprea (2024) is *not* to measure the *rate* at which subjects make mistakes in deterministic valuation tasks, or to use this rate as some kind of measure of lottery complexity. Indeed, as Oprea (2024) emphasizes, we should expect the rate at which subjects make such mistakes in mirrors to vary with framing and access to auxiliary information, meaning the rate of mistakes in any given implementation of a mirror task isn’t likely to be stable or robust enough to tell us anything very general.[18] What’s more, as Oprea (2024) also emphasizes, we have little reason to believe that mirrors are identically complex (i.e., difficult to value) as lotteries, making evidence of some given rate of mistakes in the former poor direct evidence of

[17] I suspect that the striking similarities in levels of prospect theoretic behavior in several key places in Oprea (2024)’s data, combined with the fact that Oprea (2024) confined most of his interpretation of the implications to a later section of the paper (Section III) contribute to BSWW’s impression that this similarity is central to his conclusions. I also fear that the title of the paper contributed to the confusion. “Decisions Under Risk are Decisions Under Complexity” was meant to express what I think is implied by the plain language meaning of the title, that decisions under risk tend to also be decisions under complexity, *not* that decisions under risk are *only* influenced by complexity (or worse, that decisions in lotteries are decisions in mirrors). I, of course, take ultimate responsibility for any misunderstandings of my writing and am glad to have the opportunity here to correct any such misunderstandings.

[18] Oprea (2024) writes: “It is possible that some alternative framings of deterministic mirrors might make it less difficult or costly to infer the mirror’s true value, attenuating or eliminating this effect.” (p. 3805) He also emphasizes that this is likely more true of mirrors than lotteries giving the example: “if we were to directly tell subjects in our experiment the expected value, we would expect the classical pattern to shrink much more in mirrors than in lotteries. This is because knowing the expected value directly removes the difficulties of valuation in mirrors (since the value of mirrors *just is* the expected value), but it doesn’t in any clear way remove those same difficulties in lotteries.” (p.3805)

the rate of mistakes that must also be occurring in the latter.[19]

Instead the purpose of a mirror is to study the *nature* of the mistakes people make in deterministic tasks that resemble lotteries, when they occur. In particular, the mirror is meant, in Oprea (2024), to serve as a device for measuring whether the cognitive shortcuts people use in superficially lottery-like settings tend to produce errors (when errors *do* occur) that systematically take the distinctive shape of prospect theory. Its value is that, because it contains no risk (and therefore no scope for the rational expression of tastes for risk), any deviations from expected value that occur in a mirror are unambiguous mistakes, and we can therefore use it to measure whether the mistakes people make in tasks that share superficial characteristics of lottery valuation tasks tend to be systematically biased in the direction of prospect theory (rather than being, e.g., symmetrically distributed around expected value).

To the degree errors are, indeed, biased in the direction of prospect theory in mirrors, the mirror produces some direct evidence in favor of a long-standing hypothesis in the prospect theory literature that prospect theoretic behavior may to some extent describe the cognitive shortcuts people use in response to the difficulty of proper valuation, rather than rational expressions of people's welfare-relevant tastes for risk.[20] The shortcuts proposed in this literature[21] are usually hypothesized to be responsive

[19] For instance, Oprea (2024) emphasizes that there is recent evidence that stochasticity makes reasoning more difficult even in isomorphic tasks (Martinez-Marquina et al., 2019). In particular, he writes that such results "suggest that problems involving risk are often more complex than isomorphic deterministic problems, generating more mistakes. This provides a useful caution in interpreting our results because it suggests that the complexity of lotteries may be greater than that of mirrors in at least some decision settings. To the degree this is true, we should view the decomposition afforded by our approach as providing a conservative, lower bound estimate of the role complexity plays in driving lottery anomalies—in at least some cases, we should expect lottery errors to be more severe than mirror errors, even if errors in each are produced by complexity alone." (p. 3806)

[20] This is a very old hypothesis in the prospect theory literature, and one that the literature has openly entertained since the beginning. For instance, it is difficult not to read Kahneman and Tversky (1979) as advocating this interpretation, and Tversky and Kahneman (1992) directly highlight the resemblance of prospect-theoretic patterns to psychophysical distortions from the perception literature. The ambivalence over the interpretation of prospect theory (as description of welfare-relevant tastes, or a collection of cognitive shortcuts) is reflected in recent discussions of core aspects of prospect theory like probability weighting ("there is relatively limited discussion or consensus about the psychological principles that underlie probability weighting" (O'Donoghue and Somerville, 2018)) and loss aversion ("[a]n important open question about loss aversion is whether it is a judgement error or a genuine expression of preference." (Camerer, 2005))

[21] For example, these include models that root prospect theoretic anomalies in the use of low-attention strategies (Bordalo et al., 2012), noisy valuation (Blavatskyy, 2007; Enke and Graeber,

to the same superficial characteristics (e.g., numerical information, the task's action space, etc.) mirrors are deliberately designed to share with lotteries allowing mirrors to serve as a device for empirically measuring how such shortcuts influence valuation. Oprea (2024)'s finding that, indeed, errors tend to be systematically prospect theoretic in mirrors is supportive evidence for this hypothesis and is the substance of his Finding 1, as described above.

Importantly, Oprea (2024)'s evidence for the *further* conclusion that prospect-theoretic behavior in lotteries is *also* to some degree driven by complexity, is *not* based on an assumption that mistakes rates (or deviations from expected value) are or should be identical across lotteries and mirrors. It is not, for instance, based on a claim that, because lotteries and mirrors are the same, aggregate mistakes rates in mirrors (i.e., rates of deviation from expected value) therefore serve as some kind of direct empirical measure of the rate of complexity-derived mistakes that must also be occurring in lotteries. Oprea (2024) emphasizes reasons that such a direct inference can't be supported, in part because there is no guarantee that mirrors and lotteries are identically complex (identically difficult to properly value) and in part because lotteries create additional potential scope for the rational expression of risk preferences.[22]

Instead, Oprea (2024)'s argument that prospect theoretic behavior in lotteries is at least in part a response to complexity is rooted in the extremely strong correlation he finds between subjects' average propensities for making prospect-theoretic errors in mirrors and lotteries (i.e., Finding 2).[23] If a propensity to make prospect theoretic

2023; Gabaix, 2019), noisy cognition or perception (Woodford, 2012; Steiner and Stewart, 2016; Vieider, 2024; Frydman and Jin, 2023), constrained memory sampling (Friedman, 1989; Stewart et al., 2006) or the use of heuristic rules of thumb (Rubinstein, 1988; Gabaix, 2014) – all responses to superficial descriptive characteristic of lottery choice problems that avoid the complexity of accessing or expressing one's valuations rationally in choice.

[22] "We do not claim, for instance, on the basis of these data that risk preferences or even loss preferences do not exist, but only that they are unlikely to be reliably revealed in lottery valuations. Indeed, our finding of potentially stronger loss aversion in lotteries than mirrors might even be evidence that true loss averse preferences act as a secondary driver of loss aversion in lotteries on top of the mistakes that exclusively drive the same phenomenon in mirrors..." (p. 3804)

[23] When asking whether lottery anomalies occur due to complexity, as they do in mirrors, Oprea (2024) writes "To what degree do these anomalies (the fourfold pattern and loss aversion) occur in risky lotteries and riskless mirrors for the same reason, driven by the same behavioral mechanism? To study this, we make use of our within-subjects design and examine the statistical relationship between anomalous behavior in mirrors and lotteries across subjects. Since there is no risk in mirrors, to the extent that evidence of anomalies is strongly *correlated* in lotteries and mirrors, we have

errors (which can only occur in mirrors due to the use of non-maximizing heuristic behaviors) *predict* the same propensity to deviate from expected value in lotteries from subject to subject, it suggests that similar heuristic behaviors are likely being used in lotteries too – and therefore that subjects find valuing lotteries, to some extent, complex as well. This correlation is very strong (ρ is consistently above 0.5) in all treatments in the Oprea (2024) data, and as emphasized above, is virtually unchanged when restricting to subjects that make zero training errors. This subject-to-subject correlational evidence that similar heuristic behaviors are used in lotteries and mirrors is the substance of Oprea (2024)'s Finding 2, as described above.

Crucially, this line of argumentation does not depend on an assumption that levels of mistakes in lotteries and mirrors are identical, that lotteries and mirrors are always identically complex, that lotteries and mirrors are complex for identical reasons, that preferences for risk don't exist, or that the interventions that are effective at making lotteries or mirrors easier to value are always the same. This is because a mirror is not meant to be a stand-in or complete experimental "model" for a lottery in the Oprea (2024) design. Its diagnostic purpose is far more modest – it is meant to be an empirical tool for collecting errors in superficially lottery-like choice problems and examining whether these errors tend to be biased in the distinctive direction of prospect theory, as suggested by one of the classical interpretations of prospect theory. In order to be effective in this task, a mirror need only be (i) costly or difficult enough to value to inspire significant use of cognitive shortcuts, and (ii) for the superficial shape of the mirror (i.e., the numbers used, the presentation of information, the action space) to resemble corresponding lotteries closely enough for those shortcuts to plausibly induce similar behaviors. To the degree errors (at whatever rate they occur) tend to be skewed in the direction of prospect theory, we have evidence that the shortcuts

evidence that they are likely both driven by the complexity of evaluation (the property lotteries and mirrors share) rather than by risk or risk preferences (a property absent from mirrors)." (p. 3800) When summarizing the conclusion that lotteries are influenced by complexity in his interpretation section (III), Oprea (2024) writes: "These systematic mistakes in mirrors strongly predict the same distinctive behaviors in lotteries, suggesting that the key empirical regularities typically used to measure putative components of preferences like probability weighting, reference dependence, and loss aversion in lotteries are likely to a great extent driven by heuristic mistakes as well." (p. 3804). Note, this doesn't appeal to a similarity in levels or a tight ex ante assumption about the relationship between lotteries and mirrors, but rather to the *empirical* subject-wise correlation between average subject-wise mirror and lottery errors.

subjects use in these settings tend to produce prospect-theoretic behavior. To the degree tendencies to make these errors are subject-to-subject correlated with similar behavioral tendencies in lotteries, we have some evidence that prospect-theoretic behavior in lotteries may to some extent represent a heuristic response to the difficulty of valuing lotteries as well.

Again, Oprea (2024) highlights similar levels of prospect theoretic behavior in mirrors and lotteries in *some* of his tasks (e.g., in the fourfold pattern in some treatments), but he explicitly cautions that there is not much of a basis for expecting this similarity in levels to be general.[24] He also documents several key differences between mirror and lottery behavior in his data, throughout, to emphasize these differences. For instance, he emphasizes that levels of loss aversion seem stronger in lotteries than mirrors.[25] Likewise, he emphasizes that there are larger gaps between levels of prospect theoretic errors in lotteries and mirrors in his student sample (even in the fourfold pattern).[26] Indeed, he emphasizes that prospect theoretic behaviors arise in lotteries but *not* in mirrors for a full 14% of subjects in his main treatment.[27] Thus, throughout, Oprea (2024) highlights the same sorts of differences between mirrors and lotteries that BSWW do.

As I discuss in point 1 above, the main difference between lottery and mirror valuations, is that subjects sometimes value mirrors at expected value at a higher rate than lotteries (otherwise, the distributions of valuations are remarkably similar). This is highly consistent with the key differences between lotteries and mirrors emphasized by Oprea (2024),[28] though Oprea (2024) does not attempt to fully resolve this ques-

[24] For instance Oprea (2024) writes: "It is possible that some alternative framings of deterministic mirrors might make it less difficult or costly to infer the mirror's true value, attenuating or eliminating this effect. On the other hand, we would not be surprised ... if this were less true of lotteries: risk itself may make inferring true value difficult, regardless of the framing. As a result, there may well be some settings in which there is a larger wedge between lottery and mirror behavior than in our experiment." (p. 3805)

[25] "a straight reading of the evidence suggests that loss aversion may be somewhat weaker on average (perhaps 80 percent as strong) in mirrors than it is in lotteries" (p. 3799)

[26] "However, the fourfold pattern and especially loss aversion are somewhat weaker in mirrors relative to lotteries than in our other treatments (the fourfold pattern shrinks to 82 percent as strong and loss aversion 54 percent as strong)," (p. 3803).

[27] "Standard risk preference–based theories predict that subjects' tendencies to exhibit the pattern will be risk sensitive but complexity insensitive. We find that only 14 percent of subjects can be classified this way, exhibiting the pattern in lotteries but not in mirrors," (p. 3801).

[28] First, as Oprea (2024) emphasizes, even if complexity has a significant (or even dominant) role

tion and it is therefore an open question for the future literature to answer. What is important to emphasize relative to BSWW's secondary line of critique is that the appearance of differences between mirrors and lotteries are (i) openly documented in Oprea (2024), (ii) are emphasized in his interpretation (especially in Section III) and (iii) are not confounds to the diagnostic structure of Oprea (2024)'s empirical argument.

Discussion. In this initial response to Banki et al. (2025), I have made three points: (i) Oprea (2024)'s primary claims are robust to removing subjects who made errors in the experiment's training questions, following BSWW's sampling proposal; (ii) data from the training questions in Oprea (2024) gives us little reason to think that many subjects were confused about the nature of the mirror payoff function when they entered the experiment; and (iii) Oprea (2024)'s empirical argument does not, as BSWW claim, rely on an assumption that aggregate error rates in lotteries and mirrors are reliably similar. Given this, I conclude that Oprea (2024)'s interpretation of his data is robust to the primary lines of criticism raised by BSWW.[29]

I want to close this initial response on an appreciatory note, by emphasizing several

in producing the distinctive patterns of prospect theory, this in no sense implies that humans are risk neutral. If we believe that tastes for risk remain to influence choice, we have little reason to expect even maximizing subjects to value lotteries at their EV at a high rate. By contrast, mirrors *induce* risk neutral preferences, meaning we should expect *every* rational subject to value mirrors in this way. Second, and more speculatively, there are also some reasons to believe lotteries may be more difficult to value than mirrors which, if true, might generate similar effects on rates of expected valuation. Oprea (2024) bases this speculation on the recent findings of Martinez-Marquina et al. (2019) who provide evidence that stochasticity makes reasoning significantly more difficult in otherwise isomorphic tasks. If, indeed, lotteries are more complex than their mirrors, we should expect even risk neutral subjects to have a harder time expressing those preferences properly, and for the rate of improvement with increases in sophistication to be substantially muted relative to mirrors. Either of these differences between lotteries and mirrors are potential reasons that expected valuation rates might sometimes differ across the two settings (especially for more sophisticated or effortful subjects), and both are perfectly consistent with the interpretation offered in Oprea (2024).

[29] The main argument in BSWW that I have not yet addressed in this initial reply is their claim that FOSD violations are unusually high in the Oprea (2024) data. There are two reasons for this. The first is that I wanted to address the comment's main points in a timely manner in this note, and fully assessing their claims on FOSD violation rates in the literature will take significant additional time (largely to expand on the small selection of studies they use to make their case). The second is that, as I have shown, Oprea (2024)'s findings remain similar even for the subset of the data for which FOSD violations are lowest, meaning it is likely a second order consideration for evaluating Oprea (2024)'s findings. Indeed the only other published paper that uses a "mirror design" like Oprea (2024)'s, Vieider (2024), features *low* FOSD violations, but nonetheless reports very similar findings to Oprea (2024)'s.

points on which BSWW and I agree.

First, much of BSWW is devoted to highlighting that lotteries and mirrors differ from one another in some statistics (e.g., rates of expected valuation), and that these statistics sometimes change differently with changes in subject sophistication. As I emphasize in my point 3 above, Oprea (2024) highlights similar gaps between lottery and mirror behavior and emphasizes important reasons to expect error rates in the two cases to differ. Nonetheless, because I gather that it is tempting to misread Oprea (2024)'s argument as hinging on a claim that lotteries and mirrors are "the same" I am genuinely grateful to have this potential misunderstanding punctured further, and to have an opportunity here to correct confusions about the empirical argument underlying Oprea (2024).

Second, in their conclusion, BSWW highlight that the results in Oprea (2024) may be a consequence, not of payoff confusion (as they argue elsewhere in their critique), but of noisy or inattentive subjects making capricious decisions in settings with boundaries that are asymmetrically distant from expected value – a possibility actually emphasized by Oprea (2024) as one of his own key hypotheses. Because valuation tasks typically elicit values on the support of the lottery, a range of imperfectly rational behaviors including noisy valuations, anchoring-and-adjustment heuristics, compromise heuristics and pull-to-the-center heuristics will all tend to produce prospect-theoretic patterns of behavior simply because of the nature of valuation. BSWW offer this possibility as an alternative to the Oprea (2024)'s account of his data, but in fact these are examples of *exactly* the types of cognitive shortcuts Oprea (2024) was designed to study. Indeed, Oprea (2024) lists exactly these kinds of behaviors as candidates for the shortcuts driving prospect-theoretic outcomes (e.g., Blavatskyy, 2007; Enke and Graeber, 2023) in the valuation of mirrors, and by Finding 2, to a significant extent lotteries.[30] BSWW call this kind of behavior "measurement error," but it is

[30] When offering explanations for his results, Oprea (2024) writes: "The "noisy coding" literature (Woodford, 2020; Glimcher, 2022), for instance, shows that if decision makers shade noisy evaluations of lotteries towards prior beliefs in a Bayesian manner, this can produce the fourfold pattern (Steiner and Stewart, 2016; Khaw et al., 2022; Vieider, 2024; Frydman and Jin, 2023) and even loss aversion (Woodford, 2012) under some assumptions. Enke and Graeber (2023) show, similarly, that uncertainty about the quality of value calculations, combined with cognitive defaults, can produce the fourfold pattern. Blavatskyy (2007) shows that if decision makers do nothing more in response to valuation noise than ensure that their valuations do not exceed the bounds of the lottery's sup-

in fact exactly the type of behavior that Oprea (2024) is attempting to measure with his mirrors. Given this, I agree with BSWW to a significant extent on the potential proximal mechanisms that may be driving the behaviors in Oprea (2024) – a fact which might point to many of our disagreements being semantic.

Third, in Section 2.5, BSWW again abandon the idea that mirror behavior is driven by payoff confusion, and adopt an interpretation very similar to Oprea (2024)'s by proposing to use mirrors as a device for recovering the "true" prospect theory preferences (i.e., tastes for risk) that remain in lotteries after mirror mistakes are netted out. To do this, they calculate normalized prospect-theoretic errors in mirrors at the subject level, and subtract them from the same deviations made by the same subjects in lotteries. The idea behind this decomposition seems to be that (i) some prospect-theoretic behavior is generated in common as mistakes in both lotteries and mirrors, (ii) differencing mirror mistakes from lottery choices nets these mistakes out, (iii) leaving only tastes for risk remaining in the difference ("The motivation ... is that there is evidence that lottery valuations are genuine expressions of risk preferences if mirror and lottery behavior are different," p. 25). This analysis seems to directly affirm Oprea (2024)'s key idea that a significant fraction of prospect-theoretic behavior occurs similarly as mistakes in both lotteries and mirrors, and that the latter serves as a measure of some of those common mistakes. Consistent with Oprea (2024)'s interpretation, BSWW find that the evidence of preferences that survives this decomposition is very small compared to the levels of prospect-theoretic behavior in the uncorrected lottery data, suggesting that (at least according to the decomposition they propose) most of the prospect-theoretic behavior originally observed in lotteries are not driven by standard, rationally expressed tastes for risk.[31] Although I have

port, the fourfold pattern will emerge as a result. Closely related is the literature on "decision by sampling," which roots the classical pattern in heuristics built on imprecise comparisons between past and present circumstances (Friedman, 1989; Stewart et al., 2006)." (p.3805) Indeed, he expresses particular support for this class of explanation: "As we highlight in Section IIC (and online Appendix A.5), the severity of the classical pattern in our data is strongly correlated with measures of behavioral noise (choice inconsistency) and cognitive uncertainty (expressed uncertainty about the optimality of one's own actions), which seems especially suggestive of explanations rooted in the use of imprecise valuation strategies..." (p. 3806).

[31] In Appendix Figure 13, I plot mean prospect theoretic errors in lotteries and mirrors and include also the mean *difference*, following BSWW. The Figure shows how these quantities vary with a number of cognitive measures included in the Oprea (2024) dataset. As the Figure shows, the magnitude of the difference (the green line) is generally very small (even for the most rational

some reservations about this decomposition (because, as Oprea (2024) emphasizes, there are reasons to think the levels of complexity in lotteries and mirrors may differ), it is so closely aligned to the complexity based interpretation of prospect theoretic behaviors in mirrors and lotteries offered in Oprea (2024) (and so inconsistent with the alternative interpretation of payoff confusion offered elsewhere in BSWW)[32] that it makes me wonder, again, whether BSWW's disagreements with Oprea (2024)'s interpretation are often semantic.

Finally, and most importantly, elsewhere in their critique, BSWW raise a concern that Oprea (2024)'s mirror results may be driven by the fact that subjects are *confused.* "Confusion" is a broad word that can refer to many kinds of errors, some of which I think are perfectly consistent with Oprea (2024)'s complexity-based interpretation.[33] However, I believe that BSWW are concerned especially with a specific sort of confusion that I have called "payoff confusion" – the concern that subjects mistakenly believe that they are in lotteries when they are, in fact, in mirrors. I fundamentally agree with BSWW that this type of confusion is a major threat to designs like Oprea (2024)'s. If subjects in mirrors believe they are in lotteries (i.e., are payoff confused), Oprea (2024)'s mirror results may not be due to complexity, but instead to subjects mistakenly expressing tastes for risk when valuing mirrors.

Indeed, I share this concern so strongly that much of the Oprea (2024) design was engineered to attempt to minimize scope for exactly this kind of confusion. For instance, the frequentist "box" description of lotteries and mirrors used in the design was introduced so that subjects assigned mirrors before lotteries would have minimal stochastic framing around the task. The goal with this framing was for mirrors to seem to subjects like a sort of deterministic puzzle rather than a lottery that pays

subjects) compared to the raw magnitude of prospect theoretic errors in lotteries (the blue line), suggesting that (at least according to BSWW's proposed decomposition) most prospect-theoretic behavior represents errors, not rationally expressed risk preferences.

[32] If subjects mistakenly believed mirrors were lotteries, leading them to express true risk preferences (i.e., tastes for risk) in mirrors (a'la the payoff confusion interpretation), differencing lottery and mirror behavior would not net out anything other than repeated expressions of risk preferences in the two types of tasks, meaning it would not decompose anything at all.

[33] For instance if subjects understand the payoff rule, but choose not to apply it to infer the object's value, they are in a certain sense "confused" about the true value of the object (i.e., they are left uncertain about what the object is worth). However if this kind of confusion leads subjects to use some cognitive shortcut or rule-of-thumb to value the object instead, it is perfectly consistent with the complexity interpretation offered in Oprea (2024).

its expected value, in order to minimize scope for payoff confusion especially among subjects who were randomly assigned mirrors before lotteries. The training questions, too, were included especially to target payoff confusion in subjects who were assigned mirrors after lotteries in order to make the change of payoff function especially salient. Finally, the payoff rule describing mirrors was designed to be relatively simple, clearly non-stochastic, and is shown in a highlighted color to subjects at the top of every mirror task so that subjects were constantly reminded of the payoff rule in every task – an effort to make it difficult for a subject to actually mistake a mirror task for a lottery.

Were these steps sufficient to prevent payoff confusion in mirrors in Oprea (2024)? The evidence outlined above suggests so (e.g., error rates drop sharply over the course of the training questions suggesting that subjects learned away most initial confusion, and the central claims in Oprea (2024) survive removing subjects who made training errors, pointing away from confusion driving the results). I am further persuaded that Oprea (2024)'s results reflect complexity rather than payoff confusion especially by several lines of secondary evidence in the data. One is that (as discussed in Online Appendix A.5 of Oprea 2024), a number of indices of low cognitive effort and ability (measures of noisiness, inattention, imprecision, unreflectiveness and self-doubt) very similarly predict prospect-theoretic valuations in both lotteries and mirrors (Figure 13 in this paper's Appendix visualizes these strong and similar predictive relationships). This strong and similar correlation of cognitive measures with lottery and mirror valuations is very consistent with Oprea (2024)'s complexity interpretation, but seems starkly inconsistent with the idea (implicit in the idea of payoff confusion) that prospect theoretic behavior reflects true preferences (i.e., tastes for risk), rationally expressed in lotteries but irrationally mis-expressed in mirrors.[34] I'm also persuaded by the fact that, as Oprea (2024) emphasizes, the key results in Oprea (2024) arise in mirrors even for subjects who have not yet encountered lotteries, which seems hard to square with concerns that the result is an artifact of subjects believing mirrors are lotteries.[35] Also persuasive to me is the fact that lottery and mirror valuations *both*

[34] In particular, if lottery valuations were rational expressions of true risk tastes, we would not expect them to be similarly correlated with markers of cognitive effort and ability as mirrors.

[35] Although the mirror task looks something like a lottery to psychologists and economists familiar with typical experimental paradigms, there are reasons to doubt that this is true to an average

dramatically change (and change similarly) when we change the method of elicitation (i.e., from MPL to BDM) – an instability that seems starkly inconsistent with the idea that prospect theoretic behavior entirely reflects stable tastes for risk, deployed in both lotteries and mirrors, but highly consistent with the idea that this behavior largely reflects the use of similar, shallow cognitive shortcuts that over-respond to superficial details of the elicitation in both lotteries and mirrors, a'la the alternative complexity interpretation.

Nonetheless, in the absence of further evidence, one's interpretation of the Oprea (2024) data will unavoidably be sensitive to one's priors about the plausibility of the hypothesis that prospect theoretic behavior may be a response to complexity rather than tastes for risk. My own priors on this are shaped by a long literature (briefly reviewed in Oprea 2024) stretching back to the 1960s that shows that anomalous lottery valuations are (i) highly unstable to context (e.g., Lichtenstein and Slovic, 1971; Hertwig et al., 2004), (ii) generate anomalous behaviors (like probability weighting) that diminish or disappear with learning (e.g., Van de Kuilen, 2009), (iii) are predicted by cognitive ability (e.g., Choi et al., 2021), (iv) are responsive to overt complexity manipulations (e.g., Huck and Weizsacker, 1999; Bernheim and Sprenger, 2020; Puri, 2023), (v) are influenced by manipulations to attention (e.g., Pachur et al., 2018), (vi) are solved procedurally in ways inconsistent with preference maximization (e.g., Arieli et al., 2011; Payne et al., 1988), etc. This literature makes a strong ex ante case that lottery valuations are heavily shaped by the use of superficial, non-maximizing valuation strategies and therefore that valuing lotteries is a complex task for many decision makers.

Ultimately, however, these questions and ambiguities can only be fully resolved by further research. While BSWW's critique has not convinced me that the interpretation offered in Oprea (2024) is mistaken, I am eager to see new experiments that deepen, alter, or even overturn this interpretation. First, concerns that the Oprea (2024)'s results are a consequence of the design being too confusing to yield insight

subject. To an average subject first assigned the mirror task (without having been previously assigned the lottery task), it seems likely that the task simply looks simply like a mental box-opening exercise – a reasoning puzzle with no natural referents to risk. To the degree this is true, we have little ex ante reason to expect subjects to be payoff confused in these tasks, particularly when they have not been exposed to lotteries first.

can only really be resolved one way or another by followup experiments that vary his procedures, instructions and other design choices in such a way as to satisfy us that the Oprea (2024) results are (or are not) overfit to that design. For this it is especially important to bear in mind that, as Oprea (2024) emphasizes, it should not be difficult to make mirrors easier or harder than Oprea (2024)'s – what is important is not to test the robustness of the rate of mistakes in mirrors (which we should expect to be sensitive to details of the implementation), but rather whether the mistakes that do occur continue to be systematically prospect-theoretic (i.e., whether Finding 1 is robust to variation in design), and continue to predict lottery behavior (Finding 2). Second, future experiments prospectively designed to more deeply understand how subjects interpret mirrors (including whether they sometimes suffer from payoff confusion) will be similarly valuable for shaping our interpretation of Oprea (2024)'s findings. When assessing payoff confusion specifically, following from my discussion in point 2 above, it is important to design tests and procedures that manage to target payoff confusion (a confound to Oprea 2024) but not low cognitive effort (an important input to Oprea 2024). We should expect the two to be difficult to separate and easy to confuse, but clever tests that manage this will be quite valuable for resolving the questions raised by BSWW.

References

Arieli, A., Y. Ben-Ami, and A. Rubinstein (2011): "Tracking Decision Makers under Uncertainty," *American Economic Journal: Microeconomics*, 3, 68–76.

Banki, D., U. Simonsohn, R. Walatka, and G. Wu (2025): "Decisions Under Risk are Decisions Under Complexity: Comment," .

Bernheim, B. D. and C. Sprenger (2020): "On the empirical validity of cumulative prospect theory: experimental evidence of rank-independent probability weighting," *Econometrica*, 88, 1363–1409.

Blavatskyy, P. (2007): "Stochastic expected utility theory," *Journal of Risk and Uncertainty*, 34, 259–286.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012): "Salience theory of choice under risk," *Quarterly Journal of Economics*, 127, 1243–1285.

Camerer, C. (2005): "Three cheers – psychological, theoretical, empirical – for loss aversion," *Journal of Marketing Research*, 42, 129–133.

Choi, S., J. Kim, E. Lee, and J. Lee (2021): "Probability weighting and cognitive ability," *Management Science*, forthcoming.

Enke, B. and T. Graeber (2023): "Cognitive Uncertainty," *The Quarterly Journal of Economics*, 138, 2021–2067.

Friedman, D. (1989): "The S-shaped value function as a constrained optimum," *American Economic Review*, 79, 1243–1248.

Frydman, C. and L. J. Jin (2023): "On the Source and Instability of Probability Weighting," *Unpublished Manuscript.*

Gabaix, X. (2014): "A Sparsity-Based Model of Bounded Rationality," *The Quarterly Journal of Economics*, 129, 1661–1710.

——— (2019): "Behavioral Inattention," *Handbook of Behavioral Economics: Applications and Foundations 2*, 2, 261–343.

Glimcher, P. (2022): "Efficiently irrational: illuminating the riddle of human choice," *Trends in Cognitive Science*, 26, 669–687.

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

Huck, S. and G. Weizsacker (1999): "Risk, complexity and deviations from extpected-value maximizagion: Results of a lottery choice experiment," *Journal of Economic Psychology*, 149, 1644–1683.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–291.

Khaw, M. W., Z. Li, and M. Woodford (2022): "Cognitive imprecision and stake-dependent risk attitudes," *Unpublished manuscript.*

Lichtenstein, S. and P. Slovic (1971): "Reversals of preference between bids and choices in gambling decisions." *Journal of experimental psychology*, 89, 46.

Martinez-Marquina, A., M. Niederle, and E. Vespa (2019): "Failures in contingent reasoning: the role of uncertainty," *American Economic Review*, 109, 3437–3474.

O'Donoghue, T. and J. Somerville (2018): "Modeling risk aversion in economics," *Journal of Economic Perspectives*, 32, 91–114.

Oprea, R. D. (2024): "Decisions Under Risk are Decisions Under Complexity," 114, 3789–3811.

Pachur, T., M. Schulte-Mecklenbeck, R. O. Murphy, and R. Hertwig (2018): "Prospect theory reflects selective allocation of attention," *Journal of Experimental Psychology: General*, 147, 147–169.

Payne, J. W., J. R. Bettman, and E. J. Johnson (1988): "Adaptive strategy selection in decision making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 534–552.

Puri, I. (2023): "Simplicity and Risk," *Unpublished Manuscript.*

Rubinstein, A. (1988): "Similarity and decision-making under risk (is there a utility theory resolution to the Allais paradox?)," *Journal of Economic Theory*, 46, 145–153.

Steiner, J. and C. Stewart (2016): "Perceiving prospects properly," *American Economic Review*, 106, 1601–1631.

Stewart, N., N. Chater, and G. D. Brown (2006): "Decision by sampling," *Cognitive Psychology*, 53, 1–26.

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

Van de Kuilen, G. (2009): "Subjective probability weighting and the discovered preference hypothesis," *Theory and Decision*, 67, 1–22.

Vieider, F. (2024): "Decisions under Uncertainty as Bayesian Inference on Choice Options," *Management Science*.

Woodford, M. (2012): "Prospect Theory as Efficient Perceptual Distortion," *American Economic Review Papers and Proceedings*, 102, 41–46.

——— (2020): "Modeling imprecision perception, valuation and choice," *Annual Review of Economics*, 12.

Appendix

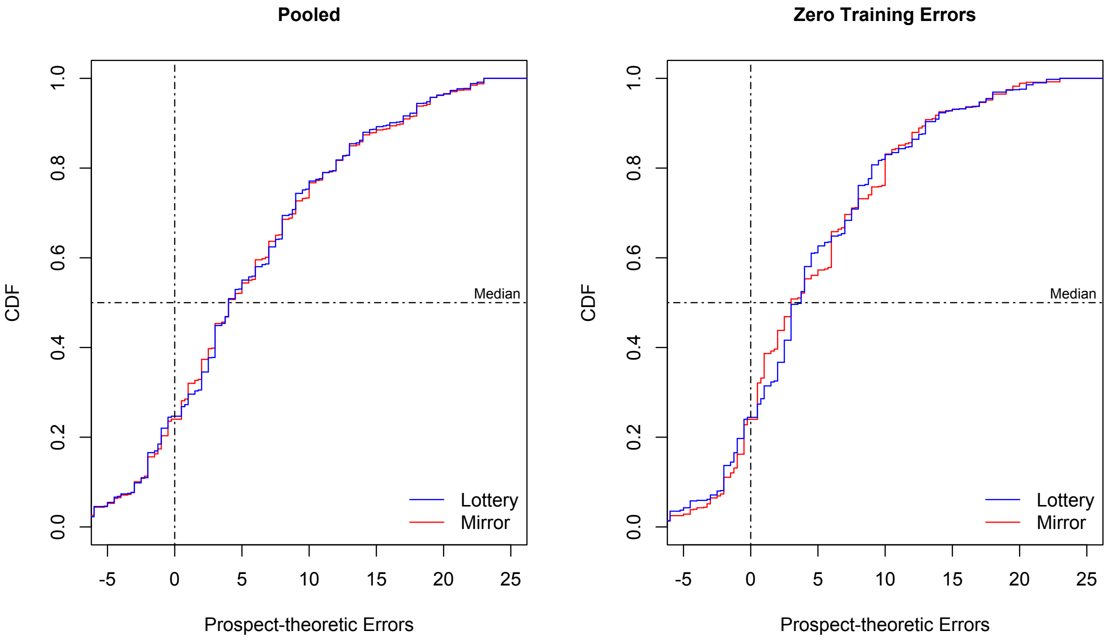


Figure 8: Empirical CDFs of deviations from expected value, normalized to be positive if consistent if in the direction of prospect-theoretic predictions with **expected valuation observations removed**. *Notes: Separate CDFs are provided for lottery and mirror deviations. The left panel plots the full pooled dataset and the right panel restricts to subjects who made zero training errors.*

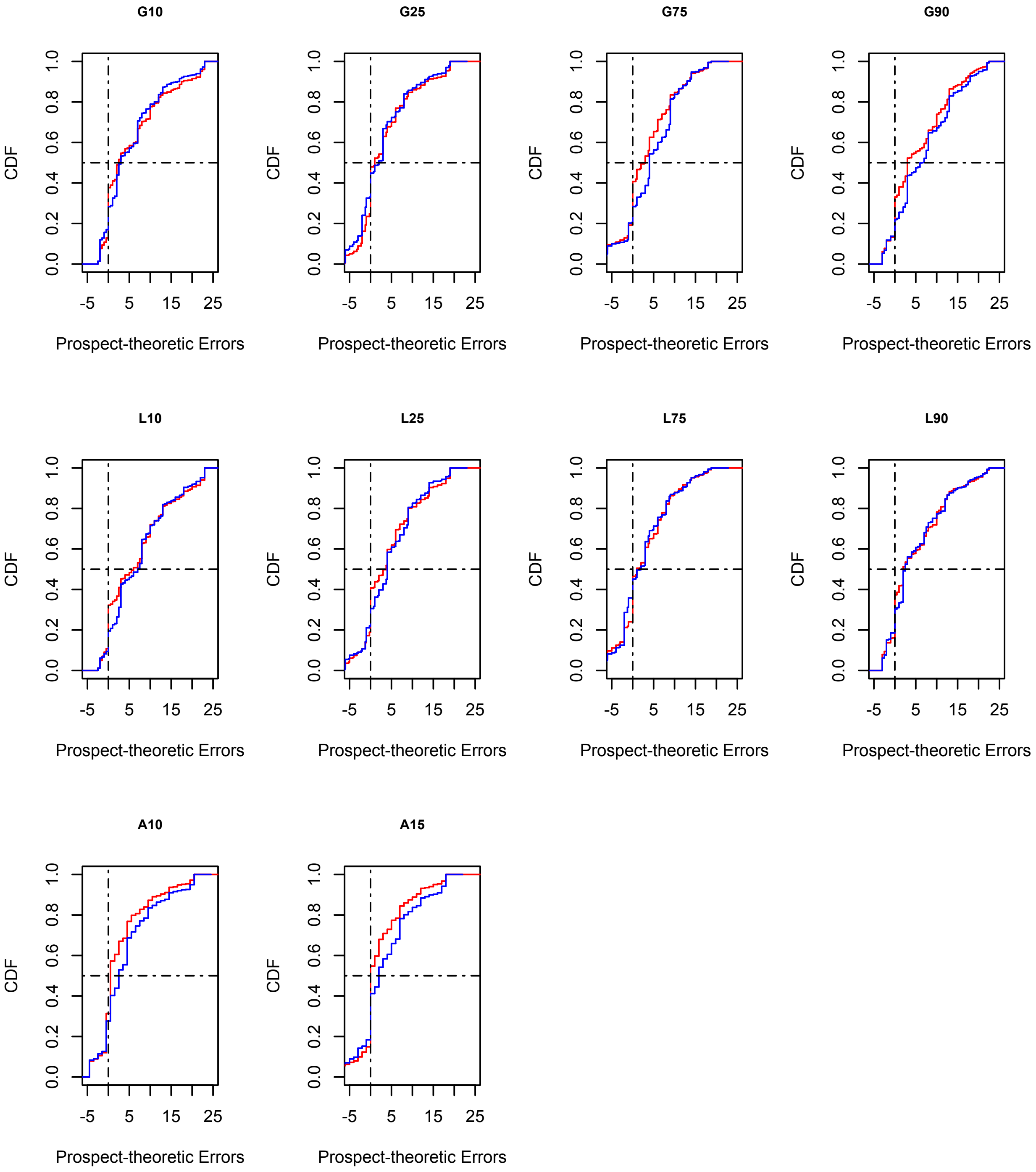


Figure 9: Empirical CDFs of deviations from expected value, normalized to be positive if in the direction of prospect-theoretic predictions, **pooled sample.**

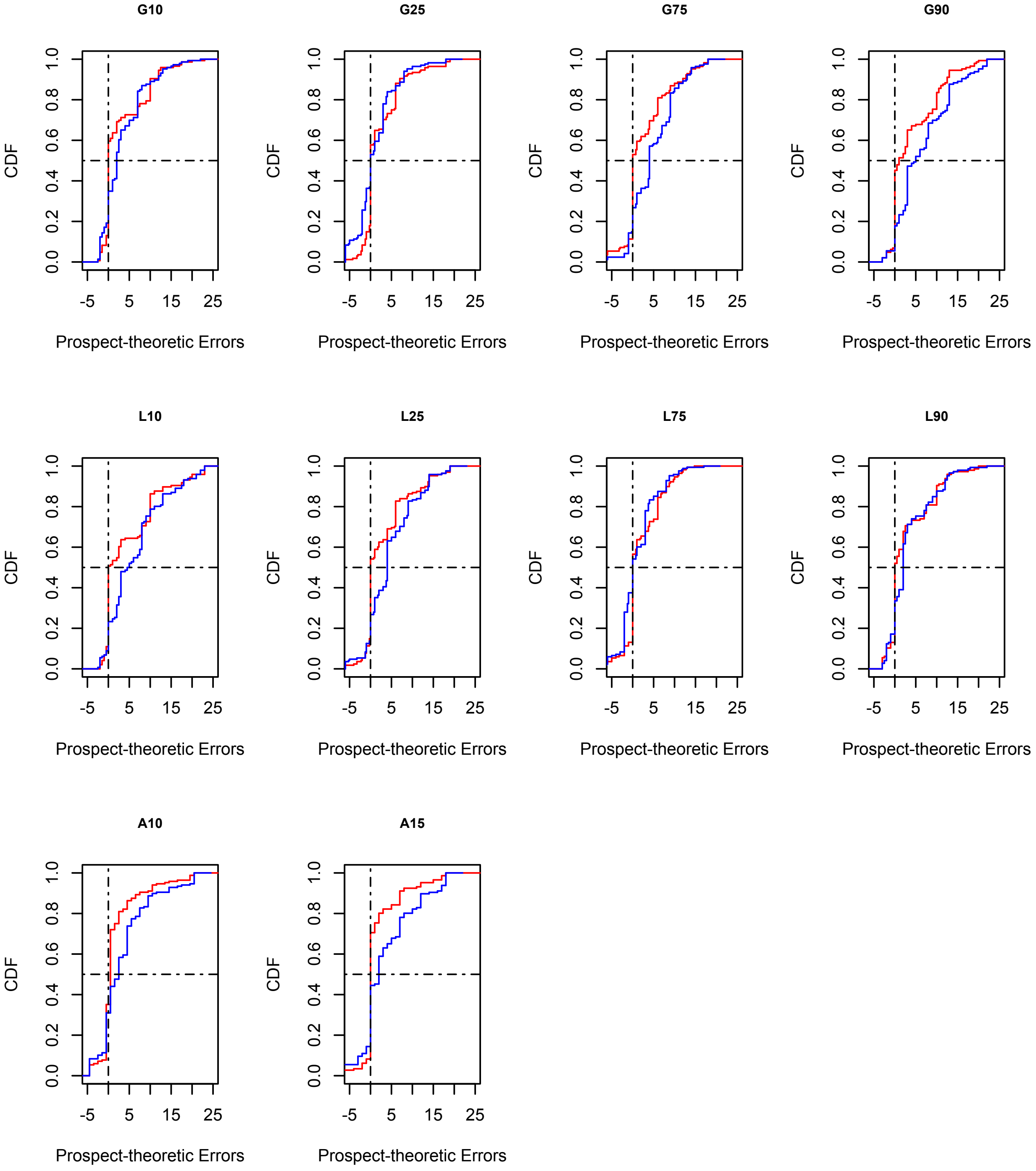


Figure 10: Empirical CDFs of deviations from expected value, normalized to be positive if in the direction of prospect-theoretic predictions, **including only subjects that made zero training errors.**

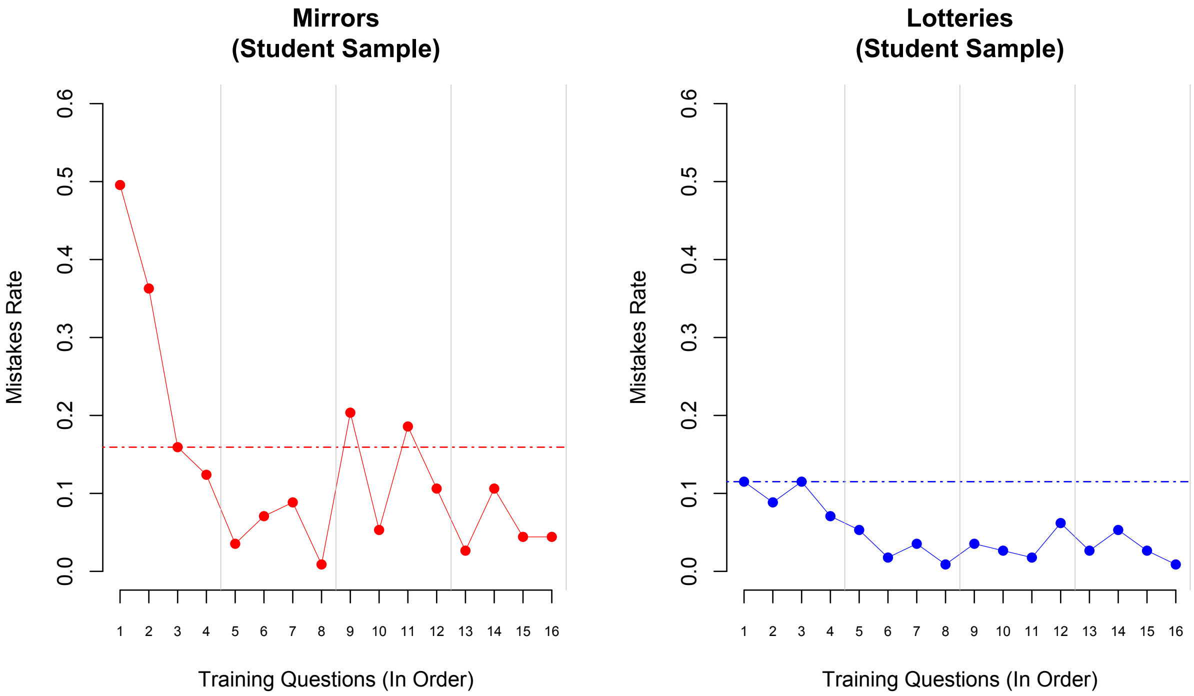


Figure 11: Time series of training question error rates over the 16 questions sequentially provided to subjects prior to mirror and lottery treatments for the student sample. *Notes: Horizontal lines mark the error rate for question #3 which measures low cognitive effort rather than payoff confusion. Vertical lines mark out blocks of questions in the training that reference the same example lottery/mirror.*

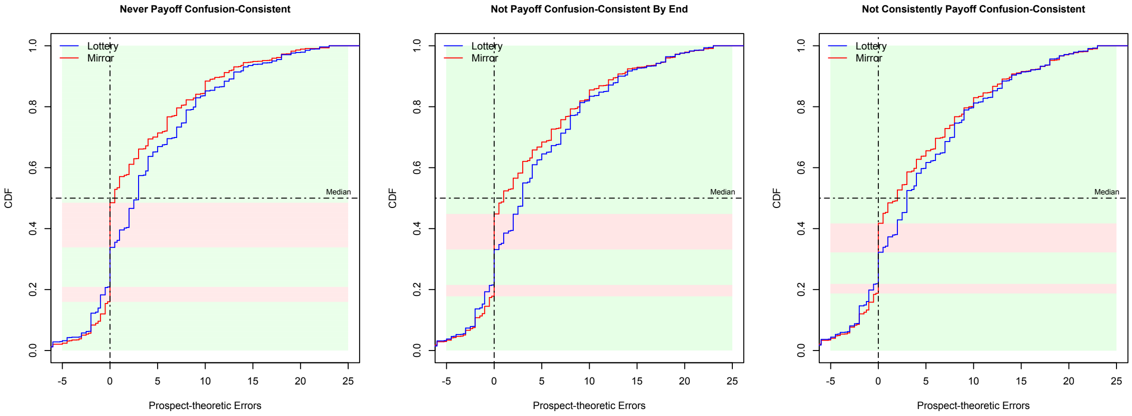


Figure 12: Empirical CDFs of deviations from expected value, normalized to be positive if in the direction of prospect-theoretic predictions. *Notes: From left to right, panels are included that exclude (i) subjects who made mistakes consistent with payoff confusion, (ii) subjects who made mistakes consistent with payoff confusion in the final training question (#4) and (iii) subjects who made mistakes consistent with payoff confusion in all training questions that measure payoff confusion. In each panel, separate CDFs are plotted for lottery and mirror deviations. Regions are shaded green where lottery and mirror distributions are qualitatively both prospect-theoretic (PT), expected value maximizing (EV) or anti-prospect theoretic (Anti-PT); areas are shaded red where lottery and mirror distributions are qualitatively different.*

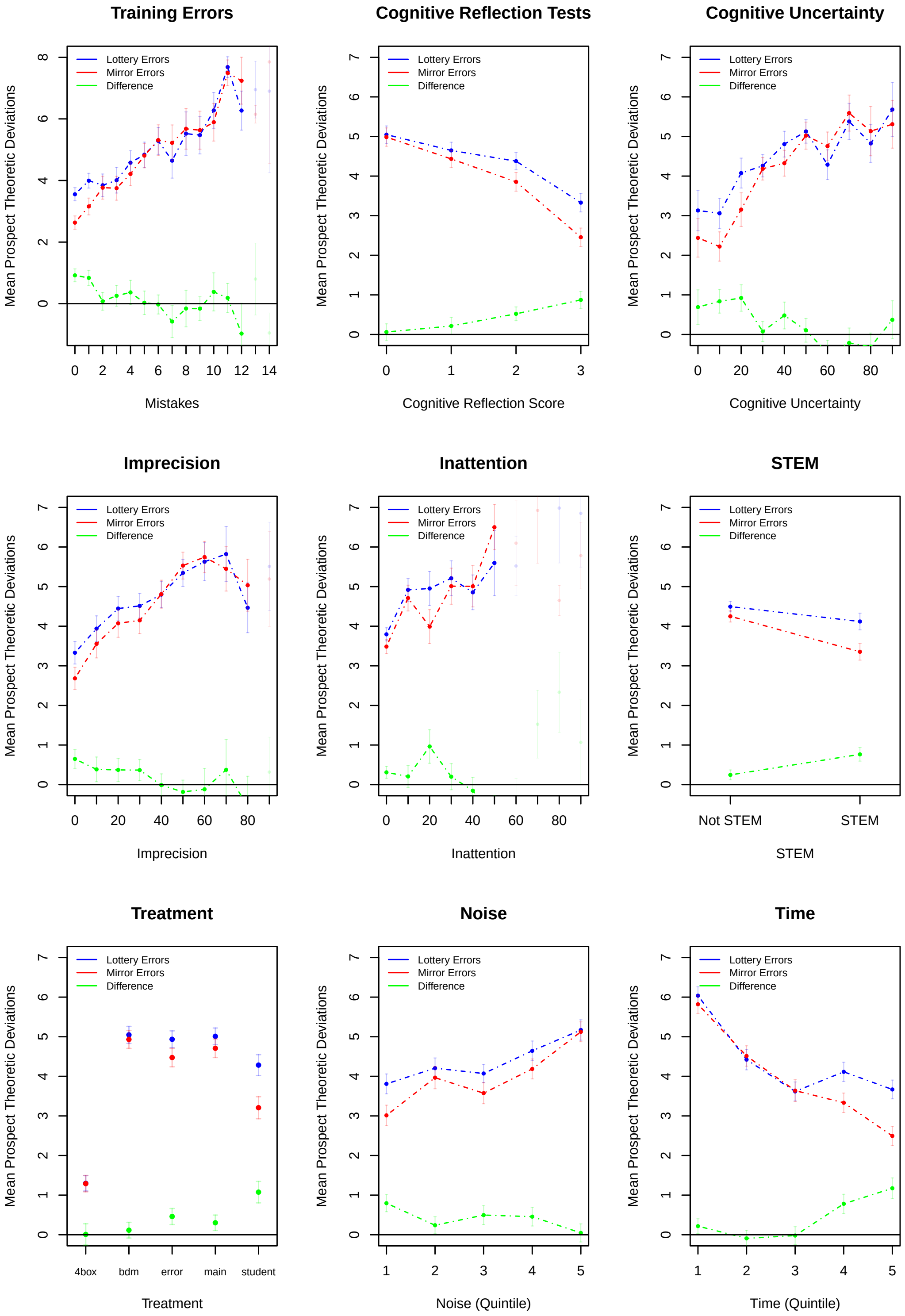


Figure 13: Mean deviations from expected value, normalized positive if in the direction of prospect-theoretic predictions. *Notes: Means of lotteries, mirrors and their difference are plotted in blue, red and green respectively. Means from bins with fewer than 10 data points are plotted separately and in a lighter shade.*