Recovering Preferences from Mistakes: An Auxiliary Task Approach*

Hanyao Zhang†

October 26, 2025

Abstract

Risk preferences recovered from lottery valuation data are not robust to unverifiable assumptions about the structure of mistakes in the valuations. To address this, we develop a novel approach utilizing Oprea's (2024) deterministic mirrors – deterministic payments that preserve key structural features of lotteries. We estimate the mistake structure in deterministic mirrors – where certain payments enable identification of mistake patterns – through a mixture model incorporating two types of mistakes frequently observed, then apply these estimates to correct lottery valuations. The corrected valuations are closer to risk neutrality than raw valuations; when they deviate from risk neutrality, they are predominantly risk averse. These patterns align the corrected valuations with expected utility theory, in contrast to the raw valuations which exhibit strong probability weighting. Our approach offers a generalizable framework for preference recovery: researchers can use auxiliary tasks with known correct answers to discipline assumptions about mistakes.

*We are beyond grateful of Mark Dean for his guidance and support. We benefited from comments by Alessandra Casella, Zhi Hao Lim, Tianhao Liu, Jiang Mao, Ryan Oprea, Mike Woodford, Yangfan Zhou, members of the Columbia Cognition and Decision Lab, and seminar audience at Columbia. We also thank Kirby Nielsen and George Wu for generously sharing their experimental instructions. This research was approved by Columbia IRB. All mistakes are my own.

†Department of Economics, Columbia University, New York. Email: hanyao.zhang@columbia.edu.

If one makes the wrong assumptions about the stochastic structure of the noise, then one usually makes wrong inferences from the data. (Hey, 2005)

1 Introduction

Risk preferences, when interpreted as a welfare-relevant ordering of alternatives, are important in designing and evaluating economic policies. These preferences are typically recovered by applying a revealed preference argument to observed choices under risk. However, one caveat of applying the recovered preferences for welfare analysis is the possibility that some choices reflect *mistakes* – failure to make the choice that reflects their preferences – due to the complexity involved in decisions under risk (Martínez-Marquina, Niederle and Vespa, 2019, Oprea, 2024), or in the context of experimentally-collected data, the complexity involved in correctly responding to experimental incentives (Plott and Zeiler, 2005, Cason and Plott, 2014, Danz, Vesterlund and Wilson, 2022). If choices fail to reflect preferences, the revealed preference argument fails to generate the true welfare-relevant ordering, which poses a threat to subsequent welfare and policy analysis based on the recovered preferences.

A by-product of the mistakes is the stochasticity of choices, i.e., making different choices facing the same decision. If mistakes are stochastic within-individual (Khaw, Li and Woodford, 2021), or heterogeneous between-individuals in pooled data sets, choices would become stochastic.[1] To recover preferences from stochastic mistakes, the dominant approach is to first make parametric assumptions over the *stochastic structure* of these choices, i.e., the distribution of choices conditional on the preference. The stochastic structure specifies the possible mistakes that individuals make, along with the probability of each specific mistake. This approach then proceeds by jointly estimating the preferences and the parameters of the stochastic structure. A problem of this approach is that the assumptions over the stochastic structure have been shown to greatly affect the recovered preferences (e.g., Buschena and Zilberman, 2000, Carbone and Hey, 2000, Alós-Ferrer, Fehr and Netzer, 2021, O'Donoghue and Somerville, 2024). Making things worse, these assumptions are typically unverifiable, since the stochastic structure can be precisely measured only when the preferences are known, which is typically not the case if one's very goal is to learn these preferences (Alós-Ferrer, Fehr and Netzer, 2021). Since the recovered preferences are not robust to unverifiable assumptions

[1]We acknowledge that stochastic choices are most commonly interpreted as fluctuations of preferences within-individual or unobserved heterogeneity of individual preferences in a pooled data set (e.g., McFadden, 1974). Both interpretations above are preference-based, implying that the stochasticity is conceptually distinct from mistakes and instead, should enter into welfare analysis. As will be made clear later, in this paper, we view stochastic choices as resulting from both mistakes and preferences. Our approach aims to correct for the mistakes in the stochastic choices while leaving the stochastic preferences intact. As a result, our final measure of preference will not be a single preference that is representative of the entire population, but a distribution of preferences in the population.

over the stochastic structure of choices, their interpretability as a welfare-relevant ordering and applicability in welfare analysis and policy design are severely limited.

In this paper, we develop an approach that reduces reliance on unverifiable assumptions about the stochastic structure by measuring it in a related context where it is observable. We achieve this by eliciting subjects' valuations of an object with a known valuation that is closely related to lotteries – namely, the deterministic mirrors of the lotteries (Oprea, 2024). The deterministic mirrors have two key advantages. First, the mirrors have *objectively correct* valuations, allowing us to directly observe the mistakes and estimate the stochastic structure. Second, as will be discussed in more details below, the mistakes in mirror valuation tasks plausibly serve as a *lower bound* of those in lottery valuation tasks, in the sense that any mistakes present in the mirror tasks should also manifest in the lottery tasks. Given this lower bound interpretation, we may use the mistakes measured in the mirror tasks to partially correct for the mistakes in the lottery tasks.

In our experiment, each subject values a set of binary lotteries $(\$X, p\%; \$0)$ (with probability $p\%$ winning a prize of $\$X$, and \$0 otherwise). In the experiment, each subject faces a fixed $X \in \{25, 100\}$ and five different $p \in \{10, 25, 50, 75, 90\}$. The valuations of lotteries are elicited with price lists (Holt and Laury, 2002) with enforced single-switching rule, yielding a unique *response* – the switching point where subjects shift from preferring the lottery to preferring a sure payment.

In addition to the lotteries, each subject also values the *deterministic mirror* of each lottery that they value. For a lottery $(\$X, p\%; \$0)$, both itself and its mirror are presented to subjects as 100 boxes, p of which containing $\$X$ and others \$0. The key difference is in the payment rule. For lotteries, one box is randomly drawn to determine the payout, creating uncertainty. For mirrors, subjects are told the payout equals the average across all 100 boxes with certainty. This means the mirror pays out the expected value of its corresponding lottery. Subjects make valuation decisions using the same price list mechanism for both the lottery and the mirror.

Crucially, mirrors are designed to hold constant the information processing required to understand the disaggregated presentation of outcomes. Both lotteries and mirrors require subjects to process the same visual information and to navigate the same price list elicitation mechanism. Mistakes arising from the complexity of processing this disaggregated format or from misunderstanding how price lists work should therefore manifest similarly in both contexts. The presence of risk in lotteries may introduce additional mistakes specific to decision-making under uncertainty (Martínez-Marquina, Niederle and Vespa, 2019), but it should not eliminate the mistakes already present in the deterministic mirror tasks. This argument supports using mirror-estimated stochastic structure to correct lottery responses: we can identify and measure the mistakes common to both contexts through mirrors and

correct for them in lottery valuations, while recognizing that lotteries may contain additional risk-specific mistakes.

Before describing our approach, we need to clarify a few key terminologies. We use *genuine valuation* to refer to the response that ranks the highest according to a subject's preference ordering facing a price list. When the response differs from the genuine valuation, we refer to this discrepancy as a *mistake*. In mirror tasks, genuine valuations are known, which, in turn, makes the mistakes observable. In lottery tasks, however, genuine valuations are unknown and heterogeneous across individuals, and the mistakes are not directly observable. The *stochastic structure* captures the relationship between genuine valuations and the distribution of responses – formally, it is the distribution of responses conditional on genuine valuation.

Our approach proceeds in two steps to recover a measure of risk preferences referred to as risk-induced valuations (RIV). All analysis in this paper works with aggregate distributions pooled across subjects. In Step 1, we estimate the stochastic structure in mirror tasks. In Step 2, we use this estimated structure to correct the lottery responses and construct the RIV. The RIV represents the distribution of responses that are attributable to risk – capturing both genuine valuations and any risk-specific mistakes absent in mirrors. Importantly, RIV is a distributional measure rather than an individual-level preference: we correct for mistakes in the aggregate distributions of responses while recognizing that the responses contain inherent stochasticity from preference-based sources, including within-individual fluctuations and between-individual heterogeneity in preferences. We do not attempt to recover a single preference or utility model that applies to the whole population.

Step 1 estimates the stochastic structure in mirror tasks. Because we ultimately need to apply this structure to lottery valuations in Step 2, we need more than just the distributions of responses for the in-sample mirrors that we elicit valuation for. Instead, we need to fully characterize the relationship between any out-of-sample mirror's genuine valuation and its distribution of responses. This allows us to predict the distribution of responses for genuine valuations we do not directly observe. This is necessary because we only elicit valuations for a limited set of mirrors that do not span the full range of genuine valuations that arise when subjects evaluate lotteries. We refer to this as the *domain expansion* problem.

We approach the domain expansion problem by observing that the vast majority of mistakes made by subjects when valuing mirrors belong to either of the two types: *difference-dependent mistakes* or *unconditional mistakes*. Difference-dependent mistakes refer to the mistake where subjects submit valuations that are close to, but not equal to, the genuine valuation, with their probabilities depending only on the difference between the response and the genuine valuation. On the other hand, unconditional mistakes refer to the mistake where subjects submit responses that follow a distribution that is invariant with the mirror. With this observation, we then build a *mixture model* assuming that the stochastic structure in

mirror tasks is a weighted average of three component distributions: the genuine valuation, difference-dependent mistakes, and unconditional mistakes. Estimating the three component distributions and their respective weights in this mixture model allows us to expand the stochastic structure from in-sample mirrors to out-of-sample mirrors, and thus solving the domain expansion problem. Moreover, the estimated mixture model provides a characterization of the stochastic structure in mirror tasks as a by-product. The unconditional mistakes mainly take the form of submitting the highest or lowest feasible responses, as well as the responses right in the middle of the price list. On the other hand, the difference dependent mistakes exhibit systematic downward bias for prize $X = 25$ and is symmetrically distributed around the genuine valuation for prize $X = 100$.

Having estimated the stochastic structure in mirror tasks, Step 2 applies the estimated stochastic structure from mirror tasks to correct lottery responses and construct RIV. For the RIV to be a good measure of risk preferences, the key condition is that the stochastic structure in lottery tasks is similar to that in mirror tasks. This condition is inherently unverifiable due to the unobservability of risk preferences, but we offer several arguments in support of our approach. First, any method for recovering risk preferences from observed choices requires assumptions about stochastic structure. The standard approach assumes parametric forms such as uniform mistake, normal mistake, or random utility models with parametric assumptions over the noises. However, we show that these standard assumptions miss important features of the stochastic structure in mirror tasks, which casts doubt on whether they would accurately capture the stochastic structure in lottery tasks either. In contrast, our approach grounds the assumption in mirror data, where we can verify that the estimated structure accurately describes observed behavior in at least one closely related decision environment. Second, both empirical evidence and theoretical considerations suggest mistakes transfer across decision contexts. Our own data reveal parallels between mirror and lottery responses. In particular, both exhibit stable masses at extreme and mid-range responses regardless of the lottery or mirror characteristics, which is a pattern commonly attributed to mistakes. Moreover, theories explaining common behavioral anomalies under risk – such as cognitive imprecision (Khaw, Li and Woodford, 2021) and salience (Bordalo, Gennaioli and Shleifer, 2012) – are based on psychological mechanisms that are not specific to risky choices. These theoretical foundations suggest that similar mistake patterns should appear in both risky and risk-free settings.

However, the condition that stochastic structures are similar across contexts may not hold perfectly. Martínez-Marquina, Niederle and Vespa (2019) find that while most mistakes in risky settings already appear in risk-free settings, people make more mistakes when facing risky choices – suggesting that mirrors capture a lower bound of mistakes in lotteries. This has implications for interpreting our results. By correcting for the mistakes that are already

present in the risk-free mirror tasks, the RIV we recover represents the distribution of valuations in the population that are attributable to risk, which captures both genuine risk preferences and any additional risk-specific mistakes. While we cannot further decompose RIV into its preference and risk-specific mistake components without additional assumptions, correcting for the mistakes we directly observe in mirrors – which plausibly persist in lottery tasks – is more disciplined than imposing parametric assumptions about the stochastic structure with limited empirical foundations.

Our findings reveal several important patterns in the recovered RIV. Compared with raw responses, the RIV distributions are substantially more concentrated around the expected values of the lotteries. When the RIV deviates from expected values, it predominantly exhibits risk-aversion, with risk-lovingness being less common. Even for lotteries involving small probabilities of gains ($p = 0.1$), the RIV shows roughly equal proportions of risk-averse and risk-loving valuations. This contrasts sharply with both the raw responses in our data and the classic fourfold pattern documented by Tversky and Kahneman (1992), where the majority of responses exhibit risk-lovingness for low-probability gains. Finally, the RIV are less heterogeneous than the raw responses, suggesting that heterogeneity in responses (Bruhin, Fehr-Duda and Epper, 2010) partly reflects heterogeneity in mistakes rather than solely heterogeneity in genuine risk preferences.

These patterns have several implications. First, since the expected utility theory (EUT) predicts approximate risk-neutrality under small stakes (Rabin, 2000), the RIV are more aligned with the EUT predictions than raw responses. The RIV also exhibit near risk-neutrality under the smaller prize of \$25 and weak risk-aversion under the larger prize of \$100, consistent with EUT under concave utility functions. Together, these patterns suggest that previous literature has overestimated the degree of deviation of risk preferences from EUT. Second, our evidence suggests that past estimates of prospect theoretic parameters – particularly the strong overweighting of small probabilities – should not be interpreted as solely reflecting genuine preferences, but at least partly reflect mistakes in responding to preference elicitation tasks.

To assess the robustness and generalizability of our findings, we apply the same two-step procedure to two additional data sets: Oprea (2024) and Zhang (2025). Both data sets collect valuations of lotteries and mirrors from the same subjects, allowing seamless applications of our approach. The key findings in this paper are robust to using alternative data sets. First, the mixture model with difference-dependent and unconditional mistakes provides an excellent fit to mirror responses in both datasets, demonstrating that these two mistake types are robust features of behavior in mirror valuation tasks across different experimental implementations. Second, the RIV recovered from both datasets exhibit patterns similar to our main results: compared to raw responses, the RIV are more concentrated around

expected values.

Moreover, we conduct a validation test using Zhang's (2025) data, which includes the same subjects valuing the same set of lotteries and mirrors twice under two different treatments – one with and one without a calculator to aid calculations. The raw responses differ significantly across treatments, as the calculator reduces calculation costs and thus reduces certain types of mistakes. However, the recovered RIV are statistically indistinguishable across treatments. This is precisely what we would expect if the RIV successfully isolate genuine risk preferences: since the same subjects face the same lotteries under both treatments, their genuine preferences should remain unchanged, and only the stochastic structure (mistakes) should differ. The fact that our procedure recovers similar RIV across treatments – despite different raw responses – provides support that the RIV capture stable preferences while successfully filtering out treatment-specific mistakes.

Topically, this study contributes to the literature studying the stochasticity of choices under risk (Hey and Orme, 1994, Hey, 1995, 2005, Loomes and Sugden, 1998, Buschena and Zilberman, 2000, Carbone and Hey, 2000, Khaw, Li and Woodford, 2021). Relatedly, there is also a literature studying the roles of non-preference factors in choices under risk (Woodford, 2012, Martínez-Marquina, Niederle and Vespa, 2019, Frydman and Jin, 2021, Nielsen and Rehbeck, 2022, Enke and Graeber, 2023, Enke and Shubatt, 2023, Oprea, 2024, McGranaghan et al., 2024, Zhang, 2025). Non-preference factors in choices under risk are also found in field settings, especially in the context of insurance choices (e.g., Barseghyan et al., 2013, Handel et al., 2024). This study contributes to these threads of literature by characterizing the stochastic structure of mistakes for the mirror tasks through the mixture model, and utilize the stochastic structure in mirror tasks to correct for the mistakes in lottery tasks, which yields a better measure of genuine risk preferences than the raw responses of lotteries.

Speaking of experimental methodology, this paper follows the leads of Cason and Plott (2014), Martínez-Marquina, Niederle and Vespa (2019), Enke and Shubatt (2023), and Oprea (2024). This literature develops the experimental technique of asking subjects to value auxiliary objects of *known* valuations (e.g., mirrors) that are closely related to the objects of interest (e.g., lotteries), which typically have *unknown* valuations. One focus of this literature has been identifying the structure of mistakes when valuing the auxiliary objects, and utilizing the identified stochastic structure to shed light on the structure of mistakes when valuing the objects of interest. This paper advances on this literature in two aspects. First, by adopting a mixture modeling approach, we provide a characterization of the stochastic structure when valuing the auxiliary objects. Second, this paper provides a formal way to correct for the identified mistakes when valuing the objects of interest.

In terms of statistical modeling, this paper is related to the literature in psychology

that uses mixture modeling to estimate and correct for unconditional mistakes[2] (Wichmann and Hill, 2001*a*,*b*, Treutwein and Strasburger, 1999, Prins, 2012, Clark and Merfeld, 2021). This literature typically applies mixture modeling to binary choice data. The current paper expands the mixture modeling to multinomial choices and even a continuum of possible options (see our application of the mixture modeling to the data set of Zhang, 2025, in Section 6). In economics, mixture modeling has been employed to study risk preference in Bruhin, Fehr-Duda and Epper (2010) and Conte, Hey and Moffatt (2011), and a related hierarchical Bayesian modeling approach is employed in Balcombe and Fraser (2025). While studies in this literature employ these modeling approaches mainly to uncover the heterogeneity of lottery valuations at within-individual or between-individual levels, we rely on mixture modeling to characterize and correct for the mistakes made by subjects.

The rest of this paper is organized as follows. Section 2 describes the experimental design. Section 3 outlines our conceptual framework. Section 4 describes our Step 1 of estimating the stochastic structure in mirror tasks. Section 5 illustrates our Step 2 of recovering risk-induced valuations. Section 6 applies the two-step procedure developed in this paper to the data sets of Oprea (2024) and Zhang (2025). Finally, Section 7 discusses the implications of our findings and concludes.

2 Experimental Design

We elicit subjects' valuations for a set of binary lotteries ($\$X, p\%; \$0$). Each subject faces five different p in the experiment: $p \in \{10, 25, 50, 75, 90\}$, and is randomized into one of the two prizes \$25 and \$100. The prize stays the same for a fixed subject throughout the experiment. In the experiment, each lottery is presented as 100 boxes, each containing some amount of money. The subjects are told that the realized payout of a lottery will be determined by the amount in a randomly drawn box.

In the same experimental session, subjects are also asked to value the deterministic mirrors of each of the five lotteries (Oprea, 2024). For a lottery, its deterministic mirror is presented visually as the same set of 100 boxes. However, the subjects are told that given a set of boxes, the mirror pays out the average amount of money contained in these 100 boxes with certainty. In other words, a mirror pays out the expected values of its corresponding lottery with certainty.

All valuations are elicited through price lists (Holt and Laury, 2002). An example price list is provided in Table 1. In each row of a price list, the subjects have to indicate whether they would like to choose a lottery/mirror, which is fixed in a price list, or a sure payment, which varies from the top to the bottom of the price list. The chosen option in each row is

[2]The terminology used in the psychology literature for what we refer to as unconditional mistakes is *lapses*.

marked with green shade, as we do in the actual experimental interface. Each price list has 26 rows with equal increments – under prize \$25, subjects choose between a lottery/mirror and 26 sure payments ranging from \$0 to \$25, with increments of \$1; under prize \$100, subjects choose between a lottery/mirror and 26 sure payments ranging from \$0 to \$100, with increments of \$4.

Decision	Option A	Option B
1	Being paid according to the boxes	**\$0.00** for sure
2	Being paid according to the boxes	**\$1.00** for sure
3	Being paid according to the boxes	**\$2.00** for sure
4	Being paid according to the boxes	**\$3.00** for sure
5	Being paid according to the boxes	**\$4.00** for sure
⋮	⋮	⋮
25	Being paid according to the boxes	**\$24.00** for sure
26	Being paid according to the boxes	**\$25.00** for sure

Table 1: An example price list

We enforce single switching in the price lists. That is to say, if a subject chooses a sure payment $\$x$ over the lottery/mirror, they have to also choose $\$x'$ over the lottery/mirror for all $x' > x$. Due to enforced single switching, for each price list, there is a highest sure payment $\$x^*$ that the subject chooses the lottery/mirror over $\$x^*$, but chooses $\$x'$ over the lottery/mirror for every $x' > x^*$.[3] We refer to this highest x^* as the *response*. For example, in Table 1, the response is \$2. The response differs slightly from the certainty equivalent, since the latter is the sure payment that makes a subject *indifferent* between the lottery and the sure payment. In our design with discrete payment increments, the certainty equivalent lies between the response and the next higher sure payment.

The experiment is split into two blocks – one containing all lottery tasks and the other containing all mirror tasks. The order of blocks is randomized at the subject level. The subjects are not informed about the second block while completing the first block, but receives instructions about the new block before starting the second block. Before each block, the subjects are first explained the payment rule (lottery or mirror) in the coming block. Then, the subjects are shown three examples. In each example, we present them with a set of 100 boxes, and generated 10 simulated payouts from the set of boxes according to the payment rule that is currently in effect. Finally, the subjects are asked to answer the five comprehension questions from Wu (2025) that tests their understanding of the payment rule

[3]Two special cases are: (1) the subject always chooses the sure payment; and (2) the subject always chooses the lottery/mirror. In the former case, the switching point is coded as \$0, while in the latter, the switching point is coded as the prize.

in effect. If a subject fails two or more comprehension questions in any of the two blocks, they will be screened out of the experiment. The subjects only start a block after passing the comprehension check. The instructions over the price lists follow McGranaghan et al. (2024). The complete experimental instructions are provided in the Appendix.

The experiment was conducted on Prolific in May 2025. Each subject received $3 payment upon completion. We incentivize the responses by the following mechanism: At the end of the experiment, with a probability $1/X$, where X is the subject-specific prize, a row in a price list is randomly selected, and the subject's choice in that row determines their payment. The median subject spent 22 minutes to complete, and the average total payment is $3.51. In total, 292 subjects finished the experiment, of which 149 have prize $X = 25$, and 143 have prize $X = 100$. In addition, 214 subjects failed the comprehension check in either of the blocks and were screened out.[4] Our data only includes the subjects who finished the experiment.

3 Conceptual Framework

This section outlines our two-step approach for recovering the distribution of genuine lottery valuations from observed responses. We apply this two-step procedure separately for each prize level $X \in \{25, 100\}$, allowing the stochastic structure to differ across prizes. We begin by introducing key terminology and notation, then describe the fundamental identification problem that necessitates our approach, and finally explain how we solve this problem using data from mirror tasks.

Consider a subject facing a lottery $l \in L$ who submits a response $r(l) \in R$ via a multiple price list, where R denotes the finite set of feasible responses (switching points). Each subject has a *risk preference*—an ordering over all feasible responses in R conditional on lottery l. We denote by $v(l) \in R$ the response that ranks the highest according to this preference ordering, which we call the *genuine valuation*. The genuine valuation represents what the subject would choose in the absence of any mistakes. Risk preferences may vary across the subject population, giving rise to a probability mass function $q^l(v)$ over genuine valuations for each lottery l. Throughout this paper, we work with aggregate distributions pooled across all subjects, analyzing behavior at the population level rather than the individual level. To simplify notation, we drop the dependence of r and v on l when context makes it clear.

Subjects may fail to submit responses that coincide with their genuine valuations due to factors such as inattention, misunderstanding of price lists, calculation errors, reliance on simplifying heuristics, or perceptual noise. When a response differs from the genuine

[4] If a subject passed the comprehension check of the first block but failed the second, we paid them $1.5 to compensate for their time.

valuation, we refer to this discrepancy as a *mistake*. Conditional on genuine valuation v, the response r has a probability mass function $g(r|v)$, which we call the *stochastic structure*. The stochastic structure captures the probabilistic relationship between what subjects genuinely prefer and what they actually choose.

By the law of total probability, the marginal distribution of responses for lottery l, denoted $f^l(r)$, satisfies

$$f^l(r) = \sum_v g(r|v)q^l(v). \tag{1}$$

One interested in recovering the distribution of genuine valuations $q^l(v)$ can collect lottery valuation data revealing the distribution of responses $f^l(r)$. However, Equation (1) reveals a fundamental *identification problem*: $g(r|v)$ and $q^l(v)$ are only jointly identified. Without assumptions about the stochastic structure $g(r|v)$, the distribution of genuine valuations $q^l(v)$ cannot be recovered from observed responses alone.

One might be tempted to assume parametric forms for the stochastic structure. For instance, the *Normal Mistake* model assumes

$$g(r|v) \propto \exp\left(-\frac{(r-v)^2}{\sigma^2}\right), \quad \text{for some } \sigma > 0, \tag{2}$$

while the *Uniform Mistake* model assumes

$$g(r|v) = (1-\epsilon)\mathbb{1}[r=v] + \epsilon\frac{1}{|R|}, \quad \text{for some } \epsilon \in [0,1]. \tag{3}$$

With such parametric assumptions, one can jointly estimate $q^l(v)$ and the parameters of $g(r|v)$ (e.g., σ or ϵ). This approach may provide adequate predictive fit. However, if the functional form assumptions are incorrect, the recovered $q^l(v)$ will be biased. For example, experimental evidence shows that subjects using multiple price lists tend to bias their responses toward the middle of the list (Andersen et al., 2006, Beauchamp et al., 2020). If one incorrectly assumes Normal or Uniform Mistake structures that ignore this bias, the recovered distribution $q^l(v)$ will inherit this bias, leading to erroneous conclusions about genuine risk preferences.

Deterministic mirrors provide a solution to this identification problem. For a mirror $m \in M$, the genuine valuation $v(m) \in R$ is *known* because the mirror pays a deterministic amount with certainty. By collecting subjects' valuations for both lotteries L and mirrors M, we can employ a two-step procedure to recover the distribution of genuine lottery valuations.

Step 1: Estimate the stochastic structure in mirror tasks. Since we know $v(m)$ for each mirror m, we can estimate the stochastic structure in mirror tasks, $g_M(r|v)$, from mirror valuation data. We need the estimated stochastic structure to apply not just to the

specific mirrors we elicit valuations for, but to the full range of genuine valuations that may arise when subjects evaluate lotteries. To achieve this, we exploit empirical regularities in the mistakes committed by subjects and build a statistical model that characterizes the complete stochastic structure $g_M(r|v)$ for all possible genuine valuations $v \in R$. We defer the details of this estimation approach to Section 4.

Step 2: Apply the estimated structure to recover genuine lottery valuations. Having estimated the stochastic structure from mirror data, denoted $\hat{g}_M(r|v)$, we use it to correct lottery responses. Specifically, we use the estimated stochastic structure in mirror tasks $\hat{g}_M(r|v)$ to approximate the theoretical stochastic structure $g(a|v)$ in lottery tasks. Making this approximation in Equation (1) yields

$$f^l(r) = \sum_v \hat{g}_M(r|v)\tilde{q}^l(v). \tag{4}$$

Since both $f^l(r)$ (from lottery data) and $g_M(r|v)$ (from Step 1) are known, we can recover a distribution $\tilde{q}^l(v)$, referred to as the distribution of risk-induced valuations (RIV). The RIV $\tilde{q}^l(v)$ represents our estimate of the true distribution of genuine valuations $q^l(v)$. If the approximation in Equation (4) holds exactly, the RIV successfully recovers $q^l(v)$. However, to the extent that lottery tasks involve additional mistakes beyond those present in mirror tasks, the RIV may differ from the true $q^l(v)$. The validity of this approximation and the interpretation of the RIV is provided in Section 5.4.

Before we get into more details, it is worthwhile for us to summarize our use of terminologies in Table 2.

Terminology	**Notation**	**Meaning in this paper**
Response	r	The raw valuations reported by subjects in the experiment.
Risk-induced valuation (RIV)	$\tilde{q}^l$	The distribution recovered through Equation (4). Not identified at the individual task-level, but identified as a distribution over R for each lottery l.
Genuine valuation	v, with distribution q^l	The response that would optimize the underlying welfare-relevant preference ordering. Not directly observed.

Table 2: Conceptual distinction between responses, risk-induced valuations, and genuine valuations

4 The Stochastic Structure in Mirror Tasks

This section estimates the stochastic structure $g_M(r|v)$ in mirror tasks. Because mirrors have known genuine valuations, we can directly observe and measure mistakes – the foundation of our Step 1 approach.

4.1 Distributions of Responses

Figure 1 shows the distributions of responses for each deterministic mirror. The genuine valuation is indicated by the red vertical line. We highlight several consistent features of these distributions.

Obs 1 The peak of the distribution is always the genuine valuation. The probability mass at the genuine valuation is *remarkably stable* across p for the same prize, at around 0.3 under prize \$25 and around 0.5 under prize \$100.

Obs 2 Fixing a prize, the probability masses of responses adjacent to the peak are *stable* across p. For example, when the prize is \$25, the responses immediately before the peak have probability masses around 0.15.

Obs 3 Regardless of p, there are non-negligible probability masses at the extreme responses (e.g., \$0, \$24, \$25 under prize \$25), and at certain mid-range responses (e.g., \$9 and \$10 under prize \$25). The probability masses at these responses are *stable*.[5]

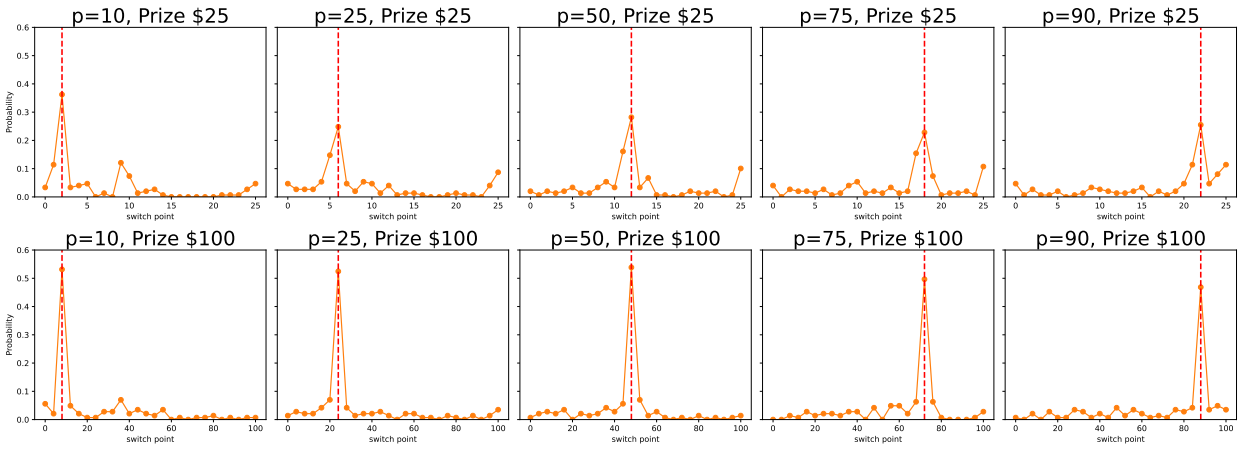

Figure 1: Distributions of responses in mirrors tasks

These observations suggest that the responses in mirror tasks mainly fall into three types, one rational choice and two types of mistakes. First, some subjects perfectly understand

[5]This observation is stronger under prize \$25, but is still visible under prize \$100.

the task and choose the genuine valuation (Obs 1). Second, instead of choosing the genuine valuation, some subjects commit *difference-dependent mistakes* by choosing a response close to the rational one (Obs 2). Such mistakes likely arise from subjects who generally understand the concept of mirrors and respond systematically to the probability parameter p of the corresponding lottery, but who occasionally miscalculate the mirror's payout or misreport their choice in the price lists. Finally, some subjects commit *unconditional mistakes* by completely ignoring changes in p (Obs 3). This type of mistake potentially reflect subjects who are inattentive throughout the experiment, or who employ some decision-making heuristic that disregards information about p. Moreover, the relative frequencies of these three types of responses are stable across mirror tasks, as indicated by the stability of the probability masses documented in Observations 1–3.

Crucially, the stochastic structure we observe in mirror tasks are inconsistent with commonly employed assumptions about the stochastic structures in valuation tasks. For example, Bruhin, Fehr-Duda and Epper (2010) assume that the stochastic structure adds a zero-mean normal noise to the genuine valuation, which would imply the distribution of response to be normally distributed and symmetric around the genuine valuation. In the meantime, Barseghyan et al. (2013) assume, following McFadden (1974), that the noise is a Type-I extreme value distribution added to the utility. This assumption would lead to a single-peaked probability mass function of responses, with the peak at the genuine valuation. Both assumptions are inconsistent with the actual stochastic structure in Figure 1. As a result, had we not known the genuine valuations in mirrors and need to estimate them from the responses, adopting these assumptions would lead to inaccurate conclusions about the genuine valuations.

4.2 Mixture Modeling of Responses

The three types of responses identified above, together with the stability of their relative frequencies, naturally motivate a statistical model that represents the distribution of responses as a probability mixture of three component distributions, each corresponding to one type of responses. This framework is commonly referred to as a *mixture model* in statistics. Now, we introduce some notations. Each observation in our mirror valuation data set $D_M = \{(m, r_m)\}$ consists of a response $r_m \in R$ when facing a mirror $m \in M$. For each mirror m, the genuine valuation is known and denoted $v(m)$. The set of *in-sample* genuine valuation is $V_M = \{v(m) : m \in M\} \subseteq A$, while the set of *out-of-sample* genuine valuations is $V_M^c = A - V_M$. To make the definitions concrete, under prize \$25, the set of possible responses is $A = \{0, 1, 2, \ldots, 25\}$, and the set of in-sample genuine valuations is $V_M = \{2, 6, 12, 18, 22\}$. Under prize \$100, the set of possible responses is $A = \{0, 4, 8, \ldots, 100\}$, and the set of

in-sample genuine valuations is $V_M = \{8, 24, 48, 72, 88\}$.[6]

Before describing the model in detail, let us first explain why we need to estimate a statistical model describing $g_M(r|v)$. Our immediate goal is to estimate the function $g_M(r|v)$. For in-sample genuine valuations ($v \in V_M$), the conditional probabilities $\{g_M(r|v) : a \in R, v \in V_M\}$ can be estimated nonparametrically by the empirical frequencies of r conditional on v, as illustrated in Figure 1. However, for out-of-sample genuine valuations ($v \in V_M^c$), without further assumptions, we do not know the conditional probabilities $\{g_M(r|v) : a \in R, v \in V_M^c\}$. There are two possible strategies to expand the domain of $g_M(r|v)$ from in-sample to out-of-sample. First, we could expand the set of mirrors that we elicit valuation for, so that $V_M = A$. However, this approach would substantially increase the number of observations required, since there will be $|R|^2$ possible combinations of r and v, compared with $5 \times |R|$ in the current data. Alternatively, we can exploit the empirical regularities documented in Observations 1–3 by specifying a statistical model that incorporates the three observed types of responses, and use the statistical model to *predict* $g_M(r|v)$ for out-of-sample v. The mixture model serves exactly this purpose, and we will later show that it indeed achieves strong out-of-sample predictive power.

Formally, the mixture model is specified in Model 1 as a mixture of three component distributions:

1. A degenerate *rational distribution* $\mathbb{1}[r = v]$, which assigns probability one on the genuine valuation v;

2. A *difference-dependent* mistake distribution $h_{dd}(r - v)$, which allocates probability based on the difference $r - v$, restricted to a window of radius K;

3. A *unconditional* mistake distribution $h_{uc}(r)$, which allocates probability mass independent of v.

[6]For example, for the mirror of the lottery $(10\%, \$100; 0)$, which involves a sure payout of \$10, the subject should choose a response of \$8, since the subject prefers the mirror to \$8, but prefers \$12 (the sure payment in the next row) to the mirror. For the other mirrors, their corresponding $v(m)$ are computed accordingly.

Model 1 (Mixture Model for Mirror Responses)**.**

$$g_M(r|v;\Psi) = w_{rational}\mathbb{1}[r=v] + w_{dd}h_{dd}(r-v) + w_{uc}h_{uc}(r) \tag{5}$$
$$s.t \quad w_{rational} + w_{dd} + w_{uc} = 1$$

$$h_{dd}(x) = \begin{cases} \pi_x & \text{if } 1 \le |x| \le K \\ 0 & \text{otherwise} \end{cases} \tag{6}$$
$$s.t \quad \sum_{x\in\{-K,\ldots,-1,1,\ldots,K\}} \pi_x = 1$$

$$h_{uc}(r) = \lambda_r \tag{7}$$
$$s.t \quad \sum_{r\in R} \lambda_r = 1$$

The complete set of parameters in this model is

$$\Psi = \{w_{rational}, w_{dd}, w_{uc}, \{\pi_x\}_{x\in\{-K,\ldots,-1,1,\ldots,K\}}, \{\lambda_r\}_{r\in R}\}.$$

We impose no restrictions on the functional forms of the unconditional distribution $h_{uc}(a)$. For the difference-dependent mistake distribution $h_{dd}(a-v)$, the only restriction is that $h_{dd}(a-v)$ can only place probability to r that is within a window of radius K around v. Otherwise, the functional form is fully flexible. In our data, it is feasible to estimate such a flexible model because the set of possible responses R is finite. Crucially, since the mixture model is set up and estimated under flexible functional forms, we avoid any bias from misspecified functional forms in the process of domain expansion.

The difference-dependent mistake distribution $h_{dd}(a-v)$ are parameterized by probabilities $\{\pi_x\}_{x\in\{-K,\ldots,-1,1,\ldots,K\}}$ that sum to one. The unconditional distribution $h_{uc}(a)$ is parameterized by probabilities $\{\lambda_a\}_{a\in R}$, also summing to one. In the model, the window size of the difference-dependent mistake distribution, K, is a "hyperparameter," which is not estimated, but chosen by the econometrician. In our main specification, we choose $K=2$ under prize \$25, and accordingly choose $K=8$ under prize \$100.[7] Other choices of K yield similar results.

The log likelihood function of this model is given by

$$\mathcal{L}_M(\Psi; D_M) = \sum_{(m,r_m)\in D_M} \log g_M(r_m|v(m);\Psi). \tag{8}$$

[7]We choose $K=2$ under prize \$25 for two reasons. First, for the mirror of lottery $(\$25, 10\%; \$0)$ $(v(m)=2)$, if K is greater than 2, the difference-dependent distribution will place positive probability on $v(m)-K$, which is smaller than 0 and is not a possible response. This raises technical difficulty. Second, $K=2$ seems to be a good approximation based on visual inspection – outside this window, the difference-dependent mistakes seem to be negligible. The reason why $K=8$ under prize \$100 is similar.

We estimate the parameters Ψ by maximum likelihood estimation

$$\hat{\Psi} = \arg\max_{\Psi} L_M(\Psi; D_M),$$

through the expectation-maximization algorithm (Dempster, Laird and Rubin, 1977), which is a standard method of estimating mixture models.

4.3 The Estimated Mixture Model

Figure 2 presents the estimated distributions of the two mistake components, h_{dd} and h_{uc}, under prize \$25. The difference-dependent mistakes occur primarily when subjects choose the response immediately below the genuine valuation, while other deviations are relatively rare. The unconditional mistakes disproportionately places weights on the highest response (25), as well as on mid-range responses (9 and 10) and the lowest response (0). In terms of estimated mixture weights, unconditional mistakes account for $w_{uc} = 0.538$, which is more than twice the weight of the difference-dependent mistakes ($w_{dd} = 0.203$). Both types of mistakes are important, however, with magnitudes comparable to the rational component ($w_r = 0.259$).

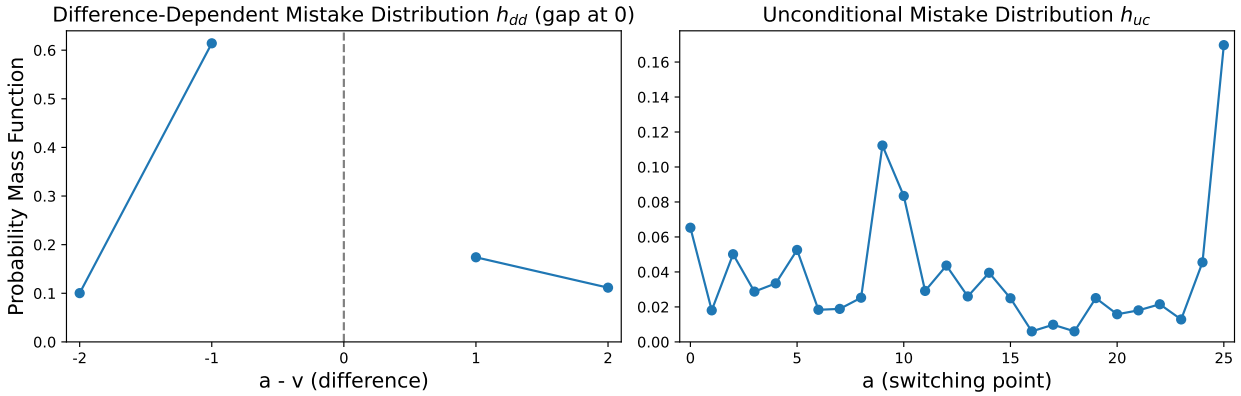

Figure 2: The distributions d_{dd} and h_{uc} under prize \$25

Note: h_{dd}'s domain is $\{-2, -1, 1, 2\}$ due to our choice of $W = 2$

Figure 3 reports the corresponding estimates under prize \$100. Difference-dependent mistakes are concentrated on responses just below or just above the genuine valuation, but not further away. Unconditional mistakes are most prominent at the mid-range points (e.g., 36, 48, 56), the low responses (e.g., 8, 16), and the two highest responses (96 and 100). The estimated mixture weights again indicate that unconditional mistakes are substantially more frequent than difference-dependent mistakes ($w_{uc} = 0.396$ vs. $w_{dd} = 0.108$).

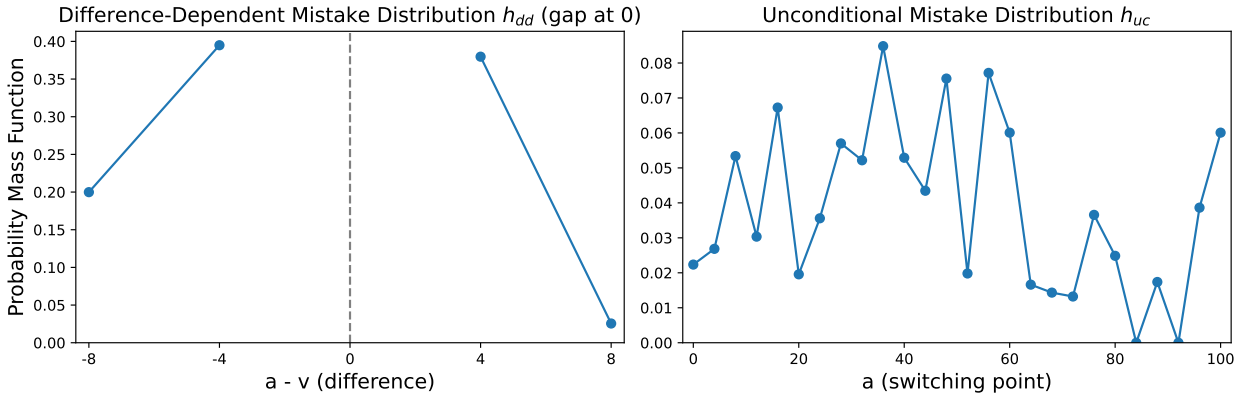

Figure 3: The distributions d_{dd} and h_{uc} under prize \$100

Note: h_{dd}'s domain is $\{-8, -4, 4, 8\}$ due to our choice of $W = 8$

To assess overall fit, Figure 4 juxtaposes the empirical distributions of responses with the distributions simulated from the estimated mixture model. The simulated distributions reproduce the key empirical features documented in Observations 1–3, delivering a great in-sample fit. Importantly, the good fit does not arise simply from flexibility of the functional forms of h_{dd} and h_{uc}. Despite the flexibility, the mixture model imposes strong restrictions on the distributions of the responses by requiring the distributions to be fixed-weight mixtures of the three types of responses. In Section 4.4, we will further show that the mixture model has not just great in-sample fit, but strong out-of-sample predictive power, which is not achievable through functional form flexibility alone. Therefore, the close alignment between simulated and empirical distributions therefore provides strong evidence in favor of the three-type interpretation of the mirror response distributions.

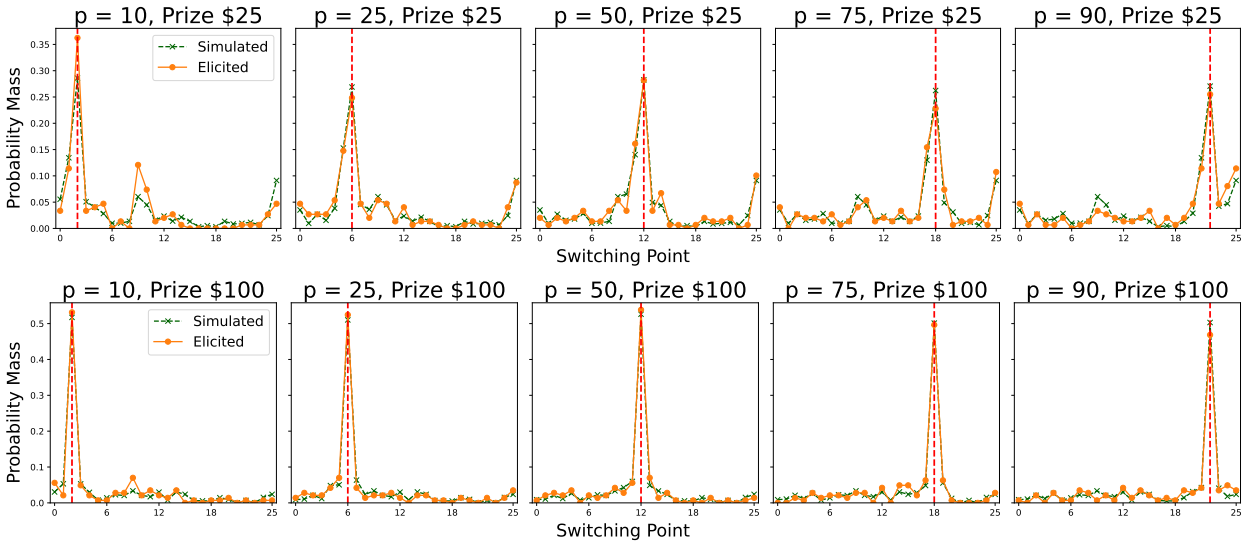

Figure 4: Simulated distribution of mirror responses against the empirical distribution

4.4 Out-of-Sample Predictive Power

The mixture model provides a great fit of the function $g_M(r|v)$ in-sample over $\{g_M(r|v) : a \in R, v \in V_M\}$. However, how well does it *predict* the function out-of-sample over $\{g_M(r|v) : a \in R, v \in V_M^c\}$? Borrowing techniques from machine learning, we now assess the out-of-sample predictive power of the mixture model, and compare its predictive power with several benchmark specifications of the stochastic structure.

The in-sample fit of the mixture model is not suitable for assessing its out-of-sample predictive power, since the model is estimated using exactly the same data set that the model is assessed on (see, e.g., Hastie, Tibshirani and Friedman, 2009, Section 7.4). Instead, to assess the out-of-sample predictive power, we need to use a different data set than the one on which the model parameters are estimated. To this end, we employ the algorithm of leave-one-m-out cross-validation. This algorithm works by first excluding all observations involving a fixed mirror m from D_M and generating a sub-data set $D_{-m}^{estimation}$. Next, the procedure estimates the parameters of the mixture model using the new data set $D_{-m}^{estimation}$ and generate parameter estimates $\hat{\Psi}_{-m}$. Then, the procedure evaluates the parameter estimates $\hat{\Psi}_{-m}$ using the sub-data set $D_m^{validation}$ that includes all observations involving the mirror m. The metric for our assessment is the log likelihood $\mathcal{L}(\hat{\Psi}_{-m}; D_m^{validation})$. Therefore, the data set that the model is evaluated on, $D_m^{validation}$, is entirely different from the data set on which the model is estimated, $D_{-m}^{estimation}$. Finally, the algorithm repeats the process above for all $m \in M$, and use the average log likelihood

$$\bar{\mathcal{L}} := \frac{1}{|M|} \sum_{m \in M} \mathcal{L}(\hat{\Psi}_{-m}; D_m^{validation}) \tag{9}$$

as the measure of the out-of-sample *predictive power*. Similar cross-validation procedures are widely adopted in machine learning in evaluating the out-of-sample performance of predictive models (Hastie, Tibshirani and Friedman, 2009, Section 7.10), and have been increasingly adopted in economics (see, e.g., Kleinberg et al., 2018, Fudenberg et al., 2022, Mullainathan and Obermeyer, 2022).

The predictive power $\bar{\mathcal{L}}$ is a relative measure that only becomes meaningful when compared between different models. To this end, we also construct the same measure, $\bar{\mathcal{L}}$, for a few benchmark specifications of the stochastic structure $g^M(a|v)$ for comparison. For each of these benchmark specifications, we first estimate its parameters, if any, and then compute the predictive power $\bar{\mathcal{L}}$ through the same leave-one-m-out cross validation explained above. The benchmark specifications include the following.

- **Uniform Choice Lower Bound** assumes uniform choice of responses conditional on any v: $g_M(r|v) = \frac{1}{|R|}$. This "guess the choice at random" model can be viewed as a

"lower bound" of $\bar{\mathcal{L}}$ since any reasonable prediction should not be worse than uniform choice.

- **Omniscient Upper Bound** uses the empirical distribution of r conditional on v observed in the data to "predict" itself:

$$g_M(r|v) = g_{M,empirical}(r|v) := \frac{\sum_{(m,r_m)\in D_M} \mathbb{1}[r_m = r, v(m) = v]}{\sum_{(m,r_m)\in D_M} \mathbb{1}[v(m) = v]},$$

  the predictive power $\bar{\mathcal{L}}$ achieves the theoretical upper bound for the data set D_M. This bound is unattainable in practice but provides a natural upper scale for $\hat{\mathcal{L}}$.

- **Uniform Mistake** assumes that with probability $1-\epsilon$, the subjects choose the genuine valuation, and with the remaining probability ϵ, they make a mistake by choosing responses uniformly in R. The formal definition is in Equation (3). The set of parameters in this stochastic structure is $\{\epsilon\}$. This stochastic structure is a generalization of Harless and Camerer (1994).

- **Normal Mistake** assumes that the mistakes are normally distributed and simply add white noise to the genuine valuation, as in Equation (2). The variance of the mistake is assumed to be heterogeneous across individual subjects. The set of parameters is $\{\sigma_i$ for each subject $i\}$. The stochastic structure is a generalization of Hey and Orme (1994) that is adopted by Bruhin, Fehr-Duda and Epper (2010).[8]

We also include two restricted versions of Model 1 in the comparison. These models are described in terms of restrictions we impose on the parameters Ψ, but otherwise they are identical to Model 1.

- **Uniform Unconditional Mistake** restricts the unconditional mistake in Model 1 to be a uniform distribution over all $r \in R$: $\lambda_r = \frac{1}{|R|}$ for all $r \in R$. This differs from the Uniform Mistake above by allowing for difference-dependent mistakes.

- **Unconditional Mistake Only** restricts Model 1 by not allowing for difference-dependent mistakes: $w_{dd} = 0$.

Table 3 shows the predictive power of each specification of stochastic structure. To make the table easier to read, the predictive power is normalized by a linear transformation so that

[8]When implementing the normal stochastic structure, we impose a restriction that $\sigma_i > \underline{\sigma}$ for all i, where $\underline{\sigma}$ is a lower bound. The rationale can be explained by an example. Suppose that a subject i chooses $r_{i,m} \neq v(m)$ for m, but $r_{i,m'} = v_{m'}$ for all other $m' \neq m$. Without the lower bound on σ_i, the $\sigma_{i,-m}$ estimated using $D_{-m}^{estimation}$ will be 0. However, with $\sigma_{i,-m} = 0$, the likelihood of the observation (m, r_{im}) is zero and the log likelihood is negative infinity. Therefore, by imposing a lower bound for σ_i, we avoid this scenario and get better predictive power. The lower bound $\underline{\sigma}$ is chosen by maximizing the predictive power $\bar{\mathcal{L}}$.

	Prize $25	Prize $100
Baseline Mixture Model 1	**0.664**	**0.891**
Uniform Choice Lower Bound	0	0
Omniscient Upper Bound	1	1
Uniform Mistake	0.463	0.832
Normal Mistake	0.289	0.388
Uniform Unconditional Mistake	0.639	0.867
Unconditional Mistake Only	0.337	0.494

Table 3: The normalized predictive power for each specification of stochastic structure

the Omniscient Upper Bound has predictive power of 1, and Uniform Choice Lower Bound has predictive power of 0.[9] Three main takeaways arise from Table 3. First, the mixture model has better predictive power than the stochastic structures of Uniform Mistake and Normal Mistake commonly assumed in the literature. Second, comparing the mixture model with its two restricted versions, the mixture model has better predictive power than the Uniform Unconditional Mistake specification, and much better than the Unconditional Mistake Only specification. These findings show that both the existence of the difference-dependent mistakes and the flexible functional form of the unconditional mistakes are important in describing the stochastic structure.[10] Third and perhaps most importantly, the predictive power of the mixture model is not too far away from that of the Omniscient Upper Bound, especially under prize $100. Thus, the mixture model cannot be improved much by adopting a different specification of the stochastic structure.

4.5 Discussion

By estimating Model 1, we obtain an accurate description of the stochastic structure $g_M(r|v)$ in mirror tasks, for both in-sample and out-of-sample. We measure how good the model describes the stochastic structure mainly through the metric of out-of-sample predictive power, not Akaike information criterion (AIC), or Bayesian information criterion (BIC). Although, generally speaking, all of these are measures of how good the model describes the stochastic

[9]The notion of normalized predictive power is related to the notion of *completeness* in Fudenberg et al. (2022). However, the key distinction is that the object that we try to predict a *distribution* of values, while most applications of Fudenberg et al. (2022) pertains to providing a *point prediction*. The fact that we predict distributions makes it very hard for us to define and estimate the "irreducible noise" as defined in Fudenberg et al. (2022), which is a key component in measuring completeness. In the meantime, the normalized predictive power can be seen as a lower bound of completeness, because any definition of irreducible noise will lead to a completeness higher than the normalized predictive power.

[10]A "Difference-Dependent Mistake Only" specification is not possible, since the combination of the rational distribution and the difference-dependent mistake only allows the support of r to be $\{v-K,\ldots,v+K\}$. However, the empirical support of r will be outside this window, and without a component distribution that allows for positive probability at all $r \in R$, the log likelihood will approach negative infinity.

structure, the out-of-sample predictive power has a few advantages. First, one of our primary goals for a model of stochastic structure is that we want to predict out-of-sample values of $g_M(r|v)$ using its in-sample values, which is a prediction task that is naturally assessed with out-of-sample predictive power. Second, compared with AIC and BIC, the out-of-sample predictive power imposes minimal assumptions and is a more robust measure of how good the model describes the stochastic structure. Meanwhile, we note that when using AIC as the model selection criterion, Model 1 is still selected as the best performing model over its two restricted versions and the two benchmarks of Uniform Mistake and Normal Mistake.

Beyond expanding the domain of $g_M(r|v)$, the mixture model provides valuable insights over the deviations of the responses from the genuine valuations. In particular, since the mixture model turns out to be an accurate description of the stochastic structure, we know that the two types of mistakes, difference-dependent and unconditional, do capture the main source of mistakes, while other types of mistakes – for example, submitting a response of $v/2$ when the genuine valuation is v – are relatively rare. In Section 6, we will further show that the same two types of mistakes can largely capture the stochastic structure in two other data sets collected with different methods. To be clear, our finding does not prove that other mistake forms don't exist, but shows that these two types of mistakes capture the dominant empirical regularities.

5 Estimating Risk-Induced Valuations

Having estimated the stochastic structure in mirror tasks in Section 4, we now apply it to lottery responses to recover risk-induced valuations (Step 2). Lottery tasks differ from mirror tasks in two important respects. First, unlike in mirror tasks, the genuine valuation is *unknown*. Second, the genuine valuation may be *heterogeneous* across subjects. In this section, we apply the estimated stochastic structure from mirror tasks to recover the distribution of risk-induced valuations from lottery responses.

5.1 Distributions of Responses

Figure 5 shows the distributions of responses in lottery tasks. The risk-neutral benchmark – the response that would have been chosen by a risk-neutral agent who makes no mistakes – is marked with a red vertical line. Three observations emerge that are parallel to Observations 1-3 in mirror tasks.

Obs 1' Unlike the mirror tasks, where the distribution features a sharp single peak at the genuine valuation, the lottery tasks exhibit a *peak area* – a cluster of responses with

relatively high probability mass. This peak area is typically centered around the expected value of the lottery.

Obs 2' For a given prize, the height of the peak area is approximately stable across different probabilities p.

Obs 3' As in the mirror tasks, there are substantial and stable probability masses at the extreme responses and at selected mid-range responses.

Figure 5: Distributions of responses in lottery tasks

Together, these observations suggest that the empirical regularities documented in the mirror tasks – the existence of peaks, stable near-peak deviations, and stable extreme or mid-range probability masses – seem to carry over to the lottery tasks in modified form. The key difference is that, the heterogeneity of genuine valuations in lottery tasks make the responses less concentrated. Similar response distributions – with stable extreme responses and near-peak concentrations – also appear in other lottery valuation datasets, including (e.g., Enke and Graeber, 2023, Zhang, 2025).

5.2 Mixture Modeling of Risk-Induced Valuations

Having estimated the stochastic structure in mirror tasks in Section 4, we now apply it to recover the distribution of genuine lottery valuations, following the approach outlined in Section 3.

Now we need some notations about the lottery valuation data. Each observation in our lottery valuation data set $D_L = \{(l, r_l)\}$ consists of response $r_l \in R$ facing mirror $l \in L$. The mixture model can be described by the set of equations in Model 2. Equation (10) writes down the probability of choosing an response r when facing a lottery l, $f^l(r)$, using the

law of total probability. When recovering the RIV $\tilde{q}^l(v)$, Equation (10) uses the stochastic structure in mirrors, $g_M(r|v;\hat{\Psi})$, to approximately the one, $g(r|v)$, in lotteries. The stochastic structure $g_M(r|v;\hat{\Psi})$ is defined in Model 1 and estimated with the mirror data D_M through Equation (8).

Model 2 (Mixture Model of Lottery Responses)**.**

$$f^l(r) = \sum_v g_M(r|v;\hat{\Psi})\tilde{q}^l(v) \tag{10}$$
$$s.t \quad \sum_{v\in R} \tilde{q}^l(v) = 1 \quad \text{for any } l$$

We impose no restrictions on the functional form of $g^l(a)$ to avoid any bias from misspecified functional forms. Therefore, the set of parameters in Model 2 includes the complete probability mass function in the RIV $\tilde{q}^l(v)$:

$$\Lambda = \{\tilde{q}^l(v) : l \in L, v \in R\}.$$

The log likelihood function of Model 2 is given by

$$\mathcal{L}_L(\Lambda;\hat{\Psi}, D_L) = \sum_{(l,r_l)\in D_L} \log f^l(r_l).$$

The notations suppress the dependence of $f^l(a_{il})$ on the parameters Λ and $\hat{\Psi}$. Again, we estimate the parameters Λ by maximum likelihood estimation:

$$\hat{\Lambda} = \arg\max_{\Lambda} \mathcal{L}_L(\Lambda;\hat{\Psi}, D_L).$$

While we adopt this approximation $g_M(r|v) \approx g(r|v)$ to proceed with estimation of the RIV, we recognize that the assumption may not hold exactly. In particular, Martínez-Marquina, Niederle and Vespa (2019) document that subjects make more mistakes in a risky environment compared with a related, similar risk-free environment. If their findings generalize to our setting, our approach of using $g_M(r|v)$ to approximate $g(r|v)$ will fail to correct for the risk-specific mistakes, biasing the RIV $\tilde{q}^l(v)$ away from the distribution of genuine valuations. This emphasizes the need to be careful when interpreting the RIV. We leave a more thorough discussion of the assumption $g_M(r|v) \approx g(r|v)$ and the interpretations of the RIV to Section 5.4.

5.3 The Risk-Induced Valuations

Figure 6 shows the distributions of RIV along with the empirical distributions of responses. Table 4 provides the averages and medians of the distributions of RIV, along with the risk-neutral benchmark for comparison.

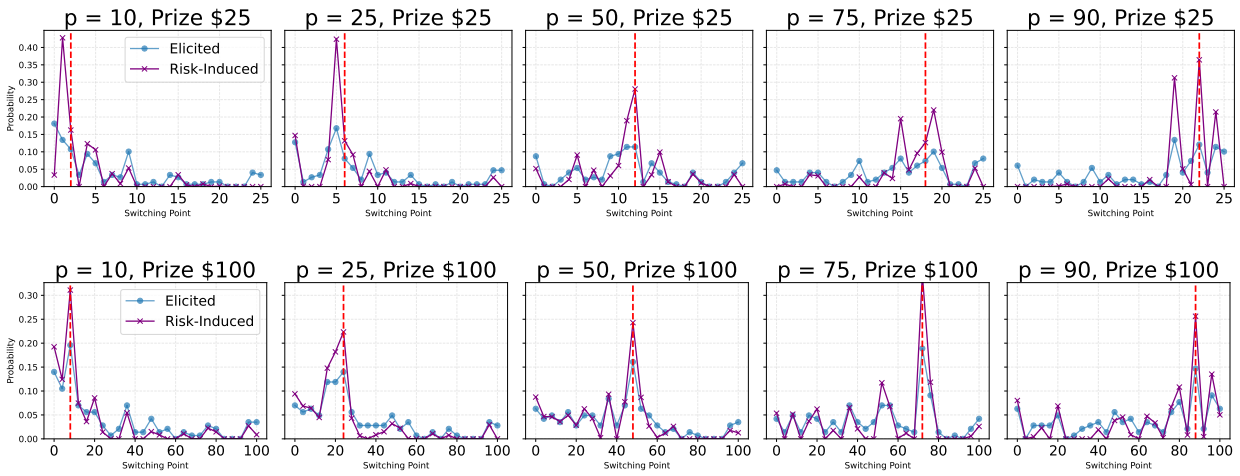

Figure 6: Distributions of risk-induced valuations

Panel A: Prize $25

p	Risk-Neutral	Mean		Median		RIV-Based Risk Attitude		
	Benchmark	RIV	Responses	RIV	Responses	Averse	Neutral	Loving
10	2	3.34	6.44	2	4	0.46	0.16	0.38
25	6	5.54	8.10	5	6	0.65	0.13	0.22
50	12	11.11	11.35	12	11	0.49	0.28	0.23
75	18	16.45	14.44	17	15	0.50	0.13	0.37
90	22	20.87	16.80	22	19	0.42	0.36	0.21

Panel B: Prize $100

p	Risk-Neutral	Mean		Median		RIV-Based Risk Attitude		
	Benchmark	RIV	Responses	RIV	Responses	Averse	Neutral	Loving
10	8	16.68	26.52	8	12	0.32	0.31	0.37
25	24	21.94	29.85	20	24	0.60	0.22	0.18
50	48	35.23	38.63	44	40	0.57	0.24	0.18
75	72	53.54	50.49	64	52	0.51	0.34	0.15
90	88	67.99	60.76	80	68	0.55	0.26	0.19

Table 4: Average and median risk-induced valuations

Comparing RIV with responses, the most striking observation is that RIV is generally closer to the risk-neutral benchmark than responses. This distinction exhibits in both the complete distributions in Figure 6 and the summary statistics in Table 4. As for the distributions, the RIV are more concentrated around the risk-neutral benchmark than the responses, while there are fewer extreme RIV than extreme responses.[11] More quantitatively, we measure the fraction of responses that fall within the window consisting of the risk-neutral benchmark and the two responses immediately above it and the two immediately below it. Under prize \$25, 65.2% of RIV are within this window, while only 42.0% of responses are. Under prize \$100, the numbers are 55.2% and 41.7%, respectively. Turning to the summary statistics for the RIV, under prize \$25, the median subject's RIV are fairly close to the risk-neutral benchmark, as well as the average subject. Under prize \$100, both the median and the average subject exhibit weak risk aversion[12] in terms of RIV, except for $p = 10$ where the average is risk-loving and the median is risk-neutral. In contrast, for the responses, under both prize levels and regardless of the choice of the summary statistic, the subjects exhibit pronounced risk-lovingness facing a low probability prize, and pronounced risk-aversion facing a high probability prize, replicating the "fourfold pattern" documented in Tversky and Kahneman (1992).

The shift from responses to risk-induced valuations in Figure 6 and Table 4 can be understood through the stochastic structure in mirror tasks. Recall from Section 4 that subjects make unconditional mistakes by placing weight on extreme or mid-range responses, regardless of p. The mixture model uses this empirical pattern of mistakes to reweight the lottery data. In particular, when we adopt the approximation $g(r|v) \approx g_M(r|v)$, an extreme or mid-range response observed in a lottery task is classified as more likely to arise from unconditional mistakes unrelated to risk than from the underlying risk-induced valuations. By contrast, a response that is not flagged as a typical mistake in the mirror tasks is more likely to be attributed to risk. In this way, the model systematically discounts probability mass at extreme and mid-range responses.

Moreover, Table 4 lists the fractions of risk-averse, risk-neutral, and risk-loving RIV. The majority of deviations from the risk-neutral benchmark are in the direction of risk-aversion, and risk-lovingness is less common. Even for the lotteries involving a prize with probability $p = 10$, risk-averse RIV are as common as risk-loving RIV. This pattern stands in clear contrast with the distributions of the responses, where under $p = 10$, the majority of responses

[11] By extreme responses, we refer to those responses that would manifest extremely risk-averse or risk-loving preferences, such as choosing a response of \$25 facing the lottery $(\$25, 10\%; \$0)$ or a response of \$0 facing the lottery $(\$25, 90\%; \$0)$.

[12] We use *risk-aversion* to refer to the scenario where the RIV is smaller than the risk-neutral benchmark, but not in a strict welfare-relevant sense due to the above-mentioned caveats of interpreting RIV as genuine valuations. The terminology is used similarly for other risk attitudes.

are risk-loving,[13] as documented in Tversky and Kahneman (1992) and replicated in countless follow-up works.

The two findings above – concentration of RIV around the expected values and predominant risk-aversion given deviating from the expected values – have a few implications. First, since the expected utility theory (EUT) predicts approximate risk-neutrality under small stakes (Rabin, 2000), the RIV are more in line with the predictions of EUT than the responses. In addition, the facts that the RIV exhibit near risk-neutrality under the smaller prize of \$25 and weak risk-aversion under the larger prize of \$100 are consistent with EUT under concave utility functions. Together, these patterns suggest that the previous behavioral economics literature has overestimated the degree of deviation of risk *preferences* from EUT. Second and relatedly, our evidence suggests that the past estimates of prospect theoretic parameters in the literature should not be interpreted as solely reflecting genuine preferences. In particular, the strong overweighting of small probabilities documented in the literature should not be solely attributed to genuine preferences for gambling, but at least partly to mistakes in responding to preference elicitation tasks. Finally, the RIV are less heterogeneous than the responses, suggesting the heterogeneity of *responses* in the population (e.g, Bruhin, Fehr-Duda and Epper, 2010) exaggerates the underlying heterogeneity of risk *preferences*.

5.4 Discussion

Facing the identification problem in lottery tasks, we choose to use the stochastic structure in mirror tasks to approximate it. How close the RIV is from the genuine valuations depends on the quality of this approximation. There are two main arguments that we believe offer support for our use of $g_M(r|v)$ to approximate $g(r|v)$.

First, we emphasize the fact that any researcher who aims to measure risk *preferences* needs to make assumptions over the stochastic structures that cannot be directly evaluated and are inevitably misspecified to some degree. From Section 4, we already know that the most common assumptions in the literature – Uniform Mistake and Normal Mistake – are clearly inaccurate in the context of mirror tasks. Moreover, from Observations 1'–3', it does not seem like these previous assumptions will become a good approximation in lottery tasks. Therefore, the best we can do in this case to get closer to the genuine valuations is to take into consideration the two types of mistakes that we already documented in mirror tasks, by using $g_M(r|v)$ to approximate $g(r|v)$. Even if $g(r|v)$ differs from $g_M(r|v)$ by bringing into *more mistakes and different types of mistakes* when making decisions under risk, as suggested by Martínez-Marquina, Niederle and Vespa (2019), we believe it is still better to account for the *mistakes we already document* in the context of mirror tasks than not, since they

[13] This can be seen from the medians of the responses.

are likely still present in the context of lottery tasks. In this sense, the mistake identified in mirror tasks can be seen as the "lower bound" of all mistakes in lottery tasks.

Second, the quality of the approximation crucially depends on the primary source of mistakes in lottery tasks. If the mistakes in lottery tasks primarily stem from a set of common mistakes that are present in both risky (lottery) and risk-free (mirror) tasks, then $g_M(r|v)$ will be a good approximation of $g(r|v)$, and the RIV will closely track the genuine valuations. However, if instead the primary source of mistake is risk-specific, the RIV will not be a good measure of the genuine valuations. In fact, various theories of lottery valuations are built on the premises that the primary source of mistakes in lottery tasks is not risk-specific. For example, the prospect theory, which is intended to describe behaviors under risk that depart from rationality, is based on psychological principles of diminishing sensitivity and reference dependence that are not risk-specific (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992, Wakker, 2025). More recently, the theory of cognitive imprecision (Woodford, 2012, Khaw, Li and Woodford, 2021, Frydman and Jin, 2021) and the theory of salience (Bordalo, Gennaioli and Shleifer, 2012) offers explanations to many well-documented anomalies in decisions under risk. However, both of these theories are based on psychological foundations that are not risk-specific – cognitive imprecision is based on imperfections in perceptions of numerical quantities, and salience is based on the context-dependent allocation of the scarce resource of attention. However, since both theories have psychological foundations that are not risk-specific, similar forces should also be present in risk-free settings.

To make a more quantitative examination of the primary source of mistakes in lottery tasks, we turn to Martínez-Marquina, Niederle and Vespa (2019). Their experiment consists of two related settings, one risk-free and another risky, similar to the mirror and lottery tasks in the current paper. Importantly, in both of their settings, there is a known correct decision. As a result, mistakes can be detected and measured independent of risk preferences. They find that in their risk-free setting, 58.5% of subjects make mistakes, while in their risky setting, an additional 21.8% of subjects make mistakes, making the total fraction 80.3%. The above figures suggest that most mistakes in the risky setting already appear in the risk-free setting, and offers additional support to the claim that the primary source of sources of mistakes is shared in risky and risk-free settings.

However, the main point of Martínez-Marquina, Niederle and Vespa (2019) is that people make more mistakes in a risky setting than a risk-free setting. Moreover, even in our data, there is also evidence that $g_M(r|v)$ may miss some important aspects of $g(r|v)$. In particular, comparing the distributions of mirror responses in Figure 1 and those of lottery responses in Figure 5 under prize \$100, there are many more cases where subjects choose the smallest response \$0 in lottery tasks than mirror tasks, regardless of p. In addition, in lottery tasks under prize \$100, the masses at \$0 is stable across p. These patterns lead to the spike at

$0 for the RIV under prize $100, as is visible in Figure 6. We believe that the most likely explanation behind these spikes at $0 is that there are more subjects making unconditional mistakes in lottery tasks than mirror tasks, in the form of choosing a response of $0 regardless of p.

Despite the above caveats, we still believe using $g_M(r|v)$ to approximate $g(r|v)$ is probably the best approach we can realistically take. First, we note that there is a level of subjectivity in interpreting the spikes of RIV at $0. Although, in our opinions, the spikes most likely reflect unconditional errors, alternatively, a group of subjects with extremely risk-averse preferences can also generate these spikes. At the end of the day, neither of the possibilities can be directly verified and we need to go back to our subjective opinions in determining the genuine valuations. In contrast, by assuming $g(r|v) \approx g_M(r|v)$, where $g_M(r|v)$ is constructed in a data-driven and purely objective fashion using the mirror tasks, we tie our hands with principles and prevent this level of subjectivity. Second, by using the stochastic structure in the risk-free mirror tasks $g_M(r|v)$, the recovered RIV can be solely attributed to risk, and not to mistakes unrelated to risk, such as those documented in Section 4. In other words, the RIV can be viewed as the *treatment effects of risk* on responses. If we instead recover $q^l(a)$ through other assumptions over the stochastic structure, the recovered $q^l(a)$ will lose this clear interpretation.

Another threat to our use of $g_M(r|v)$ to approximate $g(r|v)$ is the possibility where the subjects do not fully understand the experimental instructions. In particular, if some subjects mistake mirror tasks for lottery tasks, the subjects may perceive the mirror tasks as risky, and the interpretation of RIV as the treatment effect of risk is no longer valid. To prevent this, we adopt our experimental instructions from Wu (2025), who explains the difference between lottery and mirror tasks through "clear and straightforward instructions, training to reinforce comprehension of the payoff structure, [and] subject screening through comprehension questions" (Wu, 2025). For more information on the training and subject screening, we refer the Reader to the Appendix and Wu (2025).

6 Applying the Mixture Modeling to Other Data Sets

In this section, we apply the same two-step procedure of recovering the RIV to other data sets and see if the main results generalize. We are interested in whether the main findings in Section 4 and Section 5 are robust in the data sets of Oprea (2024) and Zhang (2025). Both of these studies conduct experiment collecting valuations of lotteries and mirrors from the same set of subjects, allowing for seamless applications of the two-step procedure.

Oprea (2024) To assess the robustness and generalizability of our findings, we apply the same two-step procedure to data from Oprea's (2024) main experiment. Oprea (2024) elicits valuations of lotteries and their deterministic mirrors are elicited through price lists, where the subjects are recruited from Prolific, and the prize is $25. However, the main difference lies in the experimental instructions and subject screening procedures. Despite procedural differences, the data allow us to test whether our methodological findings generalize across experimental implementations. To implement the mixture modeling using Oprea's (2024) data with the greatest comparability with the current paper, we exclude from his data set the lotteries that involve losses, along with the mirrors corresponding to these excluded lotteries.

Figure 7 demonstrates that the mixture model Model 1 continues to provide excellent fit to mirror responses in Oprea's data. The simulated distributions closely track the empirical distributions across all probability levels, indicating that the three-component mixture model – with difference-dependent and unconditional mistakes alongside rational responses – captures the dominant patterns of behavior in mirror valuation tasks even under different experimental protocols. This robustness suggests that these two mistake types represent fundamental features of how subjects respond in mirror valuation tasks, rather than artifacts specific to our experimental design.

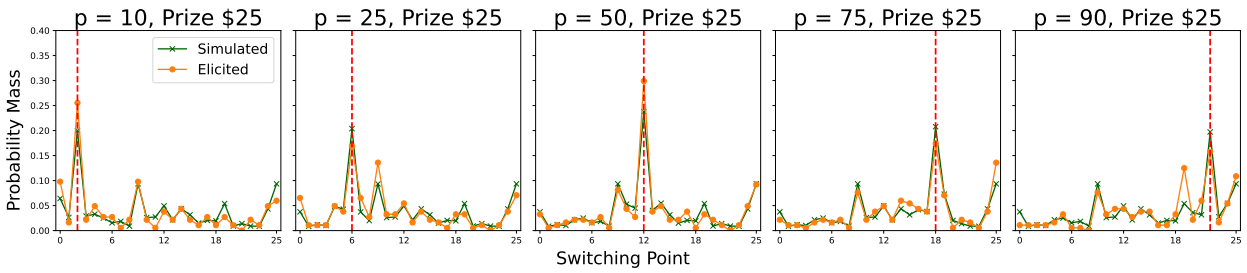

Figure 7: Model fit in mirror tasks using data from Oprea (2024)

Table 5 shows that the main substantive finding from Section 5 replicates in Oprea's data: The RIV are substantially closer to the risk-neutral benchmark than raw responses, making them more consistent with expected utility theory's prediction of approximate risk neutrality under small stakes (Rabin, 2000). This pattern is particularly striking for small-probability gains. For the lottery with $p = 10\%$, the median response is $8, reflecting strong risk-lovingness. In contrast, the median RIV equals the risk neutral benchmark ($2), and the mean RIV is only moderately above it ($3.21). This represents a dramatic shift from the pronounced risk-lovingness in raw responses toward risk neutrality after correcting for mistakes common to mirror and lottery tasks.

However, Oprea's data also reveal two important differences from our main results. First, for large-probability gains ($p = 75\%, 90\%$), the RIV exhibit risk aversion in Oprea's data but are close to risk neutral in ours, as evidenced by both means and medians. Second, for

small-probability gains ($p = 10\%$), while both datasets show RIV close to risk neutrality in terms of summary statistics, Oprea's data exhibit more risk-lovingness when RIV deviate from the expected value. Nonetheless, the core finding remains: correcting for mistakes observable in mirror tasks moves lottery valuations substantially closer to expected utility theory predictions.

p	Risk-Neutral	Mean		Median		RIV-Based Risk Attitude		
	Benchmark	RIV	Responses	RIV	Responses	Averse	Neutral	Loving
10	2	3.21	8.55	2	8	0.18	0.37	0.45
25	6	5.66	9.47	6	9	0.37	0.34	0.29
50	12	9.81	11.28	12	12	0.44	0.25	0.31
75	18	13.38	13.55	14	14	0.68	0.27	0.05
90	22	18.04	15.94	19	18	0.73	0.22	0.05

Table 5: Average and median risk-induced valuations in Oprea (2024)

Zhang (2025) We apply the mixture modeling approach to the data set of Zhang (2025), which has a few key advantages. First, Zhang's (2025) data set of lottery and mirror valuations is incentivized through the Becker-DeGroot-Marschak (BDM) mechanism (Becker, Degroot and Marschak, 1964), unlike the price lists in this paper. Therefore, a test of the mixture modeling on this data set expands the potential scope of the approach. Second and more importantly, Zhang's (2025) data set includes the *same set of subjects* performing the same set of lottery and mirror tasks twice, once with the help of a calculator (the *Calc* treatment) and once without (the *NoCalc* treatment). Zhang (2025) shows that elicited lottery valuations are somewhat different between the NoCalc and Calc treatments, and demonstrates that the differences are largely due to different stochastic structures in the two treatments, since the Calc treatment reduces mistakes through reducing the costs of performing explicit calculations. The fact that Zhang's (2025) data set includes repeated elicitations of lottery and mirror valuations under different stochastic structures provides a testing ground for whether RIV closely approximates *genuine valuations*. To be specific, if the RIV does approximate the genuine valuations, a necessary condition is that the recovered RIV should be similar across the NoCalc and Calc treatments, since both of them measure the genuine valuations of the same set of subjects, which is assumed to be the results of a stable underlying preference order and should not change across treatments. This necessary condition is testable with Zhang's (2025) data.

To maximize the comparability with the current study, we only use lotteries involving gains from (Zhang, 2025), while dropping any lotteries involving losses. Moreover, since the valuations under the BDM mechanism can take a continuum of possible values, we need to

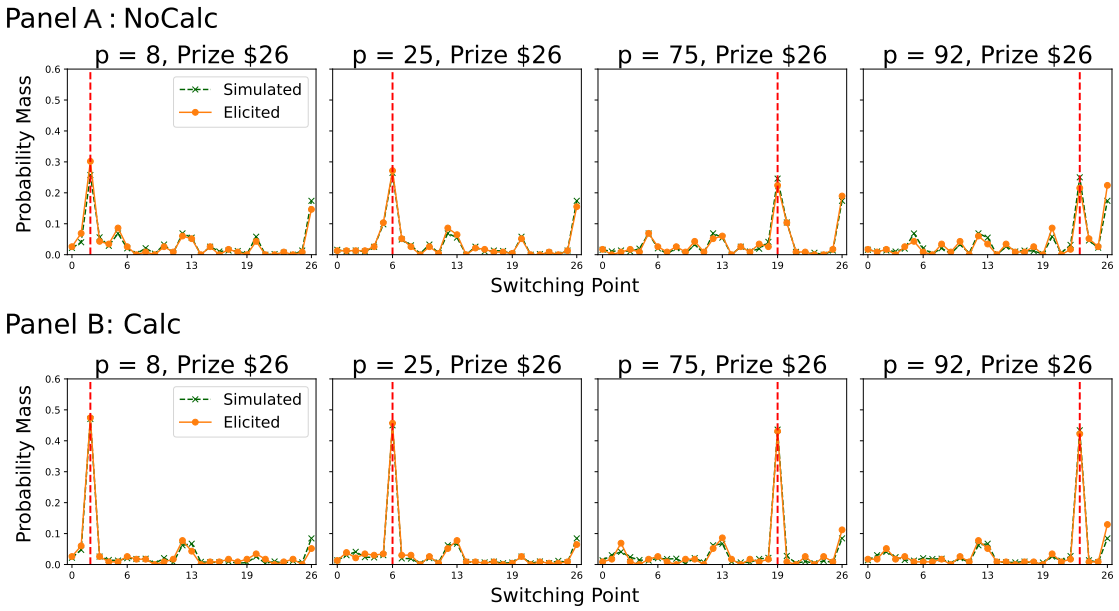

Figure 8: Model fit in mirror tasks using data from Zhang (2025)

discretize the BDM valuations, in order to make them suitable for estimating the flexible functional forms of h_{dd}, h_{uc}, and $g^l(v)$, which has been shown to be important in how well the mixture model describes the data (Section 4.4). To achieve this, we round down the BDM valuations to integers and generate the "responses." The RIV is then recovered by fitting Model 1 and Model 2 separately using the valuation data from each of the NoCalc and Calc treatments.

Figure 8 shows the fit of Model 1 using Zhang's (2025) mirror valuation data, by juxtaposing the empirical distributions of responses with the distributions simulated from the estimated mixture model, separately in the two treatments. Visibly, Model 1 again provides a great fit for the empirical distributions, which shows that the three-type decomposition of mirror responses generalize to another data set collected using a different method.

Panel A of Figure 9 shows the cumulative distribution functions (CDF) of the RIV using data from Zhang (2025). In both the NoCalc and Calc treatment, the RIV is highly concentrated around the expected values of the lotteries – For all combinations of lottery and treatment, the RIV has more than 50% probability mass at exactly the expected value of the lottery. For comparison with the RIV in Panel A, Panel B of Figure 9 juxtaposes the CDFs of the responses in the NoCalc and Calc treatments. The RIV are much more similar between the NoCalc treatment than responses are. More specifically, for the responses, the Calc treatment has much larger probability masses at the expected value of the lottery than the NoCalc treatment. In contrast, for the RIV, the two treatments have more similar probability masses at the expected value.

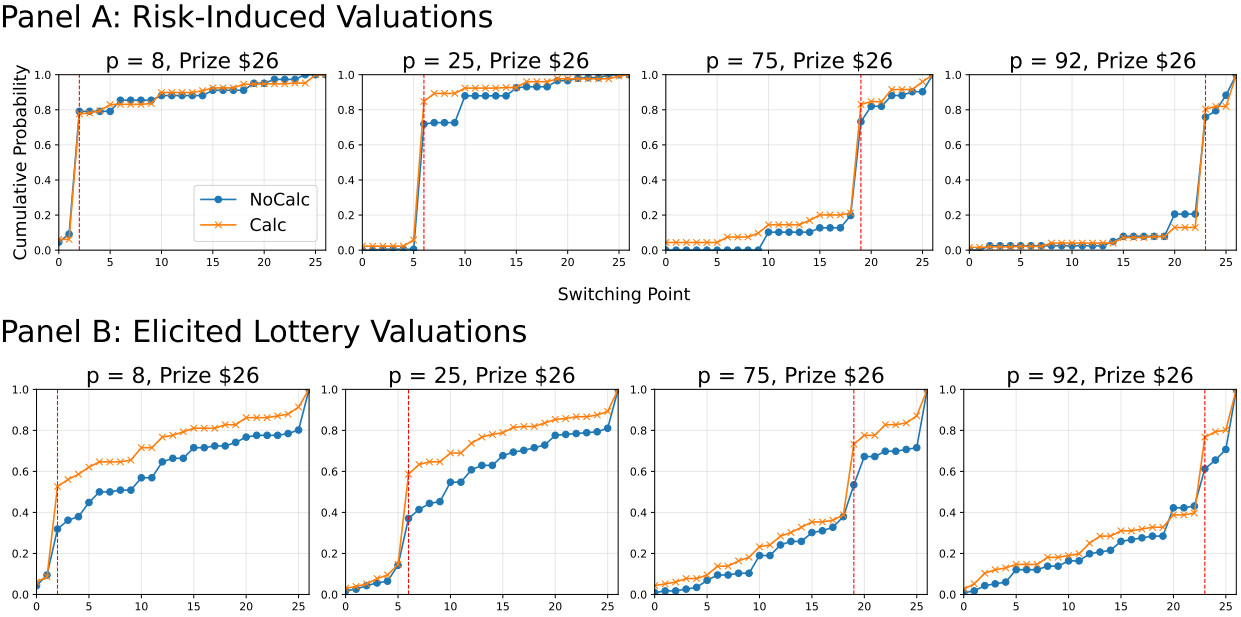

Figure 9: The cumulative distribution function of elicited lottery valuations and recovered risk-induced valuations in Zhang (2025)

Next, we test the prediction of similar RIV across treatments. Two hypotheses are tested: (1) equal distribution of RIV ($H_0^q : q^l_{NoCalc} = q^l_{Calc}$); and (2) equal distribution of responses ($H_0^f : f^l_{NoCalc} = f^l_{Calc}$). We employ a constrained-versus-unconstrained likelihood ratio test and generate the p-values through a subject-clustered bootstrap procedure. Details of this hypothesis testing procedure can be found in the Appendix. The test cannot reject the null hypothesis of equal distribution of RIV ($p = 0.92$), while it rejects the null hypothesis of equal distribution of responses ($p < 0.01$). In other words, although the empirical distributions of responses are significantly different between treatments, the RIV, which is recovered by correcting some mistakes in the responses, is statistically indistinguishable. This result provides strong support for the claim that RIV is a good approximation of the genuine valuations.

7 Conclusion

This paper contributes to our understanding of risk preferences by distinguishing between genuine preferences and mistakes in experimental responses. By correcting for mistakes that are common to both risky and risk-free decision contexts – using deterministic mirror tasks where correct valuations are known – we find that subjects' valuations are substantially closer to risk neutrality than raw responses suggest. When risk-induced valuations deviate from expected values, they predominantly exhibit risk aversion rather than the pronounced

risk-lovingness for low-probability gains documented in prior work. These findings suggest that the literature may have overestimated the extent to which risk preferences deviate from expected utility theory, and that estimates of prospect-theoretic parameters – particularly the strong overweighting of small probabilities – should not be interpreted as solely reflecting genuine preferences, but also reflect systematic mistakes in responding to preference elicitation tasks.

Beyond these substantive findings, this paper's primary contribution may be methodological. We develop a two-step procedure that reduces reliance on unverifiable parametric assumptions about stochastic choice by estimating the structure of mistakes in a related context where they are observable. The key innovation is the use of deterministic mirrors – objects with known valuations that share important features with the target decision problem. In our setting, both lotteries and their mirrors require processing the same disaggregated visual information and navigating the same price list mechanism, allowing us to identify mistake patterns that plausibly carry over to lottery tasks. By estimating a mixture model that characterizes these patterns, we obtain an empirically grounded correction for lottery responses that avoids arbitrary functional form assumptions. The validation evidence from Zhang (2025) – where recovered risk-induced valuations remain stable across treatments that alter mistake rates – suggests the method successfully isolates preferences from treatment-specific errors.

This approach may prove useful beyond the study of risk preferences. Whenever researchers seek to recover preferences from observed choices, they face the challenge that choices reflect both preferences and mistakes. Our method suggests a path forward: identify an auxiliary decision problem – a "mirror" – that shares key features with the target decision context, particularly those features likely to generate mistakes. The spirit of the mirror design is to match the cognitive demands of the original task; if disaggregation is a likely source of mistakes, the mirror should involve similar disaggregation; if the incentivizing mechanism, like price lists or the BDM mechanism, is likely to cause confusion, the mirror should be elicited with the same mechanism. Researchers can then pursue one of two strategies. If mistakes in the mirror setting are negligible, cancel out on average, or if the statistical method employed by the researchers correctly recovers preferences despite mistakes, the mirror serves as a validation device without requiring explicit correction. However, if mistakes in the mirror would lead to erroneous conclusions about preferences, then a correction procedure similar to our two-step approach becomes necessary: estimate the mistake structure in the mirror setting, then apply it to correct choices in the target domain. Researchers adopting this correction approach must be explicit about what their corrected measure captures and what it does not. In our case, the risk-induced valuations represent the treatment effect of risk on responses – capturing genuine risk preferences plus any risk-specific mistakes not present in the deterministic mirror

tasks – but cannot be interpreted as genuine preferences without additional assumptions. Any application of this methodology requires similar careful delineation of what has been corrected for (mistakes common to both contexts) and what remains (domain-specific mistakes), along with transparent discussion of the assumptions required for the correction to recover genuine preferences. While substantial work remains to understand when and how mistake structures transfer across contexts, this paper demonstrates that empirically disciplined approaches to correcting for mistakes are feasible and can materially change our conclusions about preferences.

References

Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer. 2021. "Time Will Tell: Recovering Preferences When Choices Are Noisy." *Journal of Political Economy*, 129(6): 1828–1877.

Andersen, Steffen, Glenn W. Harrison, Morten Igel Lau, and E. Elisabet Rutström. 2006. "Elicitation Using Multiple Price List Formats." *Experimental Economics*, 9(4): 383–405.

Balcombe, Kelvin, and Iain Fraser. 2025. "Flexible Estimation of Parametric Prospect Models Using Hierarchical Bayesian Methods." *Experimental Economics*, 1–25.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum. 2013. "The Nature of Risk Preferences: Evidence from Insurance Choices." *American Economic Review*, 103(6): 2499–2529.

Beauchamp, Jonathan P., Daniel J. Benjamin, David I. Laibson, and Christopher F. Chabris. 2020. "Measuring and Controlling for the Compromise Effect When Estimating Risk Preference Parameters." *Experimental Economics*, 23(4): 1069–1099.

Becker, Gordon M., Morris H. Degroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science*, 9(3): 226–232.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *The Quarterly Journal of Economics*, 127(3): 1243–1285.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica*, 78(4): 1375–1412.

Buschena, David, and David Zilberman. 2000. "Generalized Expected Utility, Heteroscedastic Error, and Path Dependence in Risky Choice." *Journal of Risk and Uncertainty*, 20(1): 67–88.

Carbone, Enrica, and John D. Hey. 2000. "Which Error Story Is Best?" *Journal of Risk and Uncertainty*, 20(2): 161–176.

Cason, Timothy N., and Charles R. Plott. 2014. "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy*, 122(6): 1235–1270.

Clark, Torin K., and Daniel M. Merfeld. 2021. "Statistical Approaches to Identifying Lapses in Psychometric Response Data." *Psychonomic Bulletin & Review*, 28(5): 1433–1457.

Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. "Mixture Models of Choice under Risk." *Journal of Econometrics*, 162(1): 79–88.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review*, 112(9): 2851–2883.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38.

Enke, Benjamin, and Cassidy Shubatt. 2023. "Quantifying Lottery Choice Complexity." National Bureau of Economic Research w31677, Cambridge, MA.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty*." *The Quarterly Journal of Economics*, 138(4): 2021–2067.

Frydman, Cary, and Lawrence J Jin. 2021. "Efficient Coding and Risky Choice." *The Quarterly Journal of Economics*, 137(1): 161–213.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy*, 130(4): 956–990.

Handel, Benjamin, Jonathan Kolstad, Thomas Minten, and Johannes Spinnewijn. 2024. "The Socioeconomic Distribution of Choice Quality: Evidence from Health Insurance in the Netherlands." *American Economic Review: Insights*, 6(3): 395–412.

Harless, David W., and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica*, 62(6): 1251–1289.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning. Springer Series in Statistics*, New York, NY:Springer.

Hey, John D. 1995. "Experimental Investigations of Errors in Decision Making under Risk." *European Economic Review*, 39(3): 633–640.

Hey, John D. 2005. "Why We Should Not Be Silent About Noise." *Experimental Economics*, 8(4): 325–345.

Hey, John D., and Chris Orme. 1994. "Investigating Generalizations of Expected Utility Theory Using Experimental Data." *Econometrica*, 62(6): 1291–1326.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–1655.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263–291.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2021. "Cognitive Imprecision and Small-Stakes Risk Aversion." *The Review of Economic Studies*, 88(4): 1979–2013.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions*." *The Quarterly Journal of Economics*, 133(1): 237–293.

Loomes, Graham, and Robert Sugden. 1998. "Testing Different Stochastic Specifications of Risky Choice." *Economica*, 65(260): 581–598.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa. 2019. "Failures in Contingent Reasoning: The Role of Uncertainty." *American Economic Review*, 109(10): 3437–3474.

McFadden, Daniel L. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*. 105–142. New York:Academic Press.

McGranaghan, Christina, Kirby Nielsen, Ted O'Donoghue, Jason Somerville, and Charles D. Sprenger. 2024. "Distinguishing Common Ratio Preferences from Common Ratio Effects Using Paired Valuation Tasks." *American Economic Review*, 114(2): 307–347.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care*." *The Quarterly Journal of Economics*, 137(2): 679–727.

Nielsen, Kirby, and John Rehbeck. 2022. "When Choices Are Mistakes." *American Economic Review*, 112(7): 2237–2268.

O'Donoghue, Ted, and Jason Somerville. 2024. "Investigating Risk Preferences Using Experiments."

Oprea, Ryan. 2024. "Decisions under Risk Are Decisions under Complexity." *American Economic Review*, 114(12): 3789–3811.

Plott, Charles R, and Kathryn Zeiler. 2005. "The Willingness to Pay–Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review*, 95(3): 530–545.

Prins, Nicolaas. 2012. "The Psychometric Function: The Lapse Rate Revisited." *Journal of Vision*, 12(6): 25.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–1292.

Treutwein, Bernhard, and Hans Strasburger. 1999. "Fitting the Psychometric Function." *Perception & Psychophysics*, 61(1): 87–106.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, 5(4): 297–323.

Wakker, Peter P. 2025. "Relating Risky to Riskless Preferences, and Their Joint Irrationality: A Comment on Oprea (2024)."

Wichmann, Felix A., and N. Jeremy Hill. 2001*a*. "The Psychometric Function: I. Fitting, Sampling, and Goodness of Fit." *Perception & Psychophysics*, 63(8): 1293–1313.

Wichmann, Felix A., and N. Jeremy Hill. 2001*b*. "The Psychometric Function: II. Bootstrap-based Confidence Intervals and Sampling." *Perception & Psychophysics*, 63(8): 1314–1329.

Woodford, Michael. 2012. "Inattentive Valuation and Reference-Dependent Choice."

Wu, George. 2025. "There Is No Smoke with Mirrors When Instructions Are Clear."

Zhang, Hanyao. 2025. "Calculations Behind Lottery Valuations."