

University of California
Santa Barbara

Three Essays in Behavioral and Experimental Economics

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Economics

by

Jing Zhou

Committee in charge:

Professor Ryan Oprea, Chair
Professor Erik Eyster
Professor Sevgi Yuksel

June 2024

The Dissertation of Jing Zhou is approved.

Professor Erik Eyster

Professor Sevgi Yuksel

Professor Ryan Oprea, Committee Chair

June 2024

Three Essays in Behavioral and Experimental Economics

Copyright © 2024

by

Jing Zhou

To my parents and grandparents.

Acknowledgements

The completion of this dissertation would not have been possible without the patient support and guidance of my committee: Ryan Oprea, Erik Eyster, and Sevgi Yuksel. I am incredibly fortunate and grateful to be your student. You have taught me an indescribable amount about economics, research, and life during my graduate studies.

First and foremost, I would like to extend my deepest thanks to my advisor, Ryan Oprea, who has provided countless hours of insightful comments and advice on my work. His guidance has drastically improved my ability to design experiments and present results in a compelling way. His willingness and enthusiasm to help in both my academic and personal life have been invaluable. His passion for behavioral and experimental economics has profoundly influenced my research interests and directly led to the development of my own research agenda. I am indebted to him for several breakthroughs that allowed me to see my projects through to completion. I thank him for shaping my taste in what makes a “good” paper, and refining my writing style.

I would like to express my heartfelt gratitude to Sevgi Yuksel for her solid and insightful comments on my research and her unwavering support during my job market process. I also extend my thank to Erik Eyster, who has consistently provided constructive and comprehensive comments and suggestions for my research and career development from various perspectives. I deeply appreciate all the encouragement, inspiration, and support that I received from him.

I would also like to thank the faculty members at UC Santa Barbara, especially Daniel Martin, Ignacio Esponda, Javier Birchenall, Ted Bergstrom, Kelly Bedard, Cheng-Zhong Qin, Heather Royer, Gonzalo Vazquez-Bare, and Shelly Lundberg for their kindness and feedback. I am indebted to the behavioral and experimental economics community for their careful analysis of my work throughout my time at UCSB. Additionally, I am

thankful for my colleagues in the department and the members of my cohort for providing a supportive environment. I am particularly grateful for Xin Jiang, Caroline Zhang, Yixin Chen, Kite Liu, and Yizi Lin for their friendship, support and guidance; they have always been willing to listen and lend a hand whenever needed. I had the pleasure of collaborating with my coauthors, Xin Jiang, Menglong Guan, ChienHsun Lin, and Ravi Vora, and I would not be the researcher I am today without their help and patience.

Lastly, my biggest thank goes out to my family. During the years in pursuit of my degree, I have missed many important events, and my parents, Zhipin and Deyou shouldered the entire burden of the family. Your unconditional love and endless support have endowed me with the courage to go further, and I am forever grateful for your sacrifice and understanding.

Curriculum Vitæ

Jing Zhou

EDUCATION

- 2024 Ph.D. in Economics (Expected), University of California, Santa Barbara.
- 2018 M.S. in Econometrics and Quantitative Economics, University of Wisconsin, Madison.
- 2016 M.A. in Economics, Nanjing University.
- 2013 B.A. in Economics, Nanjing Normal University.

RESEARCH INTERESTS

Behavioral and Experimental Economics, Information Economics

Research PAPERS

1. What Drives Probability Matching?

A shorter version of my job market paper titled “Does Correlation Matter in Probability Matching? A Laboratory Investigation” is under Conditional Acceptance at *Journal of Economic Behavior and Organization*

2. Sample Features and Belief Updating: Preference and Performance (with Menglong Guan, ChienHsun Lin, and Ravi Vora)
3. Dynamic Binary Belief Elicitation Method (with Xin Jiang)
4. Industrial Upgrading: Rural-urban Migrants Training Costs and Heterogeneous Labor Migration. (with Xiaochun Li and Yang Ge), *Bulletin of Economic Research*, Vol.70, 2017 NOS 4: 483-491.
5. Environmental Effects of Remittance of Rural-urban Migrant. (with Xiaochun Li), *Economic Modelling*, 2015 47: 174-179.

PROFESSIONAL EXPERIENCE

Teaching Assistant

University of California Santa Barbara

- Game Theory (Ph.D Level) 2020
- Game Theory (Ph.D Level) 2019
- Monetary Economics (undergraduate) 2023
- Financial Management (undergraduate) 2022
- Intermediate Microeconomics (undergraduate) 2018 - 2023
- Intermediate Macroeconomics (undergraduate) 2018 - 2023
- Introduction to Economics (for non-econ major) 2020
- Principles of Macroeconomics (undergraduate) 2019

Teaching Assistant Nanjing University

- Mathematical Economics 2014

Research Assistant Summer, 2022
University of California Santa Barbara

- For Professor Ignacio Esponda and Professor Emanuel Vepsa

Summer Research Group Summer 2017
University of Wisconsin Madison

- Developed causal empirical models to estimate impacts of postgraduate education on gender difference in wage.

Research Assistant 2014-2016
Nanjing University

- For Professor Xiaochun Li

AWARDS AND SCHOLARSHIPS

University of California Santa Barbara

- Research Quarter Fellowship January, 2022
- Economics Department Graduate Research Grant for Experiments 2021, 2022
- Distinction in Ph.D. Preliminary Examination in Econometrics August, 2019

University of Wisconsin Madison

- Econ Masters Program Research Scholarship February, 2017

Nanjing University

- National Scholarship (for top 5% students)

November, 2015

Nanjing Normal University

- Outstanding Graduates Awards

June, 2013

TECHNICAL SKILLS

Stata, R, Python, Matlab, Mathematica, html, Javascript, oTree, Qualtrics, L^AT_EX, MS Office.

Abstract

Three Essays in Behavioral and Experimental Economics

by

Jing Zhou

This dissertation consists of three chapters that explore why individuals make seemingly suboptimal decisions in risk management, why they fail to use information in a Bayesian manner when updating their beliefs, and how novel methodological tools can be developed to advance modelling and inference about subjective beliefs or perceptions, considering cognitive limitations.

Chapter 1 studies the underlying mechanisms behind a classical behavioral puzzle in risk management, called Probability Matching. Probability matching refers to people's tendency to randomize between different risky options, or even match their choice frequency to the outcome probability, when choosing over binary lotteries that differ only in their probabilities. Why? I present an experiment designed to distinguish between three broad classes of explanations: models of *Correlation-Invariant Stochastic Choice* (mixing due to factors orthogonal to how outcomes are jointly determined, such as non-standard preferences or errors), models of *Correlation-Sensitive Stochastic Choice* (e.g., deliberately mixing due to misperceived hedging opportunity), and *Framing Effects* (indecisiveness due to frame-sensitive heuristics e.g., similarity heuristic: attending to dissimilar but irrelevant attributes (outcomes), while ignoring relevant attributes (probabilities)). My experimental design uses a diagnostic approach, differentiating between their testable predictions over a series of treatments. The results suggest that a substantial proportion of mixing behavior aligns with models of *Correlation-Sensitive Stochastic Choice*, while the other classes have limited explanatory power.

In Chapter 2, a joint work with Menglong Guan, ChienHsun Lin, and Ravi Vora, we experimentally investigate how people value and utilize different statistical characteristics of a set of realized binary signals, referred to as sample features, to understand why individuals deviate from the Bayesian benchmark when updating beliefs. We find that, subjects systematically under-infer the information contained in each sample feature. Furthermore, the magnitude of under-inference significantly varies across sample features. Specifically, under-inference is least severe with Sample Proportion (the relative frequency of different outcomes in the realized signals), compared to more informative features such as Sample Count (the absolute number of different outcomes in the realized signals). We also find that the standard measure of informativeness used in information theory does not fully explain subjects' preferences for sample features. Subjects demonstrate a strict preference for the information contained in the Sample Proportion over those without it and undervalue the usefulness of sample size. Combining preference and belief updating behaviors, we find that subjects deviate less from the Bayesian benchmark when provided with a more-preferred feature than a less-preferred one. These results suggest that some biases in signal usage is more likely an intentional deviation rather than a result of inattentive heuristics.

In Chapter 3, a joint work with Xin Jiang, we introduce a novel elicitation method, called the *Dynamic Binary Method* (DBM), designed to address the common challenge individuals face in pinpointing the best point estimate of their beliefs, particularly when their beliefs are imprecise. Unlike Classical Methods (CM), which require respondents to make absolute judgments and form a point estimate of their true beliefs, DBM guides them through a series of binary relative judgments, enabling them to express interval beliefs by exiting the process at any step. To assess the empirical validity of DBM, we conduct both within-subject and between-subject experiments using a diverse range of perception tasks drawn from previous literature and CM as a benchmark of performances

in each task. We find that DBM does not perform significantly differently from CM at the aggregate level, regardless of whether the perception questions use artificial/laboratory settings or real-life settings, and irrespective of the measurement used. Notably, DBM outperforms CM when the objective truth is extreme. Furthermore, we find a negative correlation between the length of stated beliefs in tasks using DBM and their accuracy. Additionally, we find that the length stated in DBM can predict respondents' performance in CM tasks at the aggregate level, albeit not strictly in a monotonic manner. Finally, we explore methods to use DBM-collected data for predicting stated point beliefs in DBM, offering insights into potential applications of the method beyond its immediate implementation.

JEL Classification: D81, D91, C91

Contents

Curriculum Vitae	vii
Abstract	x
1 What Drives Probability Matching?	1
1.1 Introduction	1
1.2 Theoretical Foundations	12
1.3 Experimental Design	23
1.4 Results	28
1.5 Discussion	47
1.6 Conclusion	52
2 Preference for Sample Features and Belief Updating	55
2.1 Introduction	55
2.2 Experimental Design	63
2.3 Theoretical Predictions	72
2.4 Results	80
2.5 Conclusion	96
3 Dynamic Binary Method	99
3.1 Introduction	99
3.2 Theoretical Framework	104
3.3 Experimental Design	109
3.4 Results	115
3.5 Conclusion	129
A Appendix for “What Drives Probability Matching?”	131
A.1 Theories of Mixing	131
A.2 First-Order Stochastic Dominance	139
A.3 Parameters used in Experiments	141
A.4 The Magnitude of Mixing	142
A.5 Block 3: Transfer of Learning	143

A.6	Additional Results	145
A.7	Experimental Instructions	156
B	Appendix for “Preference for Sample Features and Belief Updating”	175
B.1	Incentive Compatibility of the Ranking-Card Method	175
B.2	Experimental Design Details	177
B.3	Grether Model + Full Data	178
B.4	Report-Whatever-You-See Heuristics	180
C	Appendix for “Dynamic Binary Method”	182
C.1	Binarized Scoring Rule and Incentive Compatibility	182
C.2	BDM with Myopic DM	185
C.3	More Results	187
C.4	Questions used in Experiments 1 and 2	188

Chapter 1

What Drives Probability Matching?

1.1 Introduction

Probability matching (PM), as a classical behavioral puzzle in risk management, refers to people's tendency to randomize between different risky options, or even match their choice frequency to the outcome probability, when choosing over binary lotteries that differ only in their probabilities. For example, suppose a project manager needs to decide which candidate project to implement, Project A or Project B, and can use a probabilistic choice strategy. Project A and Project B will succeed 75% and 25% of the time, respectively. The manager will receive a fixed bonus ($\$M$) if the implemented project is successful; otherwise, nothing. To maximize the likelihood of success, it is optimal to choose Project A with certainty, since A first-order stochastically dominates B.¹ However, there is significant empirical evidence demonstrating that the majority of people tend to mix by choosing each alternative with a positive chance, or even match their choice distributions to the outcome probabilities by choosing A 75% and B 25%

¹Definition of first-order stochastic dominance (FOSD): Option A is first-order stochastically dominant over Option B if $\forall x \in \mathbb{R}, Pr(A \geq x) \geq Pr(B \geq x)$, and $\exists x \in \mathbb{R}, Pr(A \geq x) > Pr(B \geq x)$.

of the time. Such behavior lowers their chance of success than the maximum (75%).² This distributional behavior is not only documented in many laboratory environments, including those with no value for exploration (e.g., when there is no feedback), and those with no portfolio effects (e.g., when only one choice is paid), but also observed in important life decisions such as university admissions (Dwenger, Kübler and Weizsäcker, 2018), stock price predictions (Kallir and Sonsino, 2009), adherence to the long-term therapy for chronic disease (AlHewiti, 2014), etc.³

Given its prevalence, this finding has long puzzled economists and psychologists and led to the development of many theoretical explanations aiming to account for this behavior. However, the empirical evidence remains limited, as few explanations are tested in isolated studies with mixed results (Literature are discussed below). Understanding the sources of this mixing behavior is fundamental to improving our understanding on how individuals make decisions and developing informed risk management strategies and consumer protection policies.

In this paper, I use a series of diagnostic laboratory experiments to study the origin of this mixing behavior, and, more specifically, to distinguish between three broad classes of explanations. To illustrate, let's revisit the previous example of project management using the commonly-used payoff structure, which I will call as the Classical Payoff Structure (CPS), as shown in Table 1.1. Each project's outcome is determined by one of the four

²It is supported by the empirical evidence that when repeatedly facing the same binary choice multiple times, individuals tend to choose each alternative in a way that replicates their preferred choice distribution over them (Feldman and Rehbeck, 2022).

³For laboratory evidence, see Martínez-Marquina, Niederle and Vespa (2019); Rubinstein (2002); Vulkan (2000). For empirical evidence, Dwenger, Kübler and Weizsäcker (2018) analyze data from the centralized clearinghouse for university admissions in Germany, which requires students to submit multiple rankings of universities; these rankings are submitted at the same time, and only a randomly chosen one matters. They find that many students report inconsistent rankings, which reduces their probabilities of getting into a more desirable university, even when there are no strategic reasons to do so. Kallir and Sonsino (2009) find most financial analysts fail to maximize their prediction accuracy due to this distributional behavioral pattern. Similarly, AlHewiti (2014) find that patients with higher education report lower adherence rate as they are concerned about the sided effects of medications which occurs with small chance.

equally likely states of the world, ω_i , where $i \in \{1, 2, 3, 4\}$. Correlation between options is determined by the joint distribution of the outcomes across states. Payoff framing represents the way outcomes and probabilities of alternatives are presented in the payoff table/matrix. In the CPS, the options are perfectly negatively correlated, meaning that when one option yields a good outcome, the other yields a bad one, and vice versa, and the correlation structure is explicitly represented in the payoff table/matrix, as shown in Table 1.1. The way outcomes and alternatives are presented in the CPS are referred to as the Classical Frame.

The first class of explanation described in the literature, which encompasses most preference-based and some heuristics-based models, argue that people mix due to non-Expected Utility preferences such as gaining extra utility from mixing itself (Allen and Rehbeck, 2023; Fudenberg, Iijima and Strzalecki, 2015), inattentive heuristics such as trembling hand (Ratcliff, 1978), inherent biases such as misperceived probability (Agronov, Healy and Nielsen, 2023), etc. One common feature shared by these explanations is that the sources of mixing are orthogonal to both the correlation structure and the way outcomes and alternatives are presented in the payoff framing. These models I refer to as models of *Correlation-Invariant Stochastic Choice*.⁴

The second class of explanation suggests that individuals deliberately mix due to heuristics that are sensitive to how outcomes are jointly determined. For instance, individuals might mistakenly consider that mixing between options can hedge against the risk of Project A failing when ω_4 gets realized, as Project B outperforms Project A in ω_4 . One potential reason could be that individuals may hold incorrect belief that it is a portfolio choice in which the decision maker is making multiple bets and gets paid for all of them, instead of a probabilistic choice strategy in which they choose the likelihood of

⁴As most payoff tables/matrices explicitly present the correlation structure, when the correlation structure changes, both the underlying correlation between options and the way outcomes and alternatives are presented in the payoff tables/matrices vary.

each option and get paid for the single realized option (false diversification, Rubinstein (2002)). Alternatively, this misperception may stem from an aversion to making big *ex post* mistakes if choosing A with certainty and ω_4 gets realized (min-max regret with convex cost of mistakes, Agranov, Healy and Nielsen (2023)). These models are referred to as models of *Correlation-Sensitive Stochastic Choice*.

The third class of explanations suggests that individuals mix as they do not know which option is optimal due to the frame-sensitive heuristics they use to simplify the comparison between marginal distributions. For instance, individuals may compare options based on the similarities between options: when facing the payoff framing as shown in Table 1.1, individuals attend to the dissimilar but irrelevant attributes – outcome differences, while neglecting the relevant attributes for EU maximization – probability differences. They may naively make column-wise comparisons: ignore columns with similar outcomes and use columns with dissimilar outcomes to find the optimal choice, irrespective of whether the outcomes in each column are indeed correlated. As such comparisons in Table 1.1 disagree on which option is optimal, individuals resolve it by mixing between the options (Leland, 1998; Rubinstein, 1988).⁵ I call this class of explanation *Framing Effects*.

To distinguish between the three classes of explanations, I conduct an experiment with three between-subject treatments: Baseline, Independence, and Unknown. Block 1 of each treatment captures the main treatment variations and Block 2 is a repetition of Block 1 enabling study of learning effect. The basic decision problem is the ticket-allocation task in Martínez-Marquina, Niederle and Vespa (2019). To elicit choice distribution, subjects are asked to predict which of the two payoff-relevant outcomes will be realized by allocating some tickets; one randomly selected ticket gets paid.

⁵Similarity heuristic has been established as a competing explanation against the sensitivity to correlation structure, for violations of FOSD in one-shot binary decision (Dertwinkel-Kalt and Köster, 2015; Leland, 1998; Leland, Schneider and Wilcox, 2019; Tversky and Kahneman, 1986).

Table 1.1: Classical Payoff Structure(CPS): Perfectly Negative Correlation + Classical Frame

	25%	25%	25%	25%
State	ω_1	ω_2	ω_3	ω_4
Project A	S	S	S	F
Project B	F	F	F	<u>S</u>

Table 1.2: Alternative Payoff Structure(APS): Positive Correlation + Alternative Frame

	25%	25%	25%	25%
State	ω_1	ω_2	ω_3	ω_4
Project A	S	S	S	F
Project B	<u>S</u>	F	F	F

Table 1.3: Classical Frame + Zero Correlation

	25%	25%	25%	25%
	<i>Partition 1</i>	<i>Partition 2</i>	<i>Partition 3</i>	<i>Partition 4</i>
Project A	S	S	S	F
Project B	F	F	F	<u>S</u>

Table 1.4: Alternative Frame + Zero Correlation

	25%	25%	25%	25%
	<i>Partition 1</i>	<i>Partition 2</i>	<i>Partition 3</i>	<i>Partition 4</i>
Project A	S	S	S	F
Project B	<u>S</u>	F	F	F

Notes: In each payoff table, Project A and Project B have 75% and 25% of chance to succeed, respectively. The project manager in question forms a choice distribution over them to implement and will receive a fixed monetary award ($\$M$) if success – otherwise, nothing. The underlined entries denote the main distinctions – different locations of outcomes – among the four payoff tables. In Tables 1.1 and 1.2, each column represents one state of the world, and outcomes within the same column will be jointly realized. That is, the correlation structure is explicitly presented in the payoff framing. Table 1.1 is widely used in previous literature. In this structure, outcomes are perfectly negatively correlated: either one project succeeds in each state. In Table 1.2, the two projects’ outcomes are positively correlated: jointly succeed or fail in most of the states. In Tables 1.3 and 1.4, two projects’ outcomes are independently determined and the correlation structure is not presented in the payoff framing: they are presented in a similar way as Tables 1.1 and 1.2, respectively. In these structures, the columns, denoted as partitions, do not necessarily represent the states of the world. Thus, the outcomes in the same column will not necessarily co-occur.

Firstly, to distinguish the models of *Correlation-Invariant Stochastic Choice* from the other two, in the Baseline treatment, I start with the task using the CPS, as shown in Table 1.1, and then construct a decision problem using the Alternative Payoff Structure (APS), as shown in Table 1.2. In the APS, the two options exhibit positive correlation — two options jointly have good or bad outcomes in most states, which is presented in the payoff table/matrix. The way outcomes and alternatives are presented in the APS is denoted as the Alternative Frame. I construct a series of decision problems to vary the correlation structure between the two payoff structures in a comprehensive manner. Models of *Correlation-Invariant Stochastic Choice* predict identical choices across all tasks in the Baseline, while the other two classes predict that subjects will mix between options whenever the task does not have the APS, and will choose the dominant option with certainty in the tasks APS. This is either because there is no way to “hedge” against the risk of Project A failing, or because the Alternative Frame in the APS highlights differences in the relevant attributes – probability, while downplaying differences in the irrelevant attributes – outcome. That is, even with similarity heuristic, individuals would find Project A optimal and choose it with certainty.

To further distinguish between models of *Correlation-Sensitive Stochastic Choice* and *Framing Effects*, in the Independence treatment, I fix the correlation between options at zero by letting the outcome of each option be independently determined. Meanwhile, I comprehensively vary the payoff framing across tasks in the same way as in the Baseline, from the Classical Frame as in Table 1.3 to the Alternative Frame as in Table 1.4. Models of *Correlation-Sensitive Stochastic Choice* predict that subjects will mix between options in all the tasks in the Independence treatment, since it is still likely that Project B outperforms Project A. However, *Framing Effects* predicts that subjects in the Independence treatment will behave the same way as the Baseline — they will mix between options when not presented with the Alternative Frame, and will choose the dominant

option with certainty in the Alternative Frame. To benchmark the magnitude of *Framing Effects*, I employ the Unknown treatment, in which the correlation information is not provided but all the tasks are presented with the Alternative Frame. Models of *Correlation-Sensitive Stochastic Choice* predict that subjects will mix between options in all the tasks, if they believe that each possible joint distribution is equally likely to occur, while *Framing Effects* predict no mixing here.

The results demonstrate that subjects deliberately consider the correlation between options when making decisions, which accounts for a substantial proportion of mixing behavior. First, the findings reject expected-utility maximization and many behavioral theories that predict no mixing in this environment, as 65% of choices mix between options in tasks using the CPS.⁶ Second, aggregate results across the three treatments show that subjects' choices respond to changes in the correlation between options in a manner consistent with models of *Correlation-Sensitive Stochastic Choice*. With framing effects controlled, subjects are, on average, 16.5% less likely to mix between options, and 10.8% less likely to match exactly to the outcome probability when the correlation between options increases. Moreover, results from the Independence treatment show that subjects deliberately take zero correlation between options into account: once the correlation between options is fixed, subjects' choices do not vary with the payoff framing. Combining the Independence and Unknown treatments, then, the estimated magnitude of framing effects is not significantly different from zero. In the Independence treatment, subjects are slightly more likely to mix compared to the Baseline, even with the correlation structure and payoff framing controlled. It suggests that subjects may misinterpret zero correlation when it is described in words in the Independence, in contrast to the

⁶For example, prospect theory (Kahneman, 1979), cumulative prospect theory (Tversky and Kahneman, 1992), rank-dependent expected utility (Quiggin, 1982), quadratic utility (Chew, Epstein and Segal, 1991), cautious expected utility (Cerrei-Vioglio, Dillenberger and Ortoleva, 2015), random utility (Gul and Pesendorfer, 2006), deliberate randomization (Cerrei-Vioglio et al., 2019), and recursive expected utility (Kreps and Porteus, 1978).

Baseline, where the joint distribution of zero correlation is presented in reduced form. More discussions can be found in Section 1.5. Exploring whether this result is due to misinterpretation or other confounds would be a fruitful direction for future research. Lastly, although learning has limited impacts on reducing mixing in the CPS, as they gain experience, subjects tend to be much less likely to mix or match exactly to the likelihood of occurrence when the correlation increases.

Classifying subjects based on their choices gives similar results: the vast majority of subjects (65%) mix between options in some tasks, while choosing the dominant option with certainty (allocate all tickets on the dominant option) in others in the Baseline, as most of them (73%) respond to changes in the correlation between options. There is some heterogeneity in the Baseline: a small proportion of subjects are consistent with the expected utility benchmark (17.5%), while an equal fraction align with the models of *Correlation-Invariant Stochastic Choice*. Overall, I find that the majority of subjects make decisions consistent with models of *Correlation-Sensitive Stochastic Choice* — they mix to hedge against (misperceived) risk — and the other two classes of explanations have limited powers in explaining mixing behaviors.

These findings have important theoretical and empirical implications. From a theoretical perspective, it is essential to develop frameworks that incorporate individuals' consideration of the correlation between options in stochastic settings. Neoclassical theories and most stochastic choice theories fall short in explaining the primary finding of this paper: the vast majority of subjects make different choices in response to changes in the correlation between the options featured in this study. On the one hand, most of these theories approach stochastic choice from the perspective of the analyst or econometrician, assuming that the decision-maker does not see their choice as random. In such models, randomness stems from exogenous and random shocks on preferences, attentions, and so forth. Recently, a small but growing subset of theoretical studies has started to consider

the possibility that individuals opt for stochastic choice due to non-standard preferences or trembling hand. These factors are also orthogonal to how options are correlated. In Section 1.5, I also investigate every heuristics in the models of *Correlation-Sensitive Stochastic Choice* in details and find that each of them has their own limitations and none of them can accommodate all the results.

Empirically, my findings also shed light on why individuals often deviate from utility maximizing choices especially in repeated economic decisions such as buying insurance, making medication decisions, to name a few. Given that individuals deliberately take into account the correlation between options, even though neoclassical theories suggest otherwise, it is crucial for policymakers to carefully explain how different options such as insurance contracts, or saving plans, are correlated in each circumstance. This is especially important for options whose risks are not perfectly negatively correlated in between. It helps individuals to always opt for the better option in the repeated choice environment. Educating on why the correlation does not matter might also be a good way to improve decision making.

Relation to the Literature and Contributions This paper contributes to the existing literature in several important ways. Firstly, to the best of my knowledge, this study is the first to directly test the predictions of three broad classes of theories that can explain stochastic choice over binary lotteries that differ only in their probabilities: models of *Correlation-Invariant Stochastic Choice*, models of *Correlation-Sensitive Stochastic Choice*, and *Framing Effects*, and provides direct evidence supporting models of *Correlation-Sensitive Stochastic Choice* – people deliberately mix to hedge against (misperceived) risk. Existing studies investigate few explanations in separated works and find inconclusive results. Most psychology literature focuses on investigating the effectiveness of different interventions in supporting one of the two arguments embedded in the dual-process theory – whether PM is an inattentive mistake as the outcome of the

fast and intuitive process, or a sophisticated strategy as the consequence of the slow and deliberate process, and are inconclusive on which one dominates (See Koehler and James (2014) for a review). Some economic studies, on the other hand, explain PM from the perspective of failure in contingent reasoning. For example, Martínez-Marquina, Niederle and Vespa (2019) find that the role of uncertainty can explain 8.7% of mixing between options, whose magnitude is smaller than what I observe. Agranov, Healy and Nielsen (2023) also examine the role of the failure in contingent reasoning in explaining PM with several interventions and find mixed results.⁷ As most studies are conducted within the CPS, my finding suggests a potential reason behind why these interventions have inconclusive or limited results: subjects' responsiveness to the correlation structure could be strong enough to mitigate the effectiveness of interventions.

Second, I contribute to empirical studies on stochastic choice in several ways. In terms of theoretical discussions, Agranov, Healy and Nielsen (2023) extensively explore various stochastic choice models to assess their abilities to explain the prevalence of mixing behavior across different domains observed in their study, with a particular emphasis on explanations of PM. Continuing this line of inquiry, I further categorize existing models, including those discussed in their study, into three distinct groups, and directly examine the validity of each class based on different testable predictions. From an empirical perspective, I provide direct evidence that the consideration of correlation structure, e.g., misperceived hedging opportunity, is a source of deliberate randomization. The results are empirically consistent with the model of min-max regret with convex costs of mistakes proposed by Agranov, Healy and Nielsen (2023), and false diversification (Rubinstein, 2002), while casting doubts on other stochastic choice theories.

⁷The intervention — changing the way questions are repeatedly asked and adding feedback on whether choice in last round is selected for payment, from subjects being asked to make binary choice that is repeated simultaneously on the same page without feedback, to subjects being asked to make binary choice that is repeated sequentially with feedback — works, while changing the way outcomes are realized, from one realization to multiple i.i.d. draws, does not.

Furthermore, my study contributes to the emerging literature on how individuals take into account correlations when making economic decisions by extending this investigation into this simple stochastic choice environment. Existing literature investigates this in the environments such as portfolio choices (Eyster and Weizsäcker, 2011), information structure (Hossain and Okui, 2020), where the correlation plays an essential role for optimal decision making, and find that individuals neglect the correlation structure, treating them as if there is no correlation in between when making decisions. Recent studies on how individuals choose between risky lotteries in one-shot binary choice environment, find that subjects are sensitive to the correlation structure, which is consistent with correlation-sensitive preference with salience theory and regret theory nested (Frydman and Mormann, 2018; Loewenfeld and Zheng, 2021). In the decision-making environment I consider, where most decision theories including correlation-sensitive preference posit that the correlation does not matter for optimal decision making, I find that individuals are sensitive to how outcomes are correlated. In addition, the finding that the decision to mix and the decision to match exactly with the probability reflect varying degrees of sensitivity to marginal changes in correlation suggests that responsiveness to correlation may differ across decision-making contexts.

Finally, I contribute to existing studies on why individuals violate FOSD by grounding this inquiry in a stochastic choice environment and examining the validity of a particular *Framing Effects* – the similarity heuristic (Rubinstein, 1988). It has been well established in the previous literature that certain ways that the alternatives and outcomes are presented mask the dominance relation, which leads to violations of FOSD (Dertwinkel-Kalt and Köster, 2015; Tversky and Kahneman, 1986). Thus, altering the frame to emphasize the dominance relation is effective in reducing violation of FOSD in the one-shot binary choice environment (See Kouroukous and Bauer (2019) for a review). However, the observations in the Independence treatment suggest that the effectiveness of the *Framing*

Effects is limited in its ability to reduce mixing behavior. My findings suggest that the effectiveness of changing the framing as an intervention to reduce the violation of FOSD may not be robust to choice environments.

The rest of the paper is organized as follows: Section 1.2 discusses the theoretical foundations. Section 1.3 presents the experimental design. Section 1.4 analyzes the results, followed by discussions and conclusions in Sections 1.5 and 1.6, respectively.

1.2 Theoretical Foundations

In this section, I first describe the basic setup and conceptual framework. Then, I explore three classes of models and show their distinguishable predictions of whether mixing behavior varies with changes in correlation between the options and payoff framing. For a more comprehensive discussion of all the example models mentioned, please refer to Appendix A.1.

1.2.1 Basic Setup

Consider a generalized version of the example discussed earlier: a decision problem involving two lotteries that differ only in their probabilities, namely Option A and Option B. Each option gives either a fixed monetary reward of $\$M$, or $\$0$. I denote the option pair as $(A : p; B : 1-p)$, where p (or $1-p$) is the likelihood of Option A (or Option B) yielding $\$M$. The key distinction between these two options is the likelihood of receiving $\$M$. For the sake of simplicity, assume that $p > \frac{1}{2}$. I refer to the option with a higher likelihood of receiving $\$M$, henceforth Option A, as the dominant option. Conversely, the option with the lower chance of obtaining $\$M$, i.e., Option B, is called the dominated option, given that the Option A FOSD Option B. Each option's outcome is determined by one of the four states of the world, denoted as $\omega_i \in \{1, 2, 3, 4\}$, wherein each state is equally likely to

be realized. Consequently, there are four possible joint outcome realizations, represented as $(x, y) \in \{(A : \$M, B : \$M), (A : \$M, B : \$0), (A : \$0, B : \$M), (A : \$0, B : \$0)\}$.

I am interested in the choice distribution formed by the decision maker: a map that associates a probability measure over the option pair $(A : p; B : 1 - p)$. This map represents the frequency at which the decision maker chooses each option, either shown as the choice pattern when repeatedly asked to choose between them multiple times or the probabilistic distribution formed from a linear convex lottery budget with one randomly selected choice getting paid.⁸ Let α and $1 - \alpha$ be the probabilities of the decision maker opting for Option A and Option B, respectively. I define a choice distribution as *mixed* if it assigns a positive probability to both options (i.e., $\alpha \in (0, 1)$). On the other hand, a choice distribution is referred to as *exact PM* if it perfectly aligns with $(A : p; B : 1 - p)$, i.e., $\alpha = p$. The use of the term "mixing behavior" encompasses both *mixed* choices and *exact PM* choices.

1.2.2 Conceptual Framework

As the primary differentiating features of my experimental design, I manipulate the correlation between options and frame in the payoff structure separately, to distinguish three classes of explanations. To be more specific, I group the decision problems based on whether the correlation structure and the payoff framing in these problems are varied separately into three scenarios to explain how the theoretical mechanisms differ.⁹ Each scenario has the same number of decision questions. Questions in all the scenarios share the same marginal distributions, $(A : p; B : 1 - p)$, but are distinct from one another

⁸Feldman and Rehbeck (2022) find empirical evidence that individuals' preference to choose a non-degenerate mixture of two different risky options from a linear convex lottery budget is positively related to their choice pattern in repeated discrete choices.

⁹To vary payoff framing, I separate the marginal distributions into the same number of equally probabilistic partitions, and then rank the payoff outcomes from the highest to the lowest. I prove in Appendix A.2 that any pair of FOSD options can be presented in the Alternative Frame, referred to as the "Transparency Frame" in Leland, Schneider and Wilcox (2019).

in terms of whether the correlation structure and payoff frames vary across the decision problems within each scenario:

Baseline Scenario: Varied Correlations + Varied Frames For the decision problems in this scenario, the correlation between options varies comprehensively, and so does the frame, as the frame demonstrates the underlying correlation. Decision questions with the CPS and those with the APS, as shown in Table 1.1 and Table 1.2, respectively, are example questions typifying this scenario.

Independence Scenario: Fixed Correlation + Varied Frames For the decision questions in this scenario, the correlation between options is fixed at zero, yet the frame varies. Decision problems with the Classical Frame + Zero Correlation and the Alternative Frame + Zero Correlation, as shown in Table 1.3 and in Table 1.4, respectively, are example questions adopted in this scenario.

Unknown Scenario: Fixed Correlation + Fixed Frame In this scenario, the decision maker faces various decision questions where the correlation between options is unknown to them. The frame is fixed as the Alternative Frame. If the decision-maker believes that each possible correlation structure is equally likely to occur, they would believe that the *ex ante* correlation between options is zero.¹⁰

The main interest of this study is how existing models differ in predicting the decision on whether to mix, i.e., whether $\alpha = 100\%$ or not, in decision problems across the three scenarios.¹¹

Expected Utility Benchmark Any models that respect FOSD and compound lottery reduction predict that the optimal choice is to choose the dominant option with $\alpha =$

¹⁰In the context of the prevailing example, this means that the decision maker believes that there is a 25% chance that the correlation between options is -1, and a 75% chance that it is 1/3. As a result, the expected correlation is 0.

¹¹This is because each theory's prediction of $\alpha \in (0\%, 100\%)$ varies with the functional forms such as the cost function in the min-max regret with convex cost of mistakes (Agranov, Healy and Nielsen, 2023), behavioral parameters such as risk preference in the false diversification (Rubinstein, 2002), and the parameters used in the experimental design.

100%, in all the decision problems across the three scenarios. Mixing between Options A and B generates a two-stage lottery, which can be reduced to a simple lottery in the simplex, denoted as \mathcal{L} :

$$\begin{aligned}\mathcal{L} &= \alpha \circ \textit{Option A} \oplus (1 - \alpha) \circ \textit{Option B} \\ &= [\alpha * p + (1 - \alpha) * (1 - p)] \circ \$M \oplus [\alpha * (1 - p) + (1 - \alpha) * p] \circ \$0\end{aligned}\tag{1.1}$$

where the first equation denotes the lottery in the first stage, and the second equation represents the reduced lottery over $\$M$ and $\$0$ with a corresponding probability of $[\alpha * p + (1 - \alpha) * (1 - p)]$ on $\$M$ and $[\alpha * (1 - p) + (1 - \alpha) * p]$ on $\$0$, respectively. In each decision problem, Option A FOSD Option B. Moreover, in decision problems with the APS, Option A also state-wise dominates Option B. That is, Option A is not only distribution-wise, but also state-wise, more likely to yield the better outcome $\$M$, than Option B. Given that Option A FOSD Option B in each payoff structure, if Option A is strictly preferred to Option B, it is also strictly preferred to any mixture between the two. Examples of models satisfying these two include the expected utility, prospect theory (Kahneman, 1979), cumulative prospect theory (Tversky and Kahneman, 1992), rank-dependent expected utility (Quiggin, 1982), quadratic utility (Chew, Epstein and Segal, 1991), cautious expected utility (Cerrei-Vioglio, Dillenberger and Ortoleva, 2015), random utility (Gul and Pesendorfer, 2006), deliberate randomization (Cerrei-Vioglio et al., 2019), and recursive expected utility (Kreps and Porteus, 1978).

Hypothesis 1. *Individuals who follow FOSD and compound lottery reduction will choose the dominant option with $\alpha = 100\%$ in all the decision problems across three scenarios.*

1.2.3 Models of Stochastic Choice

In this subsection, I focus on the preference-based and heuristics-based models that allow for violation of FOSD or violation of compound lottery reduction. Existing models can be divided into three categories: models of *Correlation-Invariant Stochastic Choice*, models of *Correlation-Sensitive Stochastic Choice*, and *Framing Effects*.¹² While each category includes many models, they have many features in common. Within each category/subcategory, I select one prominent model, describe it in detail, and discuss its implications for the behavior of interest.¹³

1.2.3.1 Models of *Correlation-Invariant Stochastic Choice*

The first class of explanations, which encompasses most preference-based, as well as some heuristics-based models, posit that individuals form a mixture between options due to factors such as non-Expected Utility preferences, random utility shocks, indifference between getting the bonus or not, inherent misperception of probability, inattentive and random mistakes, etc. One thing they share in common is that the sources of mixing are orthogonal to how outcomes are correlated, or the way outcomes and alternatives are presented in the frame. As most payoff framing coincides with the correlation structure in experiments, when the correlation between options varies, both the underlying correlation and the payoff framing change. Thus, these theories predict identical mixing behavior regardless of how the correlation between options presented in the payoff framing varies. For simplicity, these theories are referred to as models of *Correlation-Invariant*

¹²In the context of this study, when the correlation between options varies, the frame changes accordingly as it explicitly reveals the correlation structure. So, when stating that the correlation between options changes, I am referring to a simultaneous variation in both the interdependence of outcomes between options and the choice frame that explicitly illustrates this relationship.

¹³Please refer to Appendix A.1 for a more comprehensive discussions of these models.

Stochastic Choice.¹⁴ This category includes preference-based models, such as the probability weighting, perturbed utility (Fudenberg, Iijima and Strzalecki, 2015; Siegel, 1961), correlation-sensitive preferences (Lanzani, 2020) and the pairwise normalization model (Landry and Webb, 2021), as well as the heuristics-based models such as drift-diffusion models (Ratcliff, 1978) and expectation matching (Kogler and Kühberger, 2007).¹⁵

For instance, the perturbed utility model exemplifies the preferences over two-stage lotteries that allow violation of compound lottery reduction. It posits that decision maker provides different answers because they can gain extra utility from mixing itself, which does not change with the correlation structure presented in the payoff framing. Formally, the decision maker chooses a mixture $(\alpha, 1 - \alpha)$ with $0 < \alpha < 1$ to maximize expected utility plus a utility value from mixing (Fudenberg, Iijima and Strzalecki, 2015; Siegel, 1961) as follows:

$$\begin{aligned} \max_{\alpha} \sum_x \alpha(x)u(x) + V(\alpha) \\ = (\alpha * p + (1 - \alpha) * (1 - p)) * u(\$M) + (\alpha * (1 - p) + (1 - \alpha) * p) * u(\$0) + V(\alpha) \end{aligned} \tag{1.2}$$

where $u(\cdot)$ is the utility function and $V(\cdot)$ is the utility from mixing which is a function

¹⁴Most models in this category fail to consider the possibility that individuals evaluate each option not only based on its own outcomes but also in comparison to the outcomes of the alternative in each state. While some theories do account for state-wise comparisons of outcomes between options such as correlation-sensitive preferences (Lanzani, 2020) and the pairwise normalization model (Landry and Webb, 2021), they still fail to predict varied mixing behavior in response to different correlations between options. I refer interested readers to Appendix A.1 for more detailed discussions.

¹⁵The generalized perturbed utility models proposed by Allen and Rehbeck (2023), which makes no assumption on the utility function of the reduced lottery \mathcal{L} , predict identical mixing behavior in response to the correlation change for two reasons. On the one hand, I show in the Appendix A.1 that existing preferences that account for state-wise comparisons cannot predict varied mixing behavior. This is either due to their inherent features, or because they do not additionally assume convex preferences. Thus, it predicts identical choices, even if we apply these preference-based models here. On the other, this model posits that the source of mixing –gaining extra utility from mixing – is orthogonal to the evaluations of options.

of α and orthogonal to the correlation between options. It implies that once the marginal distributions are fixed, this condition does not vary with changes in the correlation between options or in the corresponding frame. So does the mixing behavior predicted by this model.

Hypothesis 2. *Most preference-based and some heuristics-based models, which posit that individuals mix due to factors such as non-standard preferences or errors, predict that the decision maker will make identical choice regardless of correlation or framing — either choosing the dominant option with $\alpha = 100\%$ or mixing between the two in all the decision problems across the three scenarios.*

1.2.3.2 Models of *Correlation-Sensitive Stochastic Choice*

Several heuristics propose that individuals are sensitive to how options are correlated for reasons such as deliberately using mixing as a tool to hedge against the (misperceived) risk.¹⁶ Examples in this category of models include the model of minmax regret with a convex cost of “mistakes” (Agranov, Healy and Nielsen, 2023); irrational diversification models (Baltussen and Post, 2011; Rubinstein, 2002); and the evolutionary model developed by Brennan and Lo (2012).

To illustrate this concept, let us apply Agranov, Healy and Nielsen (2023)’s model to the running example. The decision maker needs to find the optimal α to maximize the following utility function:

$$\begin{aligned} \max_{\alpha} & \alpha * u(\text{Option A}) + (1 - \alpha) * u(\text{Option B}) \\ & - \lambda \frac{1}{4} \sum_{\omega_i \in \Omega} (w(\alpha) \max\{B(\omega_i) - A(\omega_i), 0\} + w(1 - \alpha) \max\{A(\omega_i) - B(\omega_i), 0\}) \end{aligned}$$

¹⁶It is a “misperceived” risk, because it is a probabilistic choice distribution and implemented once to determine the payoff. Risk preferences play no role.

where the state space $\Omega = \{\omega_1, \dots, \omega_4\}$, $u(\cdot)$ represents the expected utility of each option, $\lambda \geq 0$ denotes an individual-specific scale parameter, $B(\omega_i)$ ($A(\omega_i)$) represents the outcome of Option B (Option A) in state ω_i , and $w(\cdot)$ is an increasing and weakly convex function satisfying $w(0) = 0$. The summation term counts, for each state, the fraction of times the decision maker might make a “mistake” in that state. In each state, choosing one option is considered as a “mistake” if it could have yielded better outcome by switching to the alternative. This count is then weighted by the convex function $w(\cdot)$ and multiplied by the payoff magnitude of the mistake. Convexity captures the idea that the decision maker finds it particularly undesirable to have states where most choices they have made turn out to be mistakes. Thus, the decision maker may tolerate a lower occurrence of mistakes in certain states in order to reduce mistakes in states where they have many.

Different correlations between options affect the extent to which choosing the dominant option with certainty is deemed as “mistakes” from an *ex post* perspective. To illustrate, suppose $w(x) = x^2$. Thus, in decision problems with the CPS, choosing the dominant option (Option A) with 100% would turn out to be a severe “mistake” when State ω_4 is realized. Maximizing the cost term with respect to α gives the *exact PM*. λ captures the level of tension between choosing the more likely option to maximize the expected utility, and matching with the probability to reduce the cost of mistakes. The decision maker who places a higher value on λ will lean more toward *exact PM*. When $\lambda = 0$, though, the decision maker will choose the dominant option with $\alpha = 100\%$.

On the contrary, when facing decision problems in the APS as shown in Table 1.2, choosing Option A with 100% will never be treated as a “mistake” regardless of which state will get realized. Therefore, in decision problems with the APS, choosing the dominant option with 100% both maximizes the expected utility and minimizes the cost of mistakes. Moreover, once the correlation is fixed at zero, the decision maker will mix

as it is still likely that Project B outperforms Project A, which does not vary with the framing in the Independence Scenario.

For the decision problems in the Unknown Scenario, if the decision maker believes that each possible joint distribution of options is equally likely to occur when the correlation between options is unknown, they will have a weighted average of the cost term in the CPS (with 25% of probability) and the one in the APS (with 25% of probability).¹⁷ It gives the same predictions as in the decision problems from the Independence Scenario. Note that it has an implicit assumption that there is no friction in perceiving the joint distribution of the two options ($A : p; B : 1 - p$) with zero correlation. The theoretical predictions of this class of explanations are summarized as the following:

Hypothesis 3. *Several heuristics and biases, arguing that the decision maker deliberately mix to, for example, hedge against (misperceived) risk, predict the following:*

- *In the Baseline Scenario, the decision maker mix between options whenever the payoff structure is not APS, while choosing the dominant option with $\alpha = 100\%$ when it is;*
- *Once the correlation between option is fixed at zero, the decision maker will mix across decision problems in the Independence Scenario;*
- *If the decision maker believes each potential joint distribution of options is equally likely to occur when the correlation is unknown in the Unknown Scenario, and there is no friction in perceiving zero correlation, they will mix in the Unknown Scenario, in the Independence Scenario, and in the tasks with zero-correlation options in the Baseline Scenario.*

¹⁷This is because, with this belief, the success of Project B can occur in each of the four states with equal probability. Thus, there is 75% of chance the joint distribution between Projects A and B has positive correlation as the APS, and 25% of chance the joint distribution has perfectly negative correlation as the CPS.

1.2.4 Framing Effects

Framing Effects posit that, instead of deliberately taking into account the interdependence between options in each state, the decision maker employs heuristics to simplify the comparison between marginal distributions, which is sensitive to the payoff framing rather than the actual correlation structure. With certain payoff framing, they attend to the irrelevant attributes – outcome differences, while ignoring the relevant attributes for decision making – probability differences. Similarity heuristic (Leland, 1998; Rubinstein, 1988) is an example in this class.

When using the similarity heuristic to compare between marginal distributions, individuals tend to naively compare outcomes in each column/partition regardless of whether they are actually correlated: cancelling out same outcomes in each column, while using the columns with dissimilar outcomes to decide. If the comparisons in the columns/partitions with dissimilar outcomes agree on which option is optimal, the decision maker will choose it with 100%. Otherwise, they will resolve by mixing between them (Dertwinkel-Kalt and Köster, 2015; Leland, 1998; Rubinstein, 1988). The Classical Frame as in Table 1.1 or Table 1.3 emphasizes the difference in outcomes:

Table 1.5: Classical Frame

	25%	25%	25%	25%
	<i>Partition 1</i>	<i>Partition 2</i>	<i>Partition 3</i>	<i>Partition 4</i>
Project A	S	S	S	F
Project B	F	F	F	S
	Favor A	Favor A	Favor A	Favor B

As the comparison in each column with dissimilar outcomes does not agree on which option is optimal, the decision maker will resolve it at random. On the contrary, the Alternative Frame (either Table 1.2 or Table 1.4) highlights the difference in probabilities, whereas downplaying the difference in outcomes:

Table 1.6: Alternative Frame

	25%	25%	25%	25%
	<i>Partition 1</i>	<i>Partition 2</i>	<i>Partition 3</i>	<i>Partition 4</i>
Project A	S	S	S	F
Project B	S	F	F	F
	Cancelled out	Favor A	Favor A	Cancelled out

As the comparison in each column with dissimilar outcomes agrees that choosing Project A is optimal, the decision maker will choose Project A with 100%. The theoretical predictions of this class of explanation are summarized below:

Hypothesis 4. *Frame-sensitive heuristics, e.g., similarity heuristic, which posit that the decision maker are indecisive on which option is optimal as they attend to dissimilar but irrelevant attributes (outcome differences), while neglecting relevant attributes (probability differences) when facing certain framing, predict the following:*

- *In the Baseline and Independence Scenarios, the decision maker will mix when facing the decision problems using the Classical Frame, while choosing the dominant option with $\alpha = 100\%$ when facing those using the Alternative Frame, regardless of the actual correlation;*
- *In the Unknown Scenario, where each decision problem is presented with the Alternative Frame, the decision maker will choose the dominant option with $\alpha = 100\%$.*

Table 1.7 below summarizes the predictions from the three classes of theories.

Table 1.7: Theoretical Predictions: Summary
Models of *Correlation-Invariant Stochastic Choice* predict:

Baseline Scenario = Independence Scenario = Unknown Scenario

		<i>Framing Effects</i>	
	models of	Varied Frames	Fixed Frame
<i>Correlation-Sensitive Stochastic Choice</i>	Varied Correlations	Baseline	–
	Fixed Correlation	Independence	Unknown

1.3 Experimental Design

To test these hypotheses, I design an experiment with three treatments corresponding to the three scenarios: Baseline, Independence, and Unknown. Each comprises of three blocks. The main parts of the experiment are Blocks 1 and 2. In each treatment, I use Block 1 to capture the main characteristics across questions in each scenario. Block 1 consists of 30 tasks, which cover six different probability categories: four tasks with ($A : 67\%$, $B : 33\%$), four tasks with ($A : 33\%$, $B : 67\%$), five tasks with ($A : 75\%$, $B : 25\%$), five tasks with ($A : 25\%$, $B : 75\%$), six tasks with ($A : 80\%$, $B : 20\%$), and six tasks with ($A : 20\%$, $B : 80\%$), where the latter three categories are identical to the first three except that Option B is the dominant option. Each of these 30 tasks is presented on a different screen and in a random order. Block 2 is a repetition of Block 1, with a random order to measure the learning effect. Subjects are informed that the computer randomly selects one block and then one choice in that block to determine their final payoffs. The instructions for each block are presented to subjects at the beginning of that block. The complete instructions and screenshots can be found in Appendix A.7.

In the experiment, subjects face a series of tasks similar to the example provided earlier. In each task, to elicit subjects' choice distributions, I follow the Martínez-Marquina, Niederle and Vespa (2019)'s design by asking subjects to allocate tickets to predict which of the two payoff-relevant outcomes, Option A or Option B, will be realized.¹⁸ Only one ticket is randomly selected for payment. If subjects' choices on that selected ticket predicts correctly, they will receive the award \$7. Otherwise, they receive \$0.

Section 1.3.1 begins by describing Blocks 1 and 2 of the Baseline treatment in detail, and then proceeds to demonstrate how the Independence and Unknown treatments differ

¹⁸In the actual experiment, subjects are asked to choose over two options associated with blue color and orange color separately.

from the *Baseline* in Section 1.3.2.¹⁹ Then, Section 1.3.3 presents implementation details.

1.3.1 Baseline

The ticket-allocation tasks in the first two blocks of the Baseline treatment are designed to capture the features of Varied Correlations + Varied Frames. To achieve this goal, in each probability category, I fix each option’s marginal distribution while varying the correlation between options across tasks in a comprehensive way and letting the frame explicitly present the correlation structure. In each task, subjects are told that there is a roll of coins and the computer will randomly draw one coin out of them. Each coin in the roll is labeled with a number to represent one state of the world and has two sides. The front side of each coin is either blue or has no color. The back side of each coin is either orange or has no color. Subjects are asked to predict which color, between blue and orange, is on the randomly drawn coin by allocating some tickets. That is, they need to decide how many tickets to designate for Option A – “Bet on Blue: the randomly drawn coin contains a blue side” and how many tickets to designate for Option B – “Bet on Orange: the randomly drawn coin contains an orange side.” Then, the computer will randomly select a ticket for payment. If the bet on that ticket matches the color of the side on the randomly drawn coin, the subject will get \$7; otherwise, they will receive \$0. After verifying and submitting their choices, subjects receive complete feedback, including which coin was drawn, which ticket was picked, the payoff they will receive, and what they could have received by choosing the alternative option for that ticket.

To fix the marginal distributions of options, the number of coins in the roll, the

¹⁹In each treatment, I also use the same Block 3 to explore the extent to which what subjects learned in the previous blocks can be transferred to a new setting. After Block 3, each subject completes an exit survey so as to record their demographic information. As the results in Block 3 are similar to the main findings in Blocks 1 and 2, I refer readers who are interested in them to Appendix A.5 for detailed design and results.

number of blue sides, and the number of orange sides are identical across tasks from the same probability category ($A : p, B : 1 - p$). In order to comprehensively vary the correlation between options from a perfectly negative correlation ($CORR(A, B) = -1$) to a positive one ($CORR(A, B) > 0$), the locations of colors of the dominant option are fixed, while the locations of colors of the dominated option vary across the tasks from the same probability category. On each coin, the blue and orange sides are not mutually exclusive from each other. Tasks with $CORR(A, B) = -1$ correspond to the CPS, and the tasks with $CORR(A, B) > 0$ map to the APS.

Take the five tasks in Category ($A : 75\%, B : 25\%$) as an example. Table 1.8 demonstrates the correlations and interfaces of the coin rolls that subjects can see in each of the five tasks. Subjects can see a roll of sixteen double-sided coins in each task. Twelve coins are blue on the front sides, and the rest have no color on the this side. Four coins are orange on the back sides, while the rest have no color on this side. The correlation among five tasks marginally increases from a perfectly negative one ($CORR(A, B) = -1$) to a positive one ($CORR(A, B) = \frac{1}{3}$), as explicitly presented in the interface.²⁰

1.3.2 Treatment Variations

I use the Independence and Unknown treatments to capture the main features of Fixed Correlation + Varied Frames and Fixed Correlation + Fixed Frame, respectively. In the Independence treatment, I fix the correlation between options to remove the models of *Correlation-Sensitive Stochastic Choices* as a potential candidate for explaining varied mixing behavior if it is observed in the Baseline treatment. However, the *Framing Effects* could still play a role in driving varied mixing behaviors across tasks. In the Unknown treatment, I fix both the correlation between options and payoff framing to benchmark

²⁰See Appendix A.3 for the payoff structures in Categories ($A : 67\%, B : 33\%$) and ($A : 80\%, B : 20\%$). The other three categories are identical except that Option B is the dominant option.

Table 1.8: Baseline: Five Tasks under Category ($A : 75\%$, $B : 25\%$)

Task	$CORR(A, B)$	Interface
1	$\frac{1}{3}$	Front: Back:
2	0	Front: Back:
3	$-\frac{1}{3}$	Front: Back:
4	$-\frac{2}{3}$	Front: Back:
5	-1	Front: Back:

Note: Each task has a roll of 16 double-sided coins denoted by a number 1 - 16. 12 out of the 16 coins are blue on the front side, and 4 out of the 16 coins are orange on the back side. Tasks 1 and 5 correspond to the APS and the CPS, respectively.

Table 1.9: Independence: Five Tasks under Category ($A : 75\%$, $B : 25\%$)

Task	$CORR(A, B)$	Interface
1	0	Roll Blue: Roll Orange:
2	0	Roll Blue: Roll Orange:
3	0	Roll Blue: Roll Orange:
4	0	Roll Blue: Roll Orange:
5	0	Roll Blue: Roll Orange:

Note: Each task has two rolls of 16 coins, Roll Blue and Roll Orange. b1-b16 represent the 16 coins in Roll Blue, and o1-o16 denote the 16 coins in Roll Orange. The computer will randomly select two coins: one from each coin roll. The locations of colors are identical to the Baseline.

Table 1.10: Unknown: Five Tasks under Category ($A : 75\%$, $B : 25\%$)

Task	Expected $CORR(A, B)$	Interface
1	0	Front: Back:
2	0	Front: Back:
3	0	Front: Back:
4	0	Front: Back:
5	0	Front: Back:

Note: Each task uses a unified frame to present the colors of the two-sided coins. Subjects are not informed about how colored sides are correlated in each coin. If they believe that each possible joint distribution is equally likely, their expected correlation between options is zero.

subjects' behavior when the *Framing Effects* does not come into play.

Independence Treatment Tasks in the Independence treatment are identical to the Baseline in most aspects, except that, in each task of the first two blocks, the outcomes of the two options are independently determined. To accomplish this, in each task, subjects can see two rolls of coins: Roll Blue and Roll Orange. Each coin in Roll Blue is either blue or has no color. Similarly, each coin in Roll Orange is either orange or has no color. The computer randomly selects two coins, one from each roll. To maintain the same variation of the frame as the Baseline treatment, under each probability category, I let the locations of colored coins vary across tasks in the same way as in the Baseline. On each ticket, subjects are asked to predict which roll the coin is drawn from has color by choosing between Option A – "Bet on Blue: the coin drawn from Roll Blue has color" and Option B – "Bet on Orange: the coin drawn from Roll Orange has color." Subjects receive feedback only on which ticket is picked and the payoff they receive. Table 1.9 demonstrates the correlation and interfaces that subjects can see across the five tasks in Category ($A : 75\%$, $B : 25\%$) within the Independence treatment.

Unknown Treatment For each task in Blocks 1 and 2 of the Unknown treatment, I fix both the correlation and the framing, as shown in Table 1.10. All the other components are identical to the Baseline treatment. To fix the correlation, subjects are not informed about how colored sides are correlated on each coin, as the information regarding correlation is irrelevant for expected utility maximization. To control the framing, I use a unified frame, the Alternative Frame, to present the possible outcomes of the two-sided coins.²¹ Subjects receive feedback only on: which ticket was picked and the payoff in that round only to prevent learning the ex post correlation from feedback.

²¹In the instruction, subjects are explicitly told that the framing is a visualization of the marginal probability that each outcome occurs.

1.3.3 Implementation Details

The experiment was conducted via online Zoom sessions from May to July 2021. I recruited 157 subjects through the EBEL laboratory at the University of California, Santa Barbara, using Online Recruitment System for Economic Experiments (ORSEE) recruiting software (Greiner, 2015). The experiment interface was programmed in oTree by the author. There were 11 sessions in total, each lasting 45–55 minutes on average. All the treatments were balanced and randomly assigned to subjects in each session, and the average payoff per subject was \$10 (including a \$5 for the participation fee).

1.4 Results

This section is organized as follows: Section 1.4.1 discusses the preliminary results. Section 1.4.2 reports the main results across the three treatments. In Section 1.4.3, I discuss how the decisions to mix differs from the decisions to match exactly to the probability. In Section 1.4.4, I explore individual heterogeneity by classifying subjects into different types, based on the choices they made.

The primary focus lies with how subjects' mixing behavior varies across the tasks in Blocks 1 and 2 of each treatment. This requires a definition of mixing behavior. I adopt the strictest definition used in the previous literature (Martínez-Marquina, Niederle and Vespa, 2019), in that an allocation choice is referred to as *mixed* if it allocates at least one ticket to the dominated option – betting on the color in the minority. An allocation choice is defined as *exact PM* if it allocates an exact fraction p of tickets to the dominant option and the remaining fraction $(1 - p)$ to the dominated option in the decision problem with $(A : p, B : 1 - p)$. Thus, the fraction of allocation choices that are *mixed* and those that are *exact PM* are denoted the *likelihood of mixing* and the *likelihood of exact PM*, respectively.

Another approach to defining mixing behavior is to calculate the fraction of tickets allocated by subjects to the dominated option. Since the qualitative findings are identical between these two definitions, and the quantitative results primarily stem from the change in the fraction of choices that allocate at least one ticket on the dominated option, I direct the reader to Appendix A.4 for the parallel results obtained using this alternative definition.

In each regression analysis, I pool decision problems with symmetric probability distributions, namely, $(A : p, B : 1 - p)$ and $(A : 1 - p, B : p)$, together into a single category, and represent tasks from Categories $(A : p, B : 1 - p)$ and $(A : 1 - p, B : p)$ as Category $(p, 1 - p)$. I use the variable “correlation parameter” to denote different things in different treatments: (1) in the Baseline treatment, it captures both the correlation and the corresponding frame in the payoff structure between options; (2) in the Independence treatment, it is the corresponding frame only; and (3) in the Unknown treatment, it denotes the *ex post* correlation, which is unknown from subjects’ perspective by design. Standard errors are clustered at the subject level in all regression analyses. Additionally, all regression models include categorical variables for the probability category $(p, 1 - p)$, gender, and school year, as well as indicator variables for the dominant option and STEM, serving as controls. Each bar graph is shown with 95% confidence intervals.

1.4.1 Preliminaries

Figures 1.1 and 1.2 plot the *likelihood of mixing* and *likelihood of exact PM* against the correlation parameters with the first two blocks combined, separated by probability categories for each treatment, respectively.²² The horizontal axis varies the correlation

²²As I pooled the tasks from $(A : p, B : 1 - p)$ and those from $(A : 1 - p, B : p)$ together, each subject has two allocation choices added when calculating the fraction of *mixed* choices given the probability category and correlation parameter. This is also true for the calculation of fraction of *exact PM* choices. Thus, the fractions of *mixed* choices and *exact PM* choices are different from the fractions of subjects who

parameter, representing different features in different treatments. As tasks in the Unknown treatment use a fixed correlation and identical frame, I use horizontal lines to represent the average *likelihood of mixing* and *likelihood of exact PM* in each probability category.

Note that models of *Correlation-Invariant Stochastic Choice* predict identical *likelihood of mixing* in all the tasks across the three treatments. Models of *Correlation-Sensitive Stochastic Choice* predict identical *likelihood of mixing* in all the tasks across the three treatments, except in cases where the two options are positively correlated, as in the APS, and in the APS, the *likelihood of mixing* shall decrease to zero. Moreover, the *Framing Effects* predict equal *likelihood of mixing* in all the tasks across the three treatments, except in cases where the two options are presented in the Alternative Frame regardless of the correlation, and in the Alternative Frame, the *likelihood of mixing* shall decrease to zero.

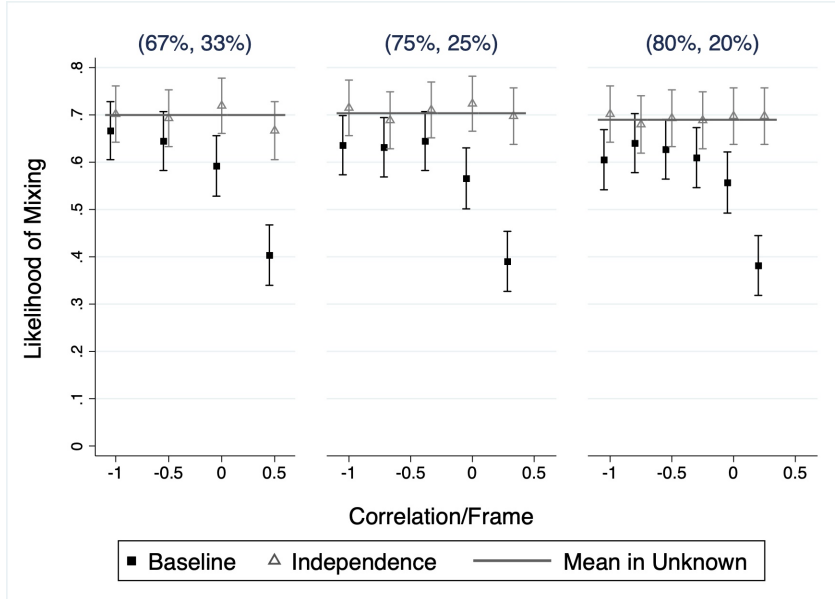
Firstly, I replicate the findings observed in the existing literature, wherein the vast majority of individuals tend to mix two options in allocation choice or even match exactly to the probability of occurrence when facing the CPS. As depicted in Figures 1.1 and 1.2, in the CPS, approximately 65% of allocation choices mix between the two options, whereas 35% of the choices match exactly to the occurrence probability in each probability category.²³

More importantly, as depicted in Figures 1.1 and 1.2, within each probability category, the average fractions of *mixed* choices and of *exact PM* choices decrease when the correlation between options increases or when the frame varies in the Baseline treatment. However, the *likelihood of mixing* and the *likelihood of exact PM* are not zero in the APS. About 30% - 40% of choices are *mixed* and 10% - 20% of choices are *exact PM*

made *mixed* choices and *exact PM* choices. I discuss the distribution of mixing types in Section 1.4.4.

²³Martínez-Marquina, Niederle and Vespa (2019) find that about 67.8% of choices are *mixed*, while 20.8% are *exact PM* choices.

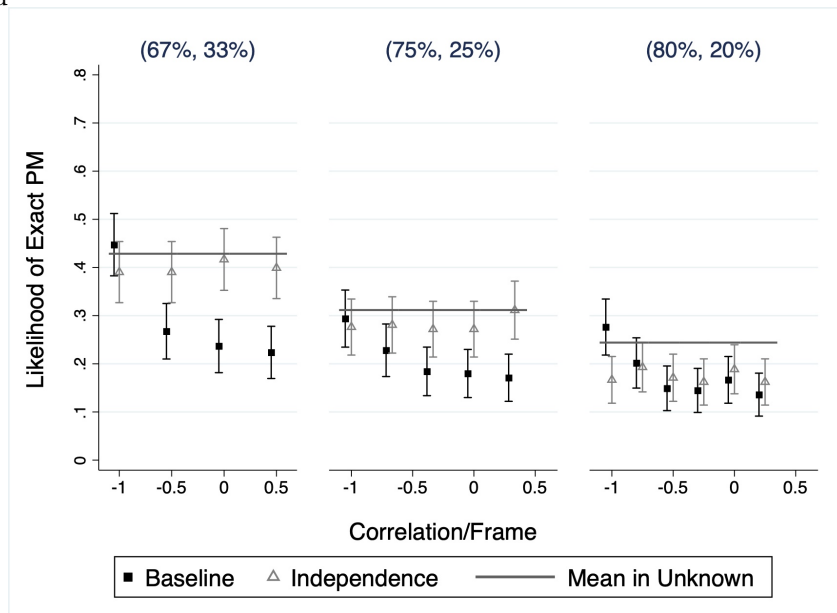
Figure 1.1: Impacts of correlation/frame on likelihood of mixing with two blocks combined



Note: The horizontal axis denotes different features in different treatments. In the Baseline, it represents varied correlations and varied frames. In the Independence, it represents varied frames. As the Unknown uses the fixed correlation and frame, green lines represent the average fractions of *mixed* allocation choices for each probability category in this treatment. Each panel represents one probability category. The error bars depict 95% confidence intervals.

in the APS where the two options are positively correlated. In contrast to Baseline, the average fractions of *mixed* choices and of *exact PM* choices do not vary with changes in the corresponding frame in the Independence treatment. Furthermore, in the Unknown treatment, the magnitude of the average fraction of *mixed* choices is nearly identical to that in the Independence treatment. In the tasks featuring the “zero-correlation frame” in the Independence treatment, the *likelihood of mixing* and the *likelihood of exact PM* are slightly larger compared to the zero-correlation tasks in the Baseline. It suggests that subjects in the former tasks might misinterpret the zero correlation, as described in words in the Independence treatment, compared to the latter tasks where the joint distribution of zero-correlation options is presented in reduced form in the Baseline. More discussions can be found in Section 1.5. Exploring whether this is a result of misinterpretation or other confounds would be a fruitful direction for future research.

Figure 1.2: Impacts of correlation/frame on likelihood of Exact PM with two blocks combined



Note: The horizontal axis denotes different features in different treatments. In the Baseline, it represents varied correlations and varied frames. In the Independence, it represents varied frames. As the Unknown uses the fixed correlation and frame, green lines represent the average fractions of *exact PM* choices for each probability category in this treatment. Each panel represents one probability category. The error bars depict 95% confidence intervals.

1.4.2 Treatment Level Results

In order to investigate the empirical validity of each class of models, I estimate the impacts of correlation change and frame change on mixing behavior by making pairwise comparisons across the three treatments.

More specifically, to estimate the impact of correlation change on mixing behavior, I regress each of the dependent variables – indicators of whether the allocation choice is *mixed*, and whether the allocation choice is *exact PM* – on: (1) the indicator variable of treatments: Independence vs Baseline; (2) correlation parameter; and (3) the interaction term between the first two variables. The dependent variables capture the *likelihood of mixing* and *likelihood of exact PM*, respectively. The coefficient for the interaction term estimates the impact of correlation changes in the Baseline by cancelling out the impact of frame changes in the Baseline with those in the Independence treatment.

Table 1.11 makes the comparison between the Baseline and Independence treatments. As illustrated in Table 1.11, after controlling for the framing effects, subjects are 15.8% (OLS, $p = 0.000$) less likely to make *mixed* choices and 12.5% (OLS, $p = 0.000$) less likely to make *exact PM* choices when the correlation increases in the Baseline with the two blocks combined. More significantly, the estimated correlation effects are more substantial in Block 2 than in Block 1. The estimated impact of correlation on the *likelihood of mixing* changes from -9.8% in Block 1 to -21.8% in Block 2. Similarly, the estimated impact of correlation on the *likelihood of exact PM* changes from -10.8% in Block 1 to -14.2% in Block 2. This suggests that learning amplifies subjects' responsiveness to changes in the correlation between options. Such a finding thus indicates that subjects, on average, are responsive to changes in the correlation between options when making decisions, which cannot be fully accommodated by models of *Correlation-Invariant Stochastic Choices*, which contains most preference-based models and some heuristics.

Table 1.11: Baseline VS Independence: Impacts of Correlation/Frame on Mixing Behavior

	Blocks 1 & 2 <i>mixed</i>	Block 1 <i>mixed</i>	Block 2 <i>mixed</i>	Blocks 1 & 2 <i>exact PM</i>	Block 1 <i>exact PM</i>	Block 2 <i>exact PM</i>
IvsB X Correlation/Frame	-0.158*** (0.0366)	-0.0977*** (0.0334)	-0.218*** (0.0452)	-0.125*** (0.0286)	-0.108*** (0.0311)	-0.142*** (0.0380)
Correlation/Frame	-0.00586 (0.0114)	-0.00990 (0.0106)	-0.00182 (0.0161)	0.0121 (0.0114)	0.0199 (0.0183)	0.00439 (0.0128)
IvsB(Baseline=1)	-0.175*** (0.0661)	-0.144** (0.0663)	-0.205*** (0.0721)	-0.0877* (0.0518)	-0.0564 (0.0515)	-0.119** (0.0565)
Constant	0.241 (0.224)	0.261 (0.214)	0.221 (0.252)	0.165* (0.0868)	0.159* (0.0839)	0.170 (0.104)
Observations	6840	3420	3420	6840	3420	3420

Note: Results from OLS regression. The dependent variable takes the value of 1 if the allocation choice in a task is classified as *mixed*, or as *exact PM*, respectively. CORR/Frame represents the variable of correlation parameters in the Baseline and Independence treatments. In the Baseline, it captures either the correlation change or the associated frame change. In the Independence, it denotes the frame change only. IvsB is the dummy variable for whether the task comes from the Baseline or Independence. It takes the value of 1 if the task comes from the Baseline. Each regression model also includes the categorical variables of probability categories, gender, and school year, as well as the indicator variables of dominant color, and STEM, as controls. Standard errors are clustered at the subject level and listed in parentheses. Full regression results can be found in Appendix A.6.1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Result 1. *On average, subjects are sensitive to changes in the correlation between options: they are less likely to mix between options or even match precisely to the probability of occurrence when the correlation between options increases. And learning amplifies the responsiveness to the correlation change: with some experience, the mixing behavior decreases further when the correlation increases.*

Along these same lines, I compare the Independence and Unknown treatments to estimate the impact of frame change on mixing behavior. As subjects are not informed about the *ex post* correlation between options, the mixing behavior does not vary across problems in the Unknown treatment by design. The coefficient on the interaction term between the indicator variable of Independence VS Unknown and the variable of correlation parameter estimates the impact of frame change on mixing behavior.

As shown in Table 1.12, the estimated impact of frame change on the *likelihood of*

mixing is not significantly different from zero ($p = 0.777$). Moreover, for the *likelihood of exact PM*, the impact is limited and is altogether absent in Block 2. Regression results show that, when combining Blocks 1 and 2 together, subjects are 3% less likely to make *exact PM* choices in the Independence treatment when the frame changes. The impact is significant in Block 1 (-6.5% , $p = 0.000$). However, as subjects gain experience in Block 2, this impact disappears. Therefore, our result indicates that subjects are not responsive to the change in frame, which rules out the *Framing Effects* — mixing due to some frame-sensitive heuristics employed to simplify the comparison between marginal distributions — as the leading explanation behind mixing behavior in decision problems using the CPS.

Result 2. *Subjects, on average, are not responsive to variations in the framing when making decisions and learning mitigates this impact even further.*

Combined with the previous findings between the Baseline and Independence treatments, the aggregate results are more consistent with models of *Correlation-Sensitive Stochastic Choices* than with models of *Correlation-Invariant Stochastic Choices* or *Framing Effects*. These results suggest that subjects deliberately take into account how outcomes of options are correlated with one another in each state of the world when making decisions. The observed mixing behavior in decision problems using the CPS therefore tends to be subjects' responses to the perfectly negative correlation between options for reasons such as mix to hedge against the (misperceived) risk.

1.4.3 Decision to Mix VS Decision to Exact PM

By zooming in on the results for the Baseline, as shown in Figure 1.3, I find that subjects' decisions to mix between the dominant and dominated options are different from their decisions to match precisely to the outcome probability in two ways: (1) different

Table 1.12: Independence VS Unknown: Framing Effects on Mixing Behavior

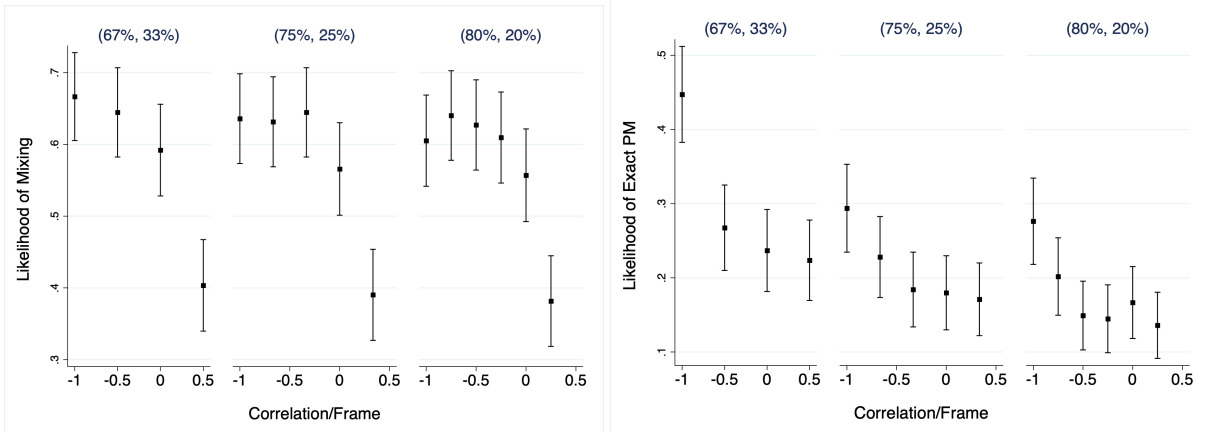
	Block 1&2 <i>mixed</i>	Block 1 <i>mixed</i>	Block 2 <i>mixed</i>	Block 1&2 <i>exact PM</i>	Block 1 <i>exact PM</i>	Block 2 <i>exact PM</i>
IvsU X Frame/Ex post <i>CORR</i>	0.00387 (0.0136)	0.00964 (0.0148)	-0.00190 (0.0192)	-0.0318** (0.0157)	-0.0648*** (0.0244)	0.00122 (0.0190)
Frame/Ex post <i>CORR</i>	-0.00509 (0.0114)	-0.00947 (0.0106)	-0.000710 (0.0161)	0.00962 (0.0114)	0.0177 (0.0183)	0.00159 (0.0129)
IvsU(Independence=1)	0.0469 (0.0700)	0.0664 (0.0710)	0.0274 (0.0763)	0.0566 (0.0567)	0.0920 (0.0572)	0.0211 (0.0612)
Constant	0.422 (0.289)	0.373 (0.275)	0.471 (0.314)	0.140 (0.143)	0.167 (0.160)	0.112 (0.136)
Observations	6780	3390	3390	6780	3390	3390

Note: Results from OLS regression. The dependent variable takes the value of 1 if the allocation choice in a task is classified as (1) a *mixed* choice, or (2) a *exact PM* choice. Frame/Ex post *CORR* is the variable of correlation parameters in the Independence and Unknown treatments. In the Independence, it captures the frame change only. In the Unknown, it denotes no impact by design. IvsU is the dummy variable on whether the task is in the Unknown VS Independence. It takes the value of 1 if the task is in the Independence. The regression also includes probability categories, dominant color, gender, school year and STEM as controls. Standard errors are clustered at the subject level and listed in parentheses. Full regression results can be found in Appendix A.6.1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

levels of responsiveness to the marginal changes in the correlation between options; and (2) different levels of responsiveness to changes in the probability of the dominant option paying off.

First of all, although variation in the correlation between options, on average, exerts negative impacts on both the *likelihood of mixing* and the *likelihood of exact PM*, such variation are attributed to different marginal changes in the correlation between options. As demonstrated in Figure 1.3, there is a considerable decline in the *likelihood of mixing* when the correlation marginally increases from weakly negative, to zero or even to positive. However, a significant drop in the *likelihood of exact PM* occurs when the correlation marginally increases from perfectly negative to moderately negative.

Figure 1.3: Baseline: Impacts of Correlation on Decision to Mix Versus Decision to Exact PM



Notes: The error bars depict 95% confidence intervals.

To formally measure these differences, I estimate the impacts of marginal correlation change at $CORR = -1$, at $CORR = 0$, and at $CORR \geq 0$, respectively. For the impact of marginal correlation change at $CORR = -1$, I compare allocation choices between tasks featuring $CORR = -1$ and tasks with moderately negative correlations in the two blocks.²⁴ For the impacts of marginal correlation change at $CORR = 0$, I compare allocation choices between tasks with $CORR = 0$ and tasks featuring weakly negative correlations.²⁵ For the impacts of marginal correlation change at $CORR \geq 0$, I combine the tasks with non-negative correlations, and compare the allocation choices noted there with those in the tasks with weakly negative correlations.²⁶

According to the regression results presented in Table 1.13, subjects are 34.7% (OLS, $p = 0.000$) less likely to make *exact PM* choices when the correlation marginally increases

²⁴Moderately negative correlations refer to the negative correlations that are closest to the perfectly negative one. To be more specific, tasks with moderately negative correlations include those with $CORR = -0.5$ in Category (67%, 33%), those with $CORR = -0.67$ in Category (75%, 25%), and those with $CORR = -0.75$ in Category (80%, 20%).

²⁵Weakly negative correlations refer to the negative correlations that are closest to zero. Specifically, tasks with weakly negative correlations include those with $CORR = -0.5$ in Category (67%, 33%), those with $CORR = -0.33$ in Category (75%, 25%), and those with $CORR = -0.25$ in Category (80%, 20%).

²⁶Tasks with positive correlations are those with $CORR = 0.5$ in Category (67%, 33%), those with $CORR = 0.33$ in Category (75%, 25%), and those with $CORR = 0.25$ in Category (80%, 20%).

from $CORR = -1$. However, the same marginal correlation change does not have significant impacts on the fraction of *mixed* choices. Moreover, when the correlation marginally increases to $CORR = 0$, subjects are 12.6% (OLS, $p = 0.000$) less likely to make *mixed* choices. In contrast, the same marginal change does not have a significant impact on the fraction of *exact PM* choices. Subjects are 30.4% (OLS, $p = 0.000$) less likely to mix when the correlation increases from weakly negative to non-negative correlations. However, the same change does not significantly affect the fraction of *exact PM* choices.

Table 1.13: Baseline: Marginal Impacts of Correlation on Mixing Behavior

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>mixed</i>	<i>exact PM</i>	<i>mixed</i>	<i>exact PM</i>	<i>mixed</i>	<i>exact PM</i>
Marginal Change at $CORR = -1$	-0.00702 (0.0352)	-0.347*** (0.0795)				
Marginal Change at $CORR = 0$			-0.126** (0.0599)	-0.0316 (0.0616)		
Marginal Change at $CORR > 0$					-0.461*** (0.100)	-0.0367 (0.0544)
Constant	0.133 (0.308)	-0.122 (0.154)	0.0870 (0.322)	0.0243 (0.0840)	0.903*** (0.256)	0.260* (0.141)
Observations	1140	1140	1140	1140	1368	1368

Note: Results from OLS regression. The dependent variables take the value of 1 if the allocation choice in a task is classified as a *mixed* choice, or as a *exact PM* choice. The variable of marginal correlation change at $CORR = -1$ takes the value of 1 if the correlation parameter is -0.5 for Category (67%, 33%), -0.67 for Category (75%, 25%), or -0.75 for Category (80%, 20%), and takes the value of 0 if $CORR = -1$. The variable of marginal correlation change at $CORR = 0$ takes the value of 1 if $CORR(B, O) = 0$, and takes the value of 0 if the correlation parameter is -0.5 for Category (67%, 33%), -0.33 for Category (75%, 25%), or -0.25 for Category (80%, 20%). The variable of marginal correlation change at $CORR \geq 0$ takes the value of 1 if $CORR(B, O) \geq 0$ which includes the correlation parameter that is 0 for all categories, 0.5 for Category (67%, 33%), 0.33 for Category (75%, 25%), or 0.25 for Category (80%, 20%), and takes the value of 0 if the correlation parameter is -0.5 for Category (67%, 33%), -0.33 for Category (75%, 25%), or -0.25 for Category (80%, 20%). The regression also includes probability categories, dominant color, gender, school year and STEM as controls. Standard errors are clustered at the subject level and listed in parentheses. Full regression results can be found in Appendix A.6.1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

These results demonstrate a substantial distinction between the decision to mix between two options and the decision to precisely match to the probability of occurrence in

terms of responsiveness to different marginal changes in the correlation between options. The former decision is more heavily influenced by whether the two options subjects must choose between are positively correlated (i.e., the APS). Note that, when the two options are positively correlated, one option not only exhibits first-order stochastic dominance but also state-wise dominance over the other. Hence, the decision to mix is more likely to stem from an evaluation of whether one option dominates the other in a state-by-state manner. In line with the explanation of needs to hedge against (misperceived) risk, it could be the case that when one option dominates the other in each state, there is no way to hedge against the result of getting nothing with choosing the dominant option when State ω_4 occur. It is also possible that when the dominant option outperforms the alternative in each state, subjects tend to feel more certain that the option is the better one, and thus become less likely to make *mixed* choices.²⁷

On the contrary, the decision to match exactly to the outcome probability largely depends on whether the two options that subjects must consider are perfectly negatively correlated (i.e., the CPS). Put otherwise, *exact PM* tends to be a response towards the particular “perfect complementary” relationship between options – whenever one option yields a good outcome, the alternative yields a bad one, and vice versa. This finding suggests that matching precisely to the outcome probability is more likely to be a form of context-specific bias, which is triggered by the “perfect complementary relations” of the outcomes between options: either Project A succeeds or Project B succeeds, but not both.

²⁷This interpretation aligns with the notion of incomplete preference or indecisiveness, which posits that decision makers may choose to mix when they are unsure of which option to choose, and they use mixing between options as a way to resolve such uncertainty. Formalizing this requires assuming specific functional forms of preference, and thus it would revert back to the complete but non-EU preference. That is why I do not discuss this branch separately in Section 1.2. Existing theories that can formalize this fail to take into account the possibility that the decision maker cares about how options are jointly determined in each state, and thus are classified in the models of *Correlation-Invariant Stochastic Choice*. Cautious Expected Utility proposed by Cerreia-Vioglio, Dillenberger and Ortoleva (2015) is an example of this.

The findings are summarized as follows:

Result 3. *The average negative impacts of correlation on the likelihood of mixing and on the likelihood of exact PM are driven by different marginal correlation changes:*

- *Subjects are significantly less likely to make mixed choices when the correlation between options marginally varies from weakly negative to non-negative correlations. However, the same marginal change does not impact the likelihood of exact PM.*
- *Subjects are significantly less likely to make exact PM choices when the correlation marginally varies from perfectly negative to moderately negative correlations. In contrast, the same change does not significantly affect the likelihood of mixing.*

Secondly, the decision to mix differs from the decision to match exactly to the outcome probability in whether it responds to changes in the probability of the dominant option paying off. Note that the probability of the dominant option yielding $\$M$ ranges from 67% to 80% across probability categories. Agranov, Healy and Nielsen (2023) observe a monotone response among subjects to this change in probability: they are significantly less likely to make *mixed* choices when the dominant option becomes more likely to yield $\$M$ (-0.064 with $p < 0.01$). In contrast to their findings, though, subjects in this experiment exhibit varying levels of responsiveness to this change depending on which decisions they make. As shown in Figure 1.3, when the dominant option is more likely to yield $\$M$, subjects' *likelihood of mixing* does not alter, whereas they are less likely to match precisely to the probability. Furthermore, for both decisions, the magnitude of responsiveness decreases when moving from Block 1 to Block 2.

For the purposes of statistical inference, I regress the indicators of whether the allocation choice is *mixed* and whether it is *exact PM* on the correlation parameters in the Baseline and the categorical variable of probability category separately, as shown in

Table 1.14: Baseline: Impacts of Increasing p of Category ($p, 1 - p$) on Mixing Behavior

	Blocks 1 & 2 <i>mixed</i>	Block 1 <i>mixed</i>	Block 2 <i>mixed</i>	Blocks 1 & 2 <i>exact PM</i>	Block 1 <i>exact PM</i>	Block 2 <i>exact PM</i>
Correlation	-0.165*** (0.0346)	-0.108*** (0.0318)	-0.221*** (0.0420)	-0.108*** (0.0270)	-0.0843*** (0.0261)	-0.133*** (0.0369)
(75%, 25%)	-0.0168* (0.00846)	-0.0248* (0.0134)	-0.00880 (0.0103)	-0.0915*** (0.0233)	-0.109*** (0.0273)	-0.0738** (0.0294)
(80%, 20%)	-0.0272 (0.0185)	-0.0238 (0.0195)	-0.0306 (0.0225)	-0.128*** (0.0246)	-0.152*** (0.0321)	-0.105*** (0.0254)
Constant	0.0266 (0.276)	0.0753 (0.272)	-0.0221 (0.293)	0.0562 (0.0908)	0.0470 (0.0757)	0.0654 (0.117)
Observations	3420	1710	1710	3420	1710	1710

Note: Results from OLS regression. The dependent variables take the value of one if the allocation choice in a task is classified as *mixed* or *exact PM*, respectively. Correlation captures the correlation parameters, which takes values of -1, -0.5, 0, 0.5 for Category (67%, 33%); -1, -0.67, -0.33, 0, 0.33 for Category (75%, 25%); and -1, -0.75, -0.5, -0.25, 0, 0.25 for Category (80%, 20%). Each regression also includes categorical variables of probability categories, gender, and school year, as well as indicator variables of dominant color and STEM, as controls. Standard errors are clustered at the subject level and are listed in parentheses. Full regression results can be found in Appendix A.6.1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.14. The *likelihood of mixing* decreases when moving from the tasks in Category (67%, 33%) to tasks in Category (75%, 25%) with Blocks 1 and 2 combined or considering Block 1 alone, and the coefficients are significantly different from zero at a 90% confidence level. However, the difference becomes insignificant as subjects gain experience in Block 2 or when they face tasks in Category (80%, 20%). Notably, for the *likelihood of mixing*, both the estimated coefficients and confidence levels are much lower than the findings presented in Agranov, Healy and Nielsen (2023)'s study. On the contrary, the *likelihood of exact PM* reacts to this change in probability: as the dominant option is more likely to yield $\$M$, subjects become less likely to match exactly to the probability when combining both blocks. When the two blocks are combined, subjects are 9.2% ($p > 0.001$) less likely to make *exact PM* choices when moving from Category (67%, 33%) to Category (75%, 25%), and 12.8% ($p > 0.001$) less likely to match precisely to the probability when

moving from Category (67%, 33%) to Category (80%, 20%). However, the magnitude of these impacts decreases in Block 2 as compared to Block 1. In line with my previous findings, this result also suggests that subjects' decisions regarding whether to mix or not relate more to whether one option dominates the other in a state-wise manner rather than in a distribution-wise manner. In addition to that, subjects' responsiveness to this change in both decisions is not robust to learning.

Result 4. *The decision to mix and decision to exact PM respond differently to the increase in the probability of the dominant option paying off: as the dominant option becomes more likely to yield \$M, subjects are less likely to match exactly to the probability but their likelihood of mixing remains unchanged. Additionally, learning mitigates this responsiveness: with some experience, both the likelihood of mixing and the likelihood of exact PM are less responsive to this change.*

1.4.4 Mixing Types

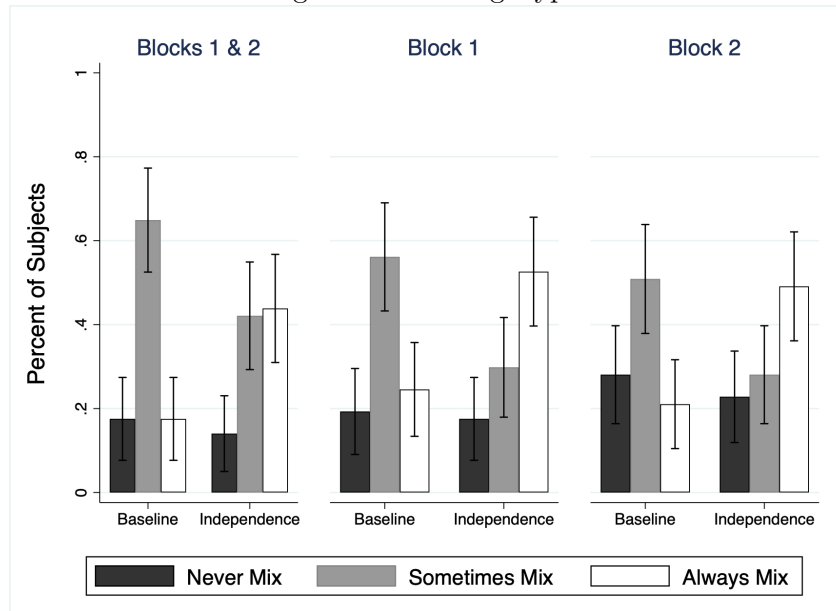
To explore individual heterogeneity, I begin by classifying subjects into three mutually exclusive types based on their behavior in Blocks 1 and 2: subjects who always allocate all the available tickets to the dominant option in all the tasks are called *Never Mix*, subjects who always allocate at least one ticket to the dominated option in all the tasks are called *Always Mix*, and subjects who are in between are called *Sometimes Mix*. Figure 1.4 demonstrates the distributions of subjects across these types based on their choices in Blocks 1 and 2, as well as their choices in Block 1 only and in Block 2 only, across the three treatments.

As shown in Figure 1.4, in the Baseline treatment, approximately 17.5% (10/57) of subjects *Never Mix*; another 17.5% (10/57) of subjects *Always Mix*. Notably, the proportion of subjects who are *Never Mix* increases from 17.5% (10/57) in Block 1 to

28.1% (16/57) in Block 2, the difference of which is significant at a 90% confidence level (OLS, $p = 0.058$). In contrast, the proportion of subjects who are *Always Mix* decreases from 24.6% (14/57) in Block 1 to 21.1% (12/57) in Block 2, although this difference is insignificantly different from zero. These results suggest that with some experience, some subjects learn to make the expected utility maximization choice — choose the dominant option with 100% — in the Baseline treatment.

The most prominent type of subjects is *Sometimes Mix*, which constitutes 64.9% (37/57) of subjects. In fact, *Sometimes Mix* is the most prominent type, not only when considering choices made with the two blocks combined, but also within each individual block of the Baseline. Unlike what is observed in the Baseline, the two most prominent types in the Independence treatment are subjects who *Always Mix* (43.9%) and those who *Sometimes Mix* (42.1%), with both blocks combined. However, within each individual block of the Independence treatment, there is a significantly larger proportion of subjects who *Always Mix* (52.6% in Block 1 and 49.1% in Block 2) compared to those who *Sometimes Mix* (29.8% in Block 1 and 28.1% in Block 2). The distribution of mixing types in the Unknown treatment is similar to that of the Independence treatment. Our results therefore indicate that the *Sometimes Mix* type in the Baseline is quite different from those in the Independence and Unknown treatments.

Figure 1.4: Mixing Types



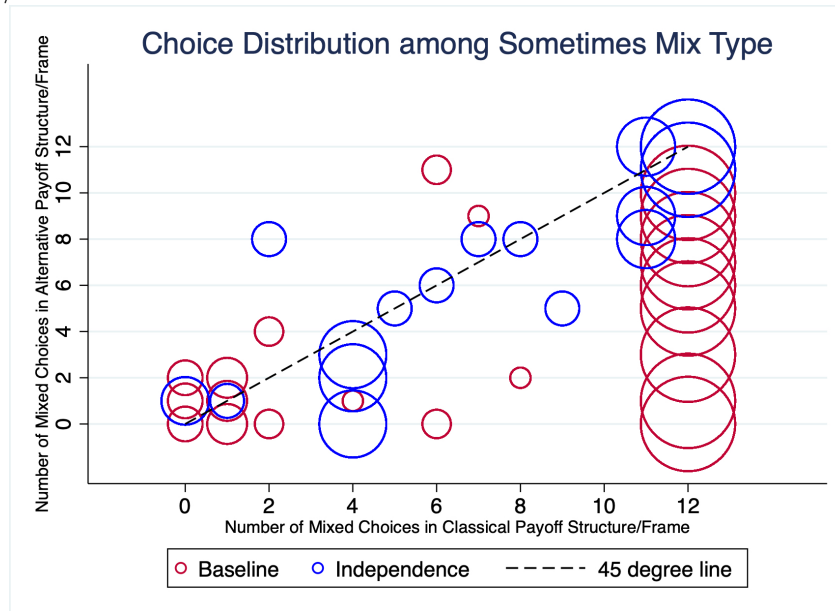
Notes: The error bars depict 95% confidence intervals.

To further investigate whether the prevalence of the *Sometimes Mix* type in the Baseline treatment primarily results from subjects' responsiveness to changes in the correlation between options, I examine the choices made by the *Sometimes Mix* type in greater detail. In essence, I seek to determine whether subjects who *Sometimes Mix* in the Baseline do so because they take the correlation between options into account and are therefore more likely to allocate all the tickets on the dominant option when the two options are positively correlated (i.e., APS), as opposed to the CPS. To achieve this goal,

I calculate, for each subject classified as *Sometimes Mix*, how many out of their choices in the 12 tasks with the APS and in the 12 tasks adopting the CPS, are *mixed*.²⁸ And then, I plot the distribution of subjects who *Sometimes Mix* based on the numbers of *mixed* choices they make in tasks with these two payoff structures, and separated them by treatment, as shown in Figure 1.5.

²⁸With Blocks 1 and 2 combined, there are 12 tasks featuring the APS and 12 tasks featuring the CPS. In each block, there are six tasks associated with each payoff structure.

Figure 1.5: Choice Distribution among *Sometimes Mix* type between CPS/Frames and APS/Frames



Notes: Each circle is weighted by the number of subjects who made the same number of *mixed* choices out of the 12 tasks with the CPS/Classical Frame and of the 12 tasks with the APS/Alternative Frame. For the Baseline treatment, the x-axis and y-axis represent the number of *mixed* choices in tasks with the CPS and those with the APS, respectively. For the Independence treatment, they denote the number of *mixed* choices in tasks with the Classical Frame and those with the Alternative Frame, respectively.

As illustrated in Figure 1.5, subjects who *Sometimes Mix* in the Baseline treatment behave differently from those in the Independence treatment. Among the 64.9% (37/57) of subjects who *Sometimes Mix* in the 60 tasks in the Baseline treatment, 73% (27/37) of them have strictly fewer tasks featuring the APS than tasks employing the CPS, in which they mix between options. In addition, as shown in Figure 1.5, most subjects who *Sometimes Mix* in the Baseline treatment are clustered at “mix in all the 12 tasks with the CPS while allocating all tickets to the dominant options in some tasks using the APS.” On the contrary, in the Independence treatment, among the 42.1% (24/57) of subjects who *Sometimes Mix* in the 60 tasks, 50% (12/24) of them make *mixed* choices in strictly fewer tasks featuring the Alternative Frame than those with the Classical Frame.

As shown in Figure 1.5, subjects who *Sometimes Mix* in the Independence treatment

are more likely to be clustered on the 45 degree line, thus indicating that the numbers of *mixed* choices they made are not significantly different in tasks with the Classical Frame, compared to those with the Alternative Frame. These findings indicate that, for subjects who *Sometimes Mix* in the Baseline treatment, their decisions to allocate all the tickets to the dominant option in some tasks is not accidental. Instead, it is a result of their deliberate consideration of the correlations between options and their corresponding responses.

In sum, subject-level analysis also indicates that most subjects tend to respond to variations in the correlation between options in the Baseline, which is in line with the models of *Correlation-Sensitive Stochastic Choice*. It is important to note that a small fraction of subjects who *Never Mix* (17.5%) is consistent with expected utility benchmark and models that respect FOSD and compound lottery reduction. The same fraction of subjects who *Always Mix* (17.5%) likewise aligns with existing models of *Correlation-Invariant Stochastic Choice*.²⁹ This suggests that the vast majority is not very responsive to frame changes once the correlation between options is fixed, and that given the zero correlation between options, most subjects tend to mix in every task.

I summarize the findings as follows:

Result 5. *In the Baseline treatment, the most prominent type in the population are those who Sometimes Mix (65%), which mainly results from the fact that the vast majority of subjects are less likely to make mixed choices when the correlation between options increases. However, once the correlation between options is fixed, most subjects tend to Always Mix in the Independence and Unknown treatments.*

²⁹It is possible that subjects who *Always Mix* tend to allocate fewer tickets, though not necessarily zero, to the dominated option when the correlation between options increases. However, by regressing the fraction of dominated options in each task on the correlation parameter in the Baseline, I reject this hypothesis by finding the opposite pattern: subjects who *Always Mix* on average allocate 2% of tickets (less than one ticket) more to the dominated options when the correlation increases.

1.5 Discussion

In this paper, I experimentally study the origin of probability matching behavior. I unearth the underlying mechanisms behind probability matching and classify them into three categories of theories: (1) models of *Correlation-Invariant Stochastic Choice*: mixing due to factors orthogonal to the correlation between options such as non-Expected Utility intrinsic preferences, inherent biases, inattentive and random mistakes, indifference between receiving the bonus or not, etc.; (2) models of *Correlation-Sensitive Stochastic Choice*: deliberately mixing to hedge against misperceived risk; and (3) *Framing Effects*: mixing due to some frame-sensitive heuristics (e.g., similarity heuristic), used to simplify comparisons between marginal distributions. Three classes of models have distinctive testable predictions on how the mixing behavior responds to variations in the correlation between options and to different framing separately. Using a novel between-subject design, I demonstrate that subjects deliberately take the correlation between options into account, which can therefore account for a substantial amount of mixing between dominant and dominated options or even matching precisely to the probability of occurrence. In this section, I discuss the implications with respect to the existing models of probability matching.

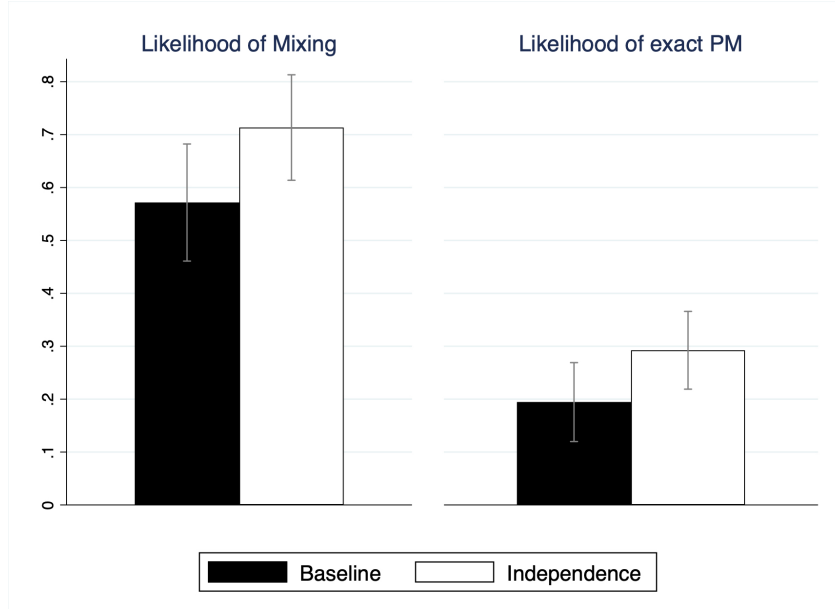
Expected utility benchmark, models of *Correlation-Invariant Stochastic Choice*, and models of *Correlation-Sensitive Stochastic Choice* can accommodate some parts of the findings, but none of them can fully explicate all of them. First of all, a minority of subjects behave consistently with the expected utility benchmark and models of *Correlation-Invariant Stochastic Choices*, which includes most preference-based models and some heuristics. To be more specific, the observation that a sizable proportion (17.5%) of subjects *Never Mix* in any of the 60 tasks is in line with expected utility benchmark: respecting FOSD and compound lottery reduction. Similarly, the finding that an equal

proportion (17.5%) of subjects *Always Mix* in each of the 60 tasks is consistent with models of *Correlation-Invariant Stochastic Choice*. That is, their mixing behavior is due to factors orthogonal to how options are correlation in between such as non-Expected Utility preference, inherent biases, random and inattentive mistakes, indifference between getting the bonus or not, etc. Secondly, although the main finding is consistent with which is argued by models of *Correlation-Sensitive Stochastic Choice*, it is worth noting that none of the existing models in this category, as discussed in Section 1.2, could be concluded to be the main mechanism behind all of these observations for two main reasons.

On the one hand, these models cannot accommodate all the findings, for example, the observation that subjects make different choices in the zero-correlation tasks between the Baseline and Independence treatments, even when the framing is controlled. I compare subjects' choices in tasks where the two options have zero correlation in the Baseline (e.g., Task 2 in Table 1.8), with tasks that employ identical frames in the Independence treatment (e.g., Task 2 in Table 1.9). As depicted in Figure 1.6, subjects are less likely to make *mixed* and *exact PM* choices in the tasks in the Baseline than those in the Independence treatment, despite both sets of tasks featuring the zero correlation between options and employing identical frames. I validate this finding by focusing on the zero-correlation tasks in the Baseline and Independence treatments, i.e., Task 2 in each category, and regressing the indicators of *mixed* choices and *exact PM* choices on the dummy variable of the treatments. As shown in Table 1.15, subjects are 14.1% ($p = 0.04$) less likely to make *mixed* choices and 9.9% ($p = 0.07$) less likely to make *exact PM* choices in the Baseline, compared to the Independence treatment.

One plausible explanation for this discrepancy is that when the zero correlation is described in words in the Independence treatment, subjects tend to interpret it differently from the actual joint distribution that is presented in the Baseline. It could be due to

Figure 1.6: Baseline VS Independence: Mixing Behavior in Zero-Correlation Tasks



Notes: This figure is based on the choice allocations in Task 2 of each treatment with two blocks combined. Task 2 in the Baseline treatment demonstrates the joint distribution of two options with zero correlation, and Task 2 in the Independence treatment employs corresponding frame as Task 2 in the Baseline. The error bars depict 95% confidence intervals.

Table 1.15: Baseline VS Independence: Mixing Behavior in Zero-correlation Tasks

	(1) <i>mixed</i>	(2) <i>exact PM</i>
IvsB(Baseline=1)	-0.141** (0.0676)	-0.0990* (0.0541)
Constant	0.331 (0.237)	0.208** (0.101)
Observations	1368	1368

Note: Results from OLS regression with observations in Task 2 and pooling Block 1 and Block 2 together. The dependent variable takes the value of 1 if the allocation choice in a task is classified as (1) *mixed*, or as (2) *exact PM*. IvsB is the dummy variable on whether the task is in the Baseline VS Independence treatment. It takes the value of 1 if the task is in the Baseline. The regression also includes categorical variables of probability categories, gender, and school year, as well as indicator variables of dominant color and STEM as controls. Standard errors are clustered at the subject level and listed in parentheses. Full regression results can be found in Appendix A.6.1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the computational difficulty in thinking through all possible joint outcomes and correctly calculating the associated probabilities. That is, given $(A : p, B : 1 - p)$, subjects might have different weights assigned on each joint outcomes:

$$\sigma(p^2) \circ (A : \$M, B : \$0) \oplus \sigma((1 - p)^2) \circ (A : \$0, B : \$M) \oplus \sigma(p(1 - p)) \circ (A : \$M, B : \$M) \oplus \sigma(p(1 - p)) \circ (A : \$0, B : \$0)$$

with the subjective weight $\sigma(\cdot)$ where $\sigma(x) \neq x$. This observation cannot be explained by any existing models of *Correlation-Sensitive Stochastic Choice*, because none of them account for the possibility that subjects might use misperceived correlation to make decision. In other world, subjects take into account the correlation structure when making decisions, but their perception of the joint distribution between options with zero correlation differs from the actual one. Existing literature defines correlation neglect as individuals' tendency to ignore the correlation between options by treating them as if there is no correlation between them (Enke and Zimmermann, 2019; Eyster and Weizsäcker, 2011). This evidence suggests that subjects might also have imprecise perception on the zero correlation, as one of the correlation structures. Investigating whether the difference is in fact due to the misperceived zero correlation or because of other confounds, theoretically and empirically, would be a promising direction for future research.

Moreover, each theory in the models of *Correlation-Sensitive Stochastic Choice* comes with their own set of concerns when considered as the underlying mechanism behind observed behavior. For instance, the irrational diversification model (Baltussen and Post, 2011; Rubinstein, 2002), which assumes that subjects incorrectly believe they will be paid for all the tickets instead of one randomly selected ticket, is somewhat unsatisfying as an explanation for the observation given the fact that substantial efforts are made in designing the instructions and interfaces to ensure that subjects correctly understand that only one ticket would get paid. In the experiment, the instruction explicitly emphasizes that only one ticket gets paid, tests subjects' understanding regarding this matter via

comprehension checks, and underscores this stipulation in the feedback provided after each decision. Thus, it is reasonable to assume that the irrational diversification model exerts a limited impact on subject's mixing behavior.

There are some concerns raised when considering the evolutionary foundation proposed by Brennan and Lo (2012) as the primary explanatory mechanism behind these findings. Evolutionary explanations posit that with a sufficient number of trials with feedback, which allow individuals to learn the joint distribution, their decisions will eventually converge to matching with the probability of occurrence, thereby rendering their mixing behavior sensitive to the correlation between options. A few questions arise in this regard. Firstly, the decision problem with unique features (same probability category, same dominant color, and same correlation parameter) only repeats twice during the entire experiment. Would this repetition be sufficient for meaningful learning to occur? Secondly, the observed responsiveness to the correlation change in Block 1 of the Baseline treatment cannot be justified by the evolutionary model, as each decision problem in Block 1 is distinct from the others and subjects only receive repeated trials in Block 2. Hence, suggesting new theoretical frameworks, especially those capable of comprehensively explaining and accommodating these observations, or conducting tests to determine which existing model in the class of *Correlation-Sensitive Stochastic Choices* better explain these results, would be a fruitful direction for future research.

Last but not least, these results could serve as a potential explanation for why some interventions aimed at reducing PM are effective while others are not. Numerous interventions have been proposed and studied across various contexts, while the evidence on their effectiveness remains inconclusive. For example, Schulze et al. (2019) fail to replicate previous findings by Wolford et al. (2004) that PM decreases when subjects have extra cognitive load. On a related note, Martínez-Marquina, Niederle and Vespa (2019) find that eliminating uncertainty about which state would occur lets subjects be 8.7%

less likely to make *mixed* choices and 5.4% less likely to make *exact PM* choices, which is smaller than the estimated impacts of correlation (16% - 22%) in this paper. One common feature shared by previous studies is that: the two options under consideration are perfectly negatively correlated (Agranov and Ortoleva, 2017; Martínez-Marquina, Niederle and Vespa, 2019; Vulkan, 2000). My findings could provide a more fundamental explanation for these phenomena: the correlation between options might interact with these interventions, potentially contributing to mixed evidence regarding their effectiveness. For example, even when uncertainty is removed, subjects in Martínez-Marquina, Niederle and Vespa (2019)'s study may still consider the relation between options and mistakenly believe that there is an opportunity to hedge, which might reduce the effectiveness of uncertainty reduction. Further theoretical and empirical studies along these lines could be a promising avenue for future investigations.

1.6 Conclusion

Individuals tend to switch between options or even match precisely to the probability of occurrence when predicting which of two payoff-relevant outcomes that differ only in their probabilities of occurrence, which is a phenomenon referred to as “probability matching.” In this paper, I experimentally study the origin of probability matching by unpacking existing theories and reclassifying them according to three categories: (1) models of *Correlation-Invariant Stochastic Choice*, which includes most preference-based and some heuristics-based models, argue that people mix due to factors orthogonal to the correlation between options and to the framing of those options such as non-Expected Utility preferences or errors, etc.; (2) models of *Correlation-Sensitive Stochastic Choice*, containing several heuristics and biases, posit that people mix due to some heuristics that are sensitive to how options are jointly determined in each state, for instance, people

deliberately use mixing as a tool to hedge against misperceived risk; and (3) *Framing Effects*, assumes that individuals mix because they use some framing-sensitive heuristics to simplify the comparison of marginal distribution between options, and with certain frames, they attend to dissimilar but irrelevant attributes (outcome differences), while neglecting relevant attributes (probability differences).

I find that the vast majority of subjects take into account the interdependence between options in each state of the world when making decisions. In response to the perfectly negative correlation between options in the CPS due to the misperceived hedging opportunity, subjects mix between options or even matching precisely to the outcome probabilities. Furthermore, I observe that although mixing behavior is robust to learning, learning amplifies subjects' responsiveness to the correlation change: with some experience, subjects are more responsive to changes in the correlation between options. I also discover that the decision to mix between the dominant and dominated options are significantly different from the decision to match exactly to the probability of occurrence.

My results highlight a number of areas for further research. First, it would be intriguing to explore the role of (misperceived) correlation between options in other domains that likewise observe seemingly "suboptimal" stochastic choice. As previously mentioned, existing models, such as correlation-sensitive preference (Lanzani, 2020), have limited predictive capabilities when it comes to stochastic choice. This holds true not only in the specific context of this paper but also in other domains where no dominance relation exists, as correlation-sensitive preference pertains to preference over reduced lottery. Without additional assumptions, this theory posits that, if one option is preferred over the other, it is always optimal to always choose the preferred one than to randomize between them, and people mix between options only when they are indifferent. Investigating how stochastic choice varies with changes in the correlation between options, especially in domains without dominance relations, could provide greater insights into

Correlation-Sensitive Stochastic Choice. Second, it would be interesting to expand the scope of this study to investigate the general distinction between state-wise dominance and first-order stochastic dominance in different choice environments such as one-shot binary choice, convex budget set, or repeated choice environment. This could shed light on the foundations of FOSD violation. Additionally, it would be worthwhile to explore when the *Framing Effects* works to reduce FOSD violation and why it is not robust across choice environments. Such an investigation could help us gain a better understanding of how the choice environment prompts individuals' violation of FOSD when making decisions.

Chapter 2

Preference for Sample Features and Belief Updating

Joint work with Menglong Guan, ChienHsun Lin, and Ravi Vora

2.1 Introduction

Different sources, such as the media, government reports, and scientific studies, often emphasize distinct statistical characteristics of the raw data about the same event, which we call sample features, to inform and influence public opinions. This requires people to interpret and incorporate the information conveyed by certain sample features for decision-making. For example, individuals who subscribe to different newspapers adjust their beliefs about a politician's favorability based on the specific statistical characteristics of the same poll results emphasized by their respective newspapers. Similarly, investors receiving financial reports from different analysts need to modify their beliefs according to the specific sample features of the same stock outcomes emphasized by the analyst whose report they receive. During the 2020 United States presidential election, some

media emphasized that Biden won Georgia by a narrow margin of 0.23% (49.47% versus 49.24% between Biden and Trump), while others highlighted the significant difference in the number of votes (12,284).¹

An important question is how people employ and perceive the usefulness of different sample features embedded in the realized signals (raw data) for belief updating, which we know surprisingly little about.² While there could be various reasons from the supply side as to why different sample features are adopted, it is essential to understand the demand side: Are people better at using certain features than others? Do they perceive some features as more useful than others? Are they sophisticated about their biased use, if present?

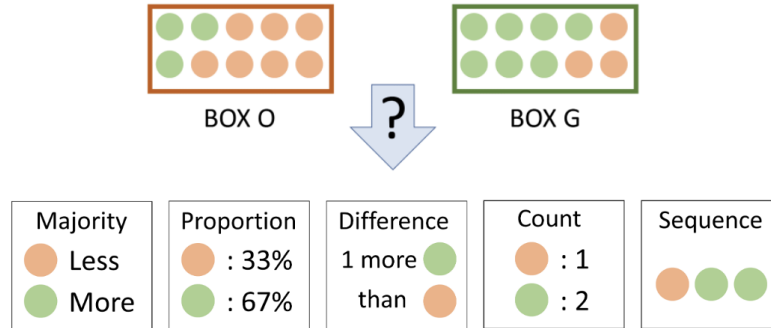
On the one hand, highlighting different sample features might not matter if people are equally good at processing each sample feature. As presumed by standard rational models, people make statistically optimal use of the information conveyed by each sample feature through Bayesian updating. On the other hand, behavioral factors can influence how effectively people use information in sample features to update their beliefs. For instance, when predicting the election winner based on a poll result, individuals could have benefited from more informative sample features, such as observing all the votes in a poll, but struggling to do so when presented with less informative alternatives, such as only knowing the relative frequency of the votes received by the poll winner.³ For instance, if individuals know that there are 10,000 votes in total and the winner got 7000 votes, they learn that this is strong evidence indicating a high likelihood of the winner

¹Sources: CBS News. (2020, August 6). *‘Biden has edge in North Carolina and race is tight in Georgia — CBS News Battleground Tracker poll’* and Staff, A. 11Alive.com (2020, November 9). *‘Blog: Joe Biden’s Georgia lead widens to more than 12,000’*.

²While there is a large literature studying belief updating, it focuses on how people update beliefs when receiving information about the realized signals with most sample features presented (Benjamin, 2019).

³The informativeness of a sample feature is defined as how much more uncertainty about the payoff-relevant state is reduced by using the sample feature to update beliefs, compared to the no-information case. See Section 2.3 for details.

Figure 2.1: “Balls-and-Boxes” Task and Five Sample Features



Note: Existing literature studies belief updating by employing the “Balls-and-Boxes” task with *Count* or *Sequence* provided. We use a novel design by separating the sample features in *Count* into *Majority*, *Proportion*, and *Difference*.

winning the election as well. However, if they only know that the winner received 70% of the votes in the poll, without knowing the size of the poll, it becomes challenging to determine whether this is strong or weak evidence. Individuals must additionally account for this uncertainty when making inferences. This additional step of consideration could be cognitively taxing and affect how effectively they utilize the information.

In this paper, we use a laboratory experiment to study these questions. Start with the widely used “balls-and-boxes” task by existing literature for studying inference from symmetric binary signals about a binary state (Benjamin, 2019), as shown in Figure 2.1. One of the two boxes is randomly selected with equal chance. Each box has ten balls, seven of which match the color of the corresponding box, while the remaining three match the color of the alternative box. That is, $\Pr(\text{One green ball}|\text{Box G}) = \Pr(\text{One orange ball}|\text{Box O}) = 70\%$. The subjects’ objective is to assess the probability that the picked box is Box G versus Box O, and gets paid by Binarized Scoring Rule (Hossain and Okui, 2013). As a clue, a sequence of balls is drawn out of the chosen box with replacement. Existing studies on belief updating either use *Count* or *Sequence* (as illustrated in Figure 2.1) to inform subjects about drawn balls.

To answer the questions of our interest, instead of directly showing the sequence of drawn balls, we propose a novel experimental design where we use five reports to separate representative sample features extracted from the information about drawn balls. The five reports are (1) *Majority*: indicates whether the set of drawn balls has more green or more orange balls; (2) *Proportion*: displays the relative frequencies of green and orange balls among the drawn balls, respectively; (3) *Difference*: demonstrates the difference in the absolute frequency of green and orange balls among the drawn balls; (4) *Count*: illustrates the absolute frequencies of green and orange balls among the drawn balls, respectively; (5) *Sequence*: depicts the original sequence in which the balls were drawn.

Among these reports, we employ *Sequence* and *Count* to replicate the findings documented in the existing literature. *Sequence* contains all the sample features of the realized signals. From *Sequence* to *Count*, the information on the order of realized signals is excluded, which is not useful for Bayesian inference.⁴ We use *Difference*, which is the sufficient statistics of information about realized signals for Bayesian inferences in (symmetric) inference problems (Benjamin, 2019). From *Count* to *Difference*, the information on the sample size is not provided, which is not instrumental for Bayesian inference in (symmetric) inference problems. By comparing across *Difference*, *Count* and *Sequence*, we can examine the extent to which non-instrumental features matter and how agents perceive their usefulness. We use *Proportion* to isolate the “Strength” (sample proportion) from the “Weight” (sample size), as defined in the “Strength-Weight bias” by Kahneman and Tversky (1972).⁵ Without the information about “Weight,” *Proportion*

⁴Instrumental value of a report is defined as the expected payoff that a Bayesian agent can receive by using it to update beliefs in “balls-and-boxes” task, compared with the case with no information. In our setting, informativeness and instrumental value give the same prediction of the ordinal rankings among the five reports. Thus, we use informativeness (informative) and instrumental value (instrumental) interchangeably. See Section 2.3 for more details.

⁵“Strength-Weight bias” describes the bias that individuals tend to over-weight sample proportion (“Strength”) while under-weighting sample size (“Weight”) when using *Sequence* or *Count* to update beliefs in “balls-and-boxes” tasks. These studies exogenously manipulate sample proportion and sample size embedded in *Sequence* or *Count*, and structurally estimate the coefficients on sample size and on

is less informative than *Difference*, *Count*, and *Sequence*. *Majority* is the least informative feature among the five. Comparing across *Sequence/Count/Difference*, *Proportion*, and *Majority* allows us to study how the updating behaviors respond to the change in the informativeness of sample features.

The experiment consists of two parts. Part 1 uses a ranking-cards method inspired by Dustan, Koutout and Leo (2022) to elicit subjects’ *willingness-to-pay* of receiving each of the five reports in the “balls-and-boxes” task. It allows us to measure the perceived usefulness of each feature. In Part 2, we employ the strategy method with 33 pre-selected scenarios of the “balls-and-boxes” task. These scenarios are designed to capture how subjects respond and adjust their beliefs based on various signal realizations and different information conveyed by different reports.

We have two main findings regarding how well subjects *use* different reports when updating beliefs. These observations are robust to different measures of performance: average absolute deviation from the Bayesian benchmark and estimated responsiveness to information change using the Grether (1980) model. Firstly, subjects’ belief updating deviates from the Bayesian benchmark under each report. However, it is least severe under *Proportion*, despite *Proportion* being less informative compared to *Difference*, *Count*, and *Sequence*. It suggests that subjects are better at using the “Strength” (sample proportion) when used alone, rather than when combined with “Weight” (sample size). Secondly, among the reports that are equally informative, i.e., *Difference*, *Count*, and *Sequence*, subjects’ belief updating is closer to the Bayesian benchmark when using *Count* and *Sequence*, compared to *Difference*. Our findings indicate that subjects are not equally good at processing each sample feature, contrasting to what the Bayesian benchmark suggests. Moreover, the biased use does not monotonically improve with the informativeness

sample proportion, respectively. By testing whether the two coefficients are identical and equal to one, the common finding is that the coefficient on sample proportion is significantly larger than that on sample size, and both are less than one (Benjamin, 2019).

of sample features.

In terms of *perceived usefulness*, we find that, on average, the perceived usefulness of the features deviates from the predictions of instrumental value in two ways. First, there is no significant difference in the average *WTP* among *Proportion*, *Count*, and *Sequence*, despite the latter two features being more instrumentally useful than *Proportion*. Second, on average, subjects assign a significantly higher value to *Proportion/Count/Sequence* by a margin of \$0.68, compared to *Difference* or *Majority*, even though the former three features have the maximum instrumental value. These results suggest that subjects fail to fully recognize the usefulness of other features, such as *Difference* and sample size, even though incorporating either of them with *Proportion* increases the instrumental value of information.

These findings suggest that subjects, on average, have a strong preference for sample features that contain *Proportion* compared to those that do not. Features that contain *Proportion*, i.e., *Count* and *Sequence*, require subjects to conduct some calculations to get the proportion information. Features that do not contain *Proportion*, i.e., *Difference* and *Majority*, require additional inference about all the potential sample proportions that could lead to the same *Difference* or *Majority* information, along with more difficult calculations. The increased difficulties of inference and calculation required to get the proportion information might lead to the distaste for *Difference* and *Majority*.

Examining the association between subjects' perceived usefulness and the actual use of the five sample features, we observe that, on average, subjects are self-consistent between their preferences and performances, making better use of the sample feature they prefer. This finding suggests that the biased use of sample features in belief updating is more likely to be an intentional deviation rather than a result of inattentive heuristics. However, there is also non-negligible inconsistency between preferences and performances, and the most prominent pattern is that some subjects prefer a report that contains more

or more informative features than another but perform relatively worse under it. In each possible pairwise comparison of reports, among subjects whose preference for and performance with the two reports, a non-negligible inconsistency between preferences and performances, and the most prominent pattern is that some subjects prefer a report that contains more or more informative features than others are ordinally inconsistent, over 60% of them follow this pattern. It indicates that a significant portion of subjects tend to prioritize quantity (as many features as possible) over relevance (how useful they are in the actual task) while failing to take into account the cost of processing more features than necessary.

Our study is related to several strands of literature. First of all, our findings contribute to the existing literature on belief updating and learning. We are the first to show direct evidence of how subjects use and perceive the usefulness of sample features for belief updating. Most previous studies demonstrate the biased use of sample features based on indirect evidence and structural estimation. They identify “Strength-Weight bias” or “Sample Size Neglect,” by asking subjects to update beliefs with either *Count* or *Sequence* adopted to convey the information about realized signals (Griffin and Tversky, 1992).

By estimating the coefficients on sample size and sample proportion, respectively, they find that the weight on sample size is smaller than that on sample proportion.⁶ Kraemer and Weber (2004) studies how the presentation mode of the signals affects belief updating by comparing realized signals and *Proportion* plus sample size. They find that subjects’ focus on sample proportion is pronounced when they receive explicit information regarding sample proportion plus sample size compared to when receiving realized signals. When most sample features are available, it is challenging to discern whether the biased weights result from the different abilities in utilizing each feature or

⁶See Benjamin (2019) for the meta analysis.

from the inclusion of too many sample features.

We add to this literature by presenting direct evidence that individuals are not equally good at processing each sample feature embedded in the realized signals, and they value the usefulness of sample features differently from instrumental value. Specifically, we find that subjects are better at processing sample proportion alone, compared to more informative features or those with other features combined. Furthermore, we demonstrate that these biases are more likely to be intentional deviations rather than the result of inattentive heuristics.

Second, our study contributes to the existing literature that examines the impacts of coarse versus precise information. Ravaioli (2021) investigates how the coarsening of food labels affects the number of calories consumed in food choices. He proposes a bounded rationality model with precision overload to explain his main finding: coarse-categorical labels reduce the number of calories consumed in food choices. As a complement to his study, we provide direct evidence that, even in an abstract learning environment, individuals are worse at processing detailed information when all sample features are included, compared to coarse information with certain features excluded. We also show that not all forms of simplification work. Both *Difference* and *Proportion* contain a reduced number of sample features, yet subjects perform worse with *Difference* compared to *Proportion*, despite the former having a higher instrumental value. Our results suggest that the perceived usefulness may play a role in determining the effectiveness of coarse information: if the coarse information emphasizes a sample feature that individuals consider useful, they are more likely to make better use of it when updating their beliefs.

Third, our study is related to the demand for information literature. There is a growing literature on how people choose and evaluate information with instrumental value (Ambuehl and Li, 2018; Charness, Oprea and Yuksel, 2021; Guan, Oprea and Yuksel,

2023; Liang, 2023).⁷ Among them, the most closely related to our study is Ambuehl and Li (2018), which connects the under-responsiveness to instrumental value in information evaluation with the non-Bayesian use of information. We also find people’s evaluation of information broadly aligns with how well they use the information from the Bayesian perspective. In addition, our finding of people performing better with *Proportion* and overvaluing *Proportion* suggests that the non-Bayesian use of information could lead to more severe deviations from instrumental value than under-responsiveness in the demand for information.

The remainder of the paper is organized as follows. Section 2.2 describes the experiment design. Section 2.3 lists theoretical predictions. Section 2.4 presents results. Section 2.5 concludes by discussing the implications of our main findings.

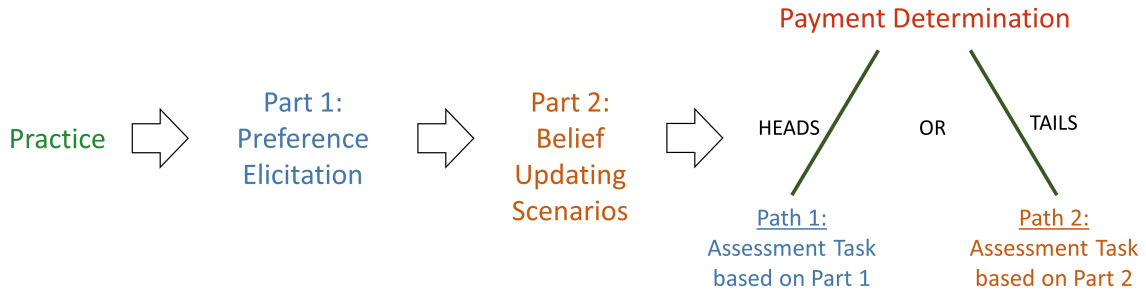
2.2 Experimental Design

We design the experiment to investigate how subjects use and perceive the usefulness of various sample features of realized signals in belief updating. To accomplish this, the experiment consists of two parts: (1) ex-ante preference elicitation; (2) belief-updating scenarios. Figure 2.2 demonstrates the experimental procedure. It starts with an introduction to the “balls-and-boxes” belief updating task, namely *Assessment Task*, and the five reports subjects may receive. This is followed by two practice rounds without feedback. Then, in Part 1, we elicit the subjects’ preference regarding the five reports. In Part 2, we use the strategy method to gauge how subjects employ the information provided for belief updating across 33 pre-selected scenarios of the *Assessment Task*.

⁷There is also a large literature focusing on non-instrumental information and showing people’s demand for information could be driven by timing preference of uncertainty resolution (Nielsen, 2020), preference for positive skewness (Masatlioglu, Orhun and Raymond, 2017), curiosity or motivated attention (Golman and Loewenstein, 2018; Golman et al., 2022), anticipatory feelings (Caplin and Leahy, 2001), etc.

Subjects face the *Assessment Task* after finishing Part 2. One of the two parts is randomly selected for payment, and subjects’ decisions in the chosen part determine their final payments in the *Assessment Task*.

Figure 2.2: Timeline of the Experiment



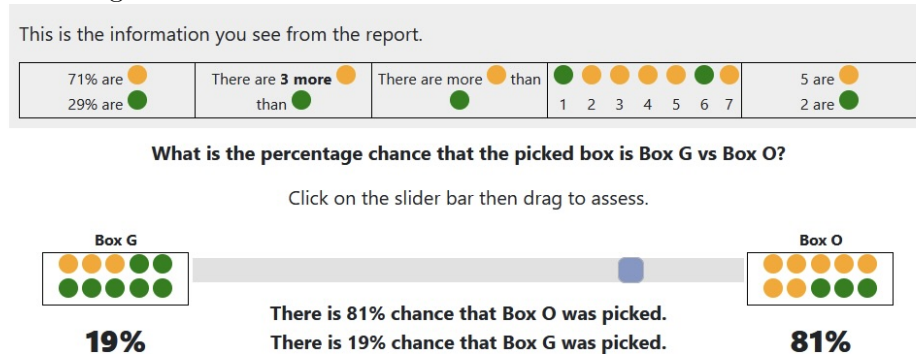
The rest of this section describes the components of the experimental design in detail. First, we outline the basic setups of the belief updating task, *Assessment Task*, and the five reports of the realized signals. Then, we demonstrate how we elicit preferences regarding the five reports and performances in the belief updating scenarios. Lastly, we discuss the choices of experimental design.

2.2.1 The “Assessment Task”

To measure how subjects use information to update beliefs, we use the stylized balls-and-boxes setting. This setting involves two boxes, each containing ten balls. *Box G* consists of seven green balls and three orange balls, while *Box O* consists of three green balls and seven orange balls. The computer randomly selects a box with equal probability. Thus, the state of the world ω is either *O* or *G*. Then, the computer independently draws balls out of the chosen box with replacement.⁸ Subjects do know which box is selected, and are asked to assess the likelihood of the selected box being Box O or Box G. This

⁸Therefore, the diagnostic rate – the likelihood of drawing a ball from the box that matches the color of the box itself – is symmetric: $P(\text{one green ball}|\text{Box } G) = P(\text{one orange ball}|\text{Box } O) = 0.7$.

Figure 2.3: Screenshot of Assessment Task: Practice Round



process of forming posterior belief is referred to as the *Assessment Task* and serves as the basis for determining the subject’s likelihood of receiving the \$10 bonus after completing Parts 1 and 2.

The computer randomly draws N balls from the chosen box with replacement, where N is a random number selected from $\{3, 5, 9, 15\}$ with equal probabilities. We use $S = (s_1, \dots, s_N)$, where for each ball, $s_n \in \{o, g\}$ with $n \in \{1, 2, \dots, N\}$, to denote the sequence of drawn balls. Instead of directly observing the exact sequence of drawn balls S , subjects receive a summary of the sequence through one of the five reports, denoted as γ_R . The report, γ_R , maps the sequence of drawn balls (S) to a statistical feature of S represented by report R , denoted as $\gamma_R(S) := S_{\gamma_R}$. Different reports capture different features of the drawn balls: (1) Sample Majority, denoted as *Majority* γ_M —“Are there more green or orange balls in the sample?”; (2) Sample Proportion, denoted as *Proportion* γ_P —“What is the fraction of green balls in the sample?”; (3) Sample Difference, denoted as *Difference* γ_D —“How many more green (orange) balls are there in the sample?”; (4) Sample Count, denoted as *Count* γ_C —“What are the total numbers of orange and green balls in the sample, respectively?”; (5) Sample Sequence, denoted as *Sequence* γ_S —“What is the sequence of drawn balls?”. Figure 2.3 shows the interface of *Assessment Task* that subjects see during the practice round. Each hypothetical scenario task in Part 2, as well

as the final *Assessment Task*, employs a similar interface. However, it should be noted that subjects are presented with a maximum of one report at a time.

To ensure incentive compatibility of posterior elicitation in the *Assessment Task*, we use the Paired-Uniform Scoring Method introduced in Wilson and Vespa (2018) as it elegantly sidesteps the need for detailed technical explanations.⁹ Although we explain the payment determination logic to the subjects, we explicitly emphasize that it is in their best interest to report their true beliefs.

2.2.2 Part 1: Preference Elicitation

Our design aims to identify both the cardinal and ordinal rankings of subjects' preferences regarding the set of reports. To achieve this, we employ a ranking-cards method whereby each subject is required to place five *Report* cards, one for each report, within an ordered list of 20 *No Report + Money* cards.¹⁰

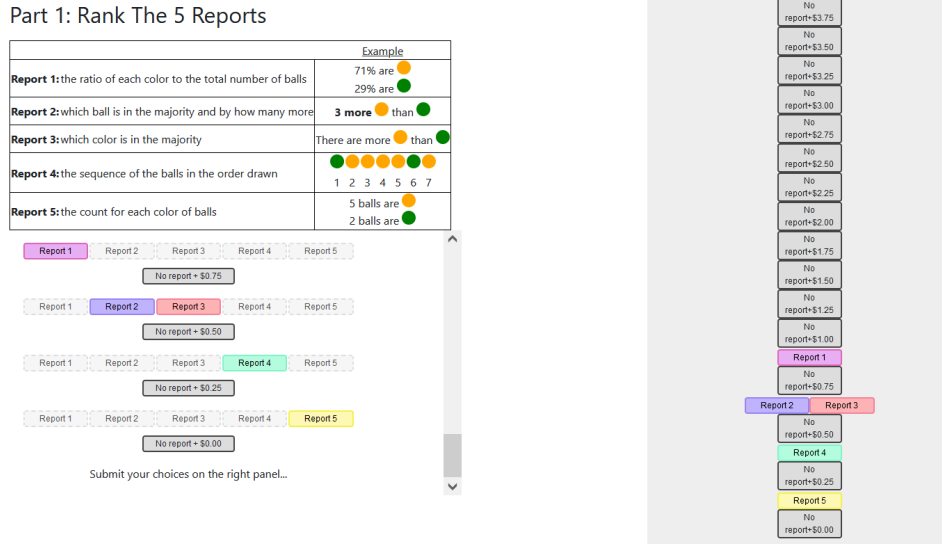
For the *No Report + Money* cards, the dollar value ranges from \$5 to \$0, descending in increments of \$0.25. To incentivize subjects to rank the cards according to their true preferences, subjects are told that, if Part 1 is randomly chosen for payments, the computer would randomly select two cards from the set of 25. The higher-ranked card would then be designated as the report that they would receive to summarize the information about the drawn balls in the *Assessment Task*.¹¹

⁹The Paired-Uniform Scoring Method is equivalent to the commonly exploited (incentive compatible) belief elicitation method, *Binary Scoring Rule (BSR)*. In the binary scoring rule, the subjects are paid according to the squared distance to the actual belief. Specifically, let p be the subject's actual belief that the true state $\omega = O$ (and $1 - p$ be the belief that $\omega = G$), and a be the *stated* belief. Then the subject will be informed of the realized state: when the realized state is $\omega = O$, the payoff is $1 - (1 - a)^2$; when when the realized state is $\omega = G$, the payoff is $1 - a^2$. Hence the expected payoff given the stated belief a is $p(1 - (1 - a)^2) + (1 - p)(1 - a^2)$. One can show that the expected payoff is maximized when $a = p$.

¹⁰This method is incentive compatible for expected utility maximizers. See Appendix B.1 for details.

¹¹For additional details about the ranking-card method and its incentive compatibility, please refer to Appendix B.1. The method is inspired by Dustan, Koutout and Leo (2022) but is different from theirs to some extent. In ours, subjects rank multiple object cards simultaneously, then two cards are randomly drawn and the one ranked higher is implemented. In Dustan, Koutout and Leo (2022), subjects insert

Figure 2.4: Screenshot of Ranking-Card Preference Elicitation Over Reports

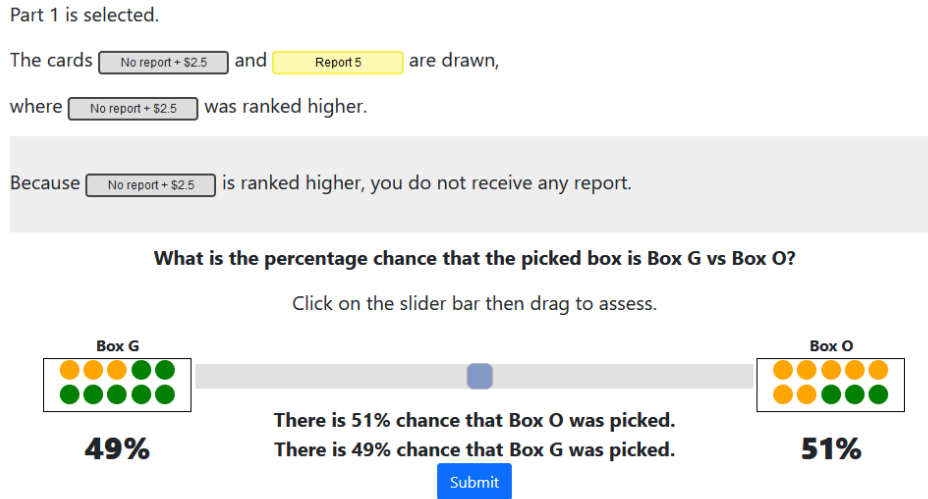


We use the same payoff method explained previously to determine subjects' final payments based on their stated beliefs in the *Assessment Task*. If the higher-ranked card is a *Report* card, denoted as γ_R , subjects will complete the *Assessment Task* with the information about the drawn balls summarized by the corresponding report, S_{γ_R} . On the other hand, if the higher-ranked card is a *No Report + Money* card, subjects will finish the *Assessment Task* without any information about the drawn balls. In addition to the payment received from the task, they will also receive the monetary compensation specified on the card. Figure 2.5 depicts an example of the *Assessment Task* when Part 1 is selected for payment and the *No Report + Money* card is ranked higher.

2.2.3 Part 2: Belief Updating Scenarios

We employ the strategy method to measure subjects' performances across 33 pre-selected scenarios of the *Assessment Task*. To be more specific, after subjects state their an object card into a list of lottery cards, then a lottery card is randomly drawn. The object card will be implemented if it is ranked higher than the drawn lottery card; otherwise, the lottery card is implemented.

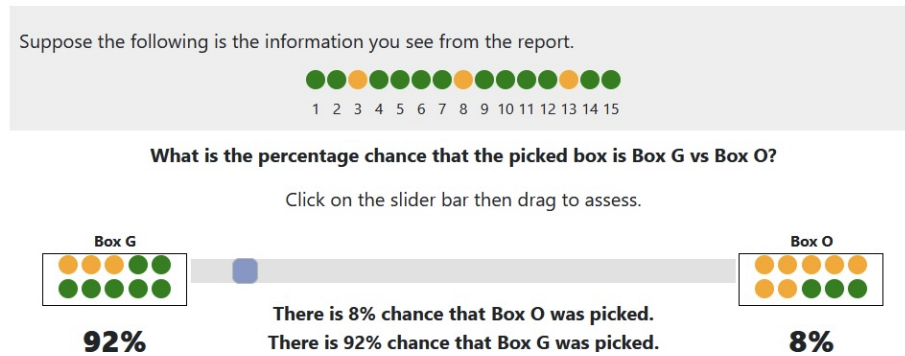
Figure 2.5: Screenshot of the *Assessment Task* when Part 1 is Selected for Payment
The Assessment Task



preferences for the five reports, they proceed to complete the hypothetical *Assessment Task* for the set of 33 predetermined scenarios. Figure 2.6 is an example of it.

In each scenario, subjects are presented with one report and are asked to state their posterior beliefs. If Part 2 is selected for payments, in the *Assessment Task*, the computer will check whether the information about the drawn balls, as summarized by report R (S_{γ_R}), matches one of the pre-selected scenarios. If a match is found, the computer will utilize the subjects' stated beliefs from that specific scenario as their posteriors in the *Assessment Task*, to determine their final payments. If there is no match with any pre-selected scenario, subjects need to manually complete the *Assessment Task* by reporting their beliefs via the slider bar. Consequently, subjects have no incentive to provide false posteriors beliefs during Part 2.

Figure 2.6: Example of a Scenario in Part 2



2.2.4 Understanding the Design

We design the experiment to answer two questions: (1) how subjects use different sample features embedded in the realized signals when updating beliefs; and (2) how they perceive the usefulness of the sample features in helping belief updating. Here we highlight the design choices made to facilitate these goals.

First, to cleanly identify how subjects use the sample features embedded in the realized signals, we employ the simple and classical “balls-and-boxes” setting with symmetric prior ($Pr(\text{Box } G) = Pr(\text{Box } O) = 50\%$) and symmetric diagnostic rate ($Pr(1 \text{ green ball} | \text{Box } G) = Pr(1 \text{ orange ball} | \text{Box } O) = 70\%$). The use of symmetric prior and symmetric diagnostic rate serves two purposes in our study. Firstly, it reduces the burden of understanding the belief updating environment, making it easier for participants to comprehend and engage with the task. Secondly, it helps mitigate any potential bias that could arise from sub-optimal utilization of prior information or an asymmetric perception of diagnostic rates. By employing symmetric priors and diagnostic rates, we aim to minimize any distortions in our objective of identifying how subjects utilize the sample features, ensuring a more accurate analysis.¹²

¹²We acknowledge that subjects may exhibit biases in aggregating prior information and the information of realized signals, and their use of sample features may also impact how they aggregate the information in general. Our study focuses on cleanly identifying the use of different sample features as the first step. We leave room for future extensions to explore variations such as asymmetric priors and

Second, we carefully choose five reports to capture representative sample features. Firstly, we use *Count* and *Sequence* as benchmarks to replicate findings from existing literature on belief updating (Benjamin, 2019). Secondly, we employ *Proportion*, which indicates the “Strength” (representativeness of the signals) in the “Strength-Weight bias” or “Sample Size Neglect” described by Kahneman and Tversky (1972), to isolate “Strength” (sample proportion) from “Weight” (sample size). Furthermore, we include *Difference*, which serves as the sufficient statistics of the information about realized signals S for Bayesian inferences in (symmetric) inference problems (Benjamin, 2019).¹³ By directly measuring subjects’ belief updating when presented with one feature at a time, we can explore whether subjects are equally good at processing each feature but struggle when processing the information with multiple features combined. Or alternatively, their abilities to process each feature fundamentally differ and so does their perceived usefulness of each feature. This exploration may shed light on the underlying mechanisms behind biases in belief updating, such as the “Strength-Weight bias” Lastly, we employ *Majority* to maximize variations in sample features with different instrumental values. This enables us to examine the extent to which the informativeness of sample features predicts how subjects use and perceive their usefulness. For a more detailed discussion on theoretical benchmarks, please refer to Section 2.3.

Next, we intentionally select a set of 33 scenarios to achieve two goals: (1) to expose subjects to a representative range of sample outcomes for each of the five reports; and (2) to intentionally obscure the exact number of balls drawn in certain reports. Some reports require additional effort to accurately deduce the complete information about all possible realizations of drawn balls. This deliberate obscurity prompts subjects to invest thoughtful analysis in interpreting the available information, which allows us to assess

asymmetric diagnostic rates, which could provide further insights into these phenomena.

¹³With asymmetric diagnostic rates, $Pr(\text{green ball}|\text{Box } G) \neq Pr(\text{orange ball}|\text{Box } O)$, *Difference* is still more informative than *Proportion*.

the impact of inferential effort on belief updating.

Furthermore, we deliberately choose the set of numbers: $\{3, 5, 9, 15\}$, from which we sample the sample size N , for three reasons. Firstly, we aim to ensure that the Bayesian posteriors, as the benchmark, are uniformly distributed between 0% and 100%. To achieve this, we restrict the maximum number of balls to prevent clustering at the extreme values (0% or 100%). Large sample sizes could otherwise lead to near-certainty Bayesian posteriors, while very small sample sizes would result in minimal variation across reports.¹⁴ Secondly, by selecting odd numbers as the sample size, we avoid situations where the Bayesian posterior equals the prior (50%). This enhances the statistical power of the experiment, as reports that yield 50% posteriors are interchangeable.¹⁵ Thirdly, we select sample sizes with common factors only, with the intention of adding the needs to consider certain information can either be strong or weak evidence. This is because multiple realizations of the balls, whether it is strong or weak evidence, can map to the same information conveyed by certain report $S_{\gamma R}$. Less informative sample features require additional steps to deal with this uncertainty which could be cognitively taxing. It allows us to investigate the extent to which this additional inferential effort predicts subjects' performance across the five reports.¹⁶

Finally, we have set the preference elicitation *before* the belief updating scenarios in order to understand how subjects evaluate the values of each report and predict the

¹⁴For instance, if a subject receives a report stating “67% of balls are orange balls,” having large sample sizes would lead to a near-certainty Bayesian posterior that the selected box is Box O (e.g. a Bayesian posterior of 99.97% for $N = 30$, 98.58% for $N = 15$, and 70% for $N = 3$). With $N = 1$, the Bayesian posterior would be equal to the diagnostic rate: $Pr(\text{Box } G | 1 \text{ green ball}) = 70\% = Pr(\text{Box } O | 1 \text{ orange ball})$, resulting in minimal variation across reports.

¹⁵For example, *Proportion* “50% of balls are orange”— *Count* “same number of balls of different colors”— *Difference* “no difference in the number of balls of different colors” give identical Bayesian posteriors.

¹⁶For instance, consider the report stating “67% of drawn balls are orange.” In this case, there are three equally likely scenarios with different levels of information strength: (1) a sample of two orange balls out of three draws, which would be relatively weak evidence; (2) a sample of six orange balls out of nine draws, which would be the evidence of intermediate strength; or (3) a sample of ten orange balls out of fifteen draws, which would be relatively strong evidence.

usefulness before experiencing the different reports in the belief update tasks. This ordering minimizes the impact of relative frequency on the evaluation, as subjects will be exposed to reports with varying frequencies during the belief updating scenarios.¹⁷

2.3 Theoretical Predictions

In this paper, we focus on two main aspects of the belief elicitation problem: the performance in the updating tasks and the preference over the reports. The following sections will describe the primary predictions of each aspect.

2.3.1 Performances in the Updating Task with Reports

2.3.1.1 Setup and Bayesian Inference

We first discuss the Bayesian benchmark in the updating tasks with reports. We use $\omega \in \{O, G\}$ to denote the state of the world (which box is selected), and the objective prior belief is $P(\omega = G) = \frac{1}{2}$. Given the realized state $\omega \in \{O, G\}$ (selected box), $N \in \{3, 5, 9, 15\}$ and is randomly determined with equal probability and a sequence of N balls are drawn independently with replacement. The drawn sequence of balls is denoted as $S = (s_1, \dots, s_N)$, where for each ball, $s_n \in \{o, g\}$ with $n \in \{1, 2, \dots, N\}$. The diagnostic rates, probabilities that a ball o is drawn from Box O and a ball g is drawn from Box G , are symmetric,

$$p(s_n = o | \omega = O) = p(s_n = g | \omega = G) = \theta = 0.7$$

The *Report*, γ_R , maps the sequence of drawn balls (S) to some statistical feature of the sample S summarized by report R . We denote $\gamma_R(S) := S_{\gamma_R}$.¹⁸ A Bayesian agent

¹⁷By the nature of our design, there is one scenario question under *Majority* and 15 questions under *Sequence*.

¹⁸For example, let $S = (o, o, o, g, g)$. As in our design, with *Majority*, i.e. γ_M , then $\gamma_M(S) =$

forms the posterior belief conditional on the feature of the drawn balls (S) summarized by report R , S_{γ_R} :

$$\frac{p(O | S_{\gamma_R})}{p(G | S_{\gamma_R})} = \frac{p(S_{\gamma_R} | O) p(O)}{p(S_{\gamma_R} | G) p(G)} \quad (2.1)$$

where $\frac{p(O)}{p(G)}$ is the ratio of prior beliefs, $\frac{p(S_{\gamma_R}|O)}{p(S_{\gamma_R}|G)}$ is the ratio of conditional likelihood of receiving S_{γ_R} given state, and $\frac{p(O|S_{\gamma_R})}{p(G|S_{\gamma_R})}$ is the ratio of posterior beliefs. With symmetric prior belief of states O and G , the Bayesian posterior can be reduced to

$$\frac{p(O | S_{\gamma_R})}{p(G | S_{\gamma_R})} = \frac{p(S_{\gamma_R} | O)}{p(S_{\gamma_R} | G)} \quad (2.2)$$

When a Bayesian agent observes the features of S summarized by reports *Sequence*, *Count*, or *Difference*, it is sufficient to use the information about the difference between the numbers of o and g balls in the sequence of drawn balls S to find the Bayesian posterior as shown below:¹⁹

$$\frac{p(O | S_{\gamma_R})}{p(G | S_{\gamma_R})} = \frac{p(S_{\gamma_R} | O)}{p(S_{\gamma_R} | G)} = \frac{\binom{N_o + N_g}{N_o} \theta^{N_o} (1 - \theta)^{N_g}}{\binom{N_o + N_g}{N_g} (1 - \theta)^{N_o} \theta^{N_g}} = \left(\frac{\theta}{1 - \theta} \right)^{N_o - N_g} \quad (2.3)$$

where N_o and N_g are the numbers of o and g in the sequence of drawn balls S , respectively. The Bayesian posterior is a function of the difference in the numbers of o and g balls in the drawn balls S , $N_o - N_g$, and the diagnostic rate, θ .

For reports *Proportion* and *Majority*, however, the drawn balls with different sample size N can map to the same S_{γ_R} . Thus, a Bayesian agent needs to take into account the fact that, given the realized state ω , the likelihood of receiving S_{γ_R} , $Pr(S_{\gamma_R} | Box \omega, N)$,

¹⁹“More o than g ,” with *Proportion*, i.e. γ_P , $\gamma_P(S) = “60\% o \text{ and } 40\% g.”$

¹⁹By sufficient, we mean no additional inference is needed before applying the Bayes’ rule.

varies with the number of drawn balls, N . For instance, when S_{γ_R} says “33% o and 67% g ”, the actual drawn sequence S can be under one of the following equally-likely cases: (1) $N = 3$: 1 o and 2 g , (2) $N = 9$: 3 o and 6 g , or (3) $N = 15$: 5 o and 10 g . Then, she needs to form expected likelihood of S_{γ_R} , given the realized state ω , over all possible N . Thus, we further extend Equation (2.2) into

$$\frac{p(O | S_{\gamma_R})}{p(G | S_{\gamma_R})} = \frac{p(S_{\gamma_R} | O)}{p(S_{\gamma_R} | G)} = \frac{\sum_{N \in \{3,5,9,15\}} p(N)p(S_{\gamma_R} | O, N)}{\sum_{N \in \{3,5,9,15\}} p(N)p(S_{\gamma_R} | G, N)} \quad (2.4)$$

where $P(N) = \frac{1}{4}$. Note that each $p(S_{\gamma_R} | O, N)$ can be found with the same method as in Equation (2.3).

2.3.1.2 Empirical Strategies and Hypotheses

We use two ways to evaluate how well agents use sample features when updating beliefs. On the one hand, we measure the absolute distance between agents’ stated posteriors and Bayesian posteriors. For a Bayesian agent, it maximizes her expected payoff by reporting the Bayesian posteriors, and there is no difference across sample features. That is, a Bayesian agent always makes the best of each sample feature. If the stated posterior deviates less from the Bayesian posterior under one report compared to another, we say that the agent performs better under the former than the latter one.

On the other hand, we follow Grether (1980)’s framework of the balls-and-boxes paradigm to measure how responsive agents are towards the change in the likelihood ratio of receiving S_{γ_R} given state ω .²⁰ Grether (1980)’s framework distinguishes the biases in using realized information from those in incorporating the prior belief by adding

²⁰It refers to the ratio of the likelihood of receiving S_{γ_R} conditional on the state, $\frac{p(S_{\gamma_R}|O)}{p(S_{\gamma_R}|G)}$.

parameters c and d to Equation (2.1) respectively

$$\frac{\pi(O | S_{\gamma_R})}{\pi(G | S_{\gamma_R})} = \left(\frac{p(S_{\gamma_R} | O)}{p(S_{\gamma_R} | G)} \right)^c \left(\frac{p(O)}{p(G)} \right)^d \quad (2.5)$$

where $\pi(\cdot | S_{\gamma_R})$ represents the subjective posterior conditional on receiving S_{γ_R} . As $p(O) = p(G)$ in our setting, the last term becomes 1, and therefore the subjective posterior becomes a function of the likelihood ratio of the signal realizations with parameter c . By taking logarithm, we have

$$\ln \left(\frac{\pi(O | S_{\gamma_R})}{\pi(G | S_{\gamma_R})} \right) = c \ln \left(\frac{p(S_{\gamma_R} | O)}{p(S_{\gamma_R} | G)} \right) = c \ln \left(\frac{p(O | S_{\gamma_R})}{p(G | S_{\gamma_R})} \right) \quad (2.6)$$

where the coefficient c measures how responsive agents are towards the change in the likelihood ratio of S_{γ_R} . A Bayesian agent has $c = 1$ in each report. $c < 1$ corresponds to updating as if S_{γ_R} provided less information about the state than it actually does (under-inference). The lower the c , the less sensitive agents are to the change, and thus the more severe under-inference. $c > 1$ means updating as if S_{γ_R} was more informative than it actually is (over-inference). The last equality follows from Equation (2.2). Specifically, we estimate the following regression model:

$$\ln \left(\frac{\pi(O | S_{\gamma_R})}{\pi(G | S_{\gamma_R})} \right) = a + c \ln \left(\frac{p(O | S_{\gamma_R})}{p(G | S_{\gamma_R})} \right) + \gamma \mathbf{X} + \varepsilon \quad (2.7)$$

where \mathbf{X} is the vector of demographic variables added as controls; a is the constant term and ε is the residual. If the estimated c from stated beliefs under some report is closer to 1 than the others, we would say that subjects perform better with the former report than the latter one.

2.3.2 Preference over Reports

2.3.2.1 Instrumental Value of Reports

We use two ways to measure the instrumental value of the reports. On the one hand, we evaluate the instrumental value of the reports by *how much the report can improve the expected payoff in the belief updating task*. Let $\mathcal{S}(\gamma_R)$ be the set of possible realizations under γ_R . As we employ the binary scoring rule (BSR) for payment, a Bayesian agent maximizes the expected payoff by reporting the Bayesian posterior given realized S_{γ_R} . Thus, the expected payoff of γ_R is

$$EP(\gamma_R) = B \cdot \sum_{S_{\gamma_R} \in \mathcal{S}(\gamma_R)} [p(O|S_{\gamma_R})(1 - (1 - p(O|S_{\gamma_R}))^2) + (1 - p(O|S_{\gamma_R}))(1 - p(O|S_{\gamma_R})^2)] p(S_{\gamma_R})$$

where $B = \$10$ is the size of the bonus, and $p(S_{\gamma_R})$ is the likelihood of receiving S_{γ_R} given γ_R . Note that without any information, the agent knows the prior only. Thus, the instrumental value is defined as the difference in the expected payoff between receiving γ_R and receiving no information:

$$V(\gamma_R) = EP(\gamma_R) - EP(P_0)$$

where $EP(P_0)$ denotes the expected payoff without the information. In our setting, for example, the prior is $P_0 = 50\%$. So the optimal guess (50%) yields the expected payoff \$7.5:

$$EP(P_0) = 10 \times [0.5(1 - (1 - 0.5)^2) + (1 - 0.5)(1 - 0.5^2)] = 10 \times 0.75.$$

If the agent receives report *Majority*, the information will increase the expected payoff to \$8.85. Thus the (expected) instrumental value of *Majority* is $\$8.85 - \$7.5 = \$1.35$.

Moreover, another widely-used measure of the usefulness of information is the reduction of the Shannon entropy (Shannon, 1948), or *informativeness* (Cabrales, Gossner and Serrano, 2013). That is, compared to the no-information case, how much more uncertainty is reduced by receiving the information about the drawn balls summarized by γ_R . Specifically, given $\omega \in \Omega = \{O, G\}$ and the probability measure $p : \Omega \rightarrow [0, 1]$, the Shannon entropy is

$$H(p) = - \sum_{\omega \in \Omega} p(\omega) \log_2 p(\omega).$$

Let $q(S_\gamma)$ be the probability that the realized S_{γ_R} is generated under report γ_R . Then informativeness is defined as the Shannon mutual information between prior and posterior beliefs

$$I(\gamma_R) = H(p_0) - \sum_{S_{\gamma_R} \in \mathcal{S}(\gamma_R)} q(S_{\gamma_R}) H(p_{S_{\gamma_R}}).$$

Table 2.1 demonstrates the informativeness of each report. Note that when there is no report, the informativeness is 0.

2.3.2.2 Hypotheses

Table 2.1 summarizes the instrumental value of the five reports measured by two definitions discussed above. Note that reports *Difference*, *Count*, and *Sequence* yield the same instrumental value, which are higher than that of *Proportion*, and *Majority* has the lowest instrumental value. In addition to that, the ordinal ranking is identical between the two evaluation approaches.²¹

Hypothesis 5. *If the agent evaluates sample features according to their instrumental value, she will rank Difference/Count/Sequence as the most preferred features, Majority*

²¹Thus, given our theoretical benchmark, we use the terms informativeness (informative) and instrumental value (instrumentally valuable) interchangeably, which captures the level of uncertainty on the information accuracy.

as the least preferred features, and *Proportion* as somewhere in between.

Table 2.1: Two Measures of the Value of Reports

	No Report	Majority	Proportion	Difference/Count/Sequence
Instrumental Value $V(\gamma_R)$	\$0	\$1.35	\$1.46	\$1.52
Informativeness $I(\gamma_R)$	0	0.44	0.51	0.55

Note: The instrumental value of each report is the difference in the expected payoff between between each report and no report. The informativeness of each report is the reduction of the Shannon entropy compared to the no-report case.

Based on the discussion above, we can identify two categories of comparisons among reports. The first category focuses on reports that have maximum instrumental value and yield identical Bayesian posteriors, namely *Difference*, *Count*, and *Sequence*. Each of them aggregates the information of the drawn balls, S , in a lossless way. When facing any of them, a Bayesian agent uses the information on the difference in counts of different-colored balls to derive Bayesian posterior. *Count*, in addition to providing difference information, also conveys the sample size of S . *Sequence*, on top of counts, provides the information about the order in which the balls in S were drawn.²² However, neither sample size nor order is necessary for Bayesian inference. The theoretical benchmark suggests that, given the drawn balls S , the Bayesian posteriors should be identical across *Difference*, *Count*, and *Sequence*. Any deviation in performance or evaluation implies that the agent might use or perceive the usefulness of the non-instrumental feature(s) in a non-standard manner.

The second category focuses on reports that *differ* in their informativeness, with the three reports mentioned in the first category being more informative than *Proportion*, while *Proportion* is more informative than *Majority*. Less informative reports require

²²Given symmetric diagnostic rates, reports *Difference*, *Count*, or *Sequence* of the drawn balls S give the same Bayesian posterior. With asymmetric diagnostic rates, *Difference* is no longer a sufficient statistics of the drawn balls S but still has a larger instrumental value than those processed by *Proportion* and *Majority*. See more details in ?? about the predictions under asymmetric diagnostic rates.

agents to additionally take into account that the information can be either strong evidence or weak evidence. For example, when receiving “two orange balls and 1 green ball are drawn out of the selected box”, agents can learn this is a relatively weak evidence. On the contrary, consider the previous example of the report stating “67% of drawn balls are orange.” Agents need to take into account that three equally likely scenarios with different levels of information strength could give the same information: (1) a sample of two orange balls out of three draws, which would be relatively weak evidence; (2) a sample of six orange balls out of nine draws, which would be the evidence of intermediate strength; or (3) a sample of ten orange balls out of fifteen draws, which would be relatively strong evidence. The additional inference required by less informative reports might be cognitively demanding, which could result in larger deviation.²³ By comparing whether the performance ranking is in line with the ranking of instrumental value, we can test whether this additional inferential effort predicts how well subjects use the sample feature for belief updating.

Lastly, if agents are sophisticated about how well they will use the sample features to update beliefs, their preference would be consistent with performance.

Hypothesis 6. *If the agent is sophisticated about how she would use each sample feature for belief updating, her perceived usefulness would be consistent with how she actual uses sample features.*

²³Studies on uncertainty in signal interpretation find that individuals tend to be more conservative or insensitive to information change when they are uncertain whether the signal is strong or weak evidence (compound diagnostic rate) (Epstein, Halevy et al., 2019; Liang, 2021).

2.4 Results

We organize our main results as follows: Section 2.4.1 documents how subjects update their beliefs using the information provided by the five reports.²⁴ In Section 2.4.2, we compare the average willingness to pay to assess how subjects perceive the usefulness of each report. Section 2.4.3 explores the relationship between the actual use and perceived usefulness of the five reports.

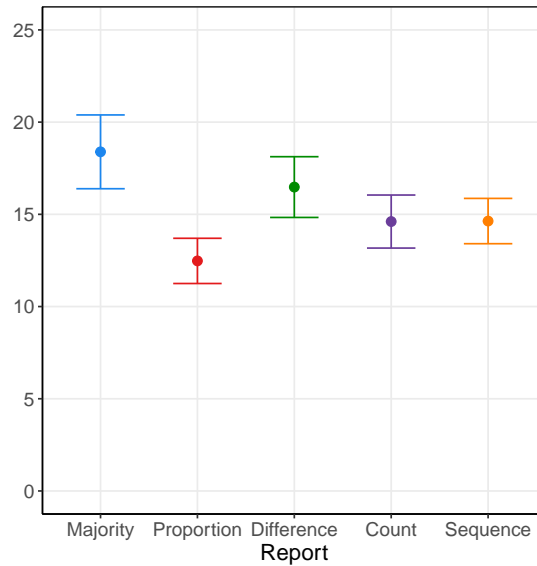
2.4.1 Performances across Sample Features

We apply two measures to assess the effectiveness of subjects in utilizing the information provided by each of the five reports when updating their beliefs.

First of all, we calculate the average absolute deviation from the Bayesian benchmark using subjects' stated beliefs, and compare them across the five reports. Figure 2.7 depicts the average absolute deviation for each report. We observe that subjects exhibit the least deviation under *Proportion*. It is significantly smaller than the deviations under *Count* and *Sequence* (t-test for each pairwise comparison, $p < 0.01$), even though the latter two are more informative than *Proportion*. The deviation under *Difference* is significantly larger than those observed in *Count* and *Sequence* (t-test for each pairwise comparison, $p < 0.01$), despite the three of them being equally informative. The largest deviation occurs under *Majority*, which are significantly larger than the deviations observed in the other reports (t-test for each pairwise comparison, $p < 0.01$). This finding provides evidence against the hypothesis that the extent to which subjects deviate from the Bayesian benchmark is identical across reports. Moreover, the observed difference in performance cannot be fully explained by variations in the informativeness of the five sample features.

²⁴We employ the terms “report” and “sample feature” interchangeably in this paper to refer to the same concept.

Figure 2.7: Average Deviation from Bayesian Benchmark by Report



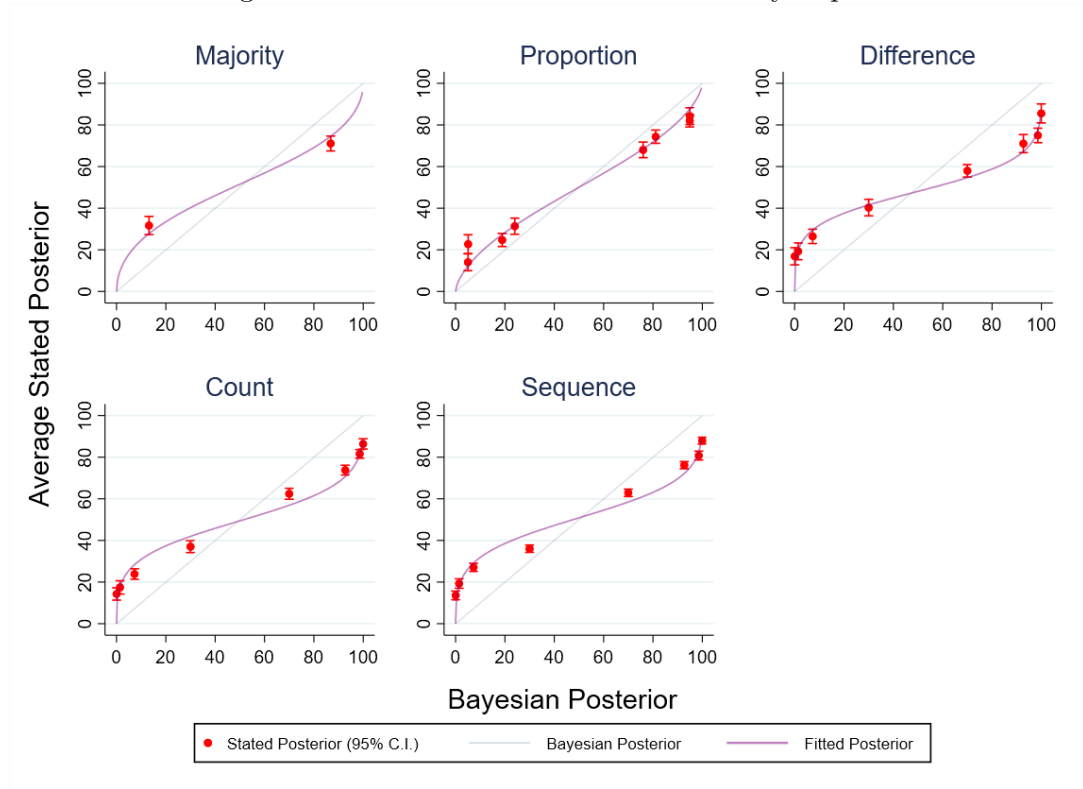
Note: The figure depicts the mean deviation of the subjects’ beliefs from the Bayesian posterior (in percentage term). For instance, if a subject assesses a belief of 80% against a scenario with the Bayesian posterior of 85%, the deviation is 5. 95% confidence intervals are included.

In addition, we use the Grether model as an alternative measure to assess performance. This model allows us to estimate the responsiveness of subjects to changes in the likelihood ratio based on the information presented in each of the five reports. Figure B.1 plots the average stated beliefs against the corresponding Bayesian posteriors for each report.²⁵ A Bayesian agent would consistently state their subjective beliefs as the Bayesian posteriors, resulting in a 45-degree line.

Remarkably, Figure B.1 demonstrates that the widely-established inverse S-shaped relationship between average stated beliefs and Bayesian posteriors, commonly observed in canonical “ball-and-box” belief updating tasks, is present across all five reports. The stated beliefs tend to be compressed closer to the 50:50 rather than aligning with the

²⁵The stated beliefs of 0% and 100% are excluded from Figure B.1 and Table 2.2 due to the logarithmic property used in the calculations. For the complete data, including these extreme beliefs, please refer to Appendix B.3, where we apply a linear approximation to accommodate the stated beliefs of 0% and 100%.

Figure 2.8: Underinference of Information by Report



Note: The stated posteriors are plotted against Bayesian posteriors and separated by reports. On each point, we plot the 95% confidence interval. The blue lines represent the 45-degree line, which denotes the Bayesian benchmark. The fitted posterior is derived from Equation (2.7) with the coefficients from Table 2.2. On the fitted lines, the stated beliefs of 0% and 100% are excluded due to the property of taking logarithm.

45-degree line. This suggests that under-inference, under-reaction to changes in the likelihood ratio, exists across all five reports. More importantly, the stated beliefs are closest to the 45-degree line under *Proportion*, indicating that subjects are the most responsive to changes in the likelihood ratio under *Proportion* compared to other sample features.

To formalize this, we estimate the coefficient of the reduced-form model proposed by Grether (1980), as shown in Equation (2.7), for each of the five sample features. Table 2.2 presents the estimated c for each sample feature. Firstly, our results replicate previous

findings where subjects receive *Count* or *Sequence* as signals. Specifically, in line with Benjamin (2019), we find that the estimated coefficients under *Count* and *Sequence* are 0.356 and 0.364, respectively.²⁶

Notably, as shown in Table 2.2, our study is the first to estimate c specifically for *Proportion* and *Difference*, and find them to be 0.679 and 0.311 separately. With a pooled analysis that combines all the observations and includes interaction terms for each sample feature, we find that the estimated c under *Proportion* is closer to 1 and significantly larger than any other sample feature (t-test for each pairwise comparison, $p < 0.01$). It indicates that subjects are more responsive to changes in the information conveyed by *Proportion* compared to the other features.²⁷ The estimated c for *Difference* is significantly smaller than that for *Count* and *Sequence* separately (t-test for each pairwise comparison, $p < 0.01$). This implies that subjects are less sensitive to changes in the likelihood ratio when using *Difference*, despite it being equally informative as *Count* and *Sequence*.²⁸ ²⁹ By measuring subjects' responsiveness to changes in the likelihood ratio, we observe similar patterns as with the average absolute deviation from the Bayesian benchmark: subjects are not equally responsive to the information change across the five reports, and this

²⁶In his meta-analysis, Benjamin uses the data from previous literature, where participants receive *Count* or *Sequence* as signals and elicit their beliefs to study belief updating. He finds that the estimated coefficient of c is 0.383 with a standard error of 0.028.

²⁷See Appendix B.3 for more details.

²⁸Due to the limited number of observations available for *Majority*, we are cautious in drawing conclusions about subjects' responsiveness to information changes under *Majority*. Each subject only receives one information under *Majority*, either indicating more orange or more green balls. Therefore, we acknowledge the need for further investigation and caution in interpreting the results regarding subjects' responsiveness to information changes under *Majority*.

²⁹One potential explanation for the subjects' improved performance under *Proportion* is that subjects may naively report the observed proportion information as their stated beliefs, resulting in a higher estimated c . To test this hypothesis, we categorize the stated beliefs into two groups: beliefs within a 5% range of the sample proportion and beliefs outside of this range. We find that 67% of the stated beliefs fall *outside* of the 5% range of the sample proportion. Moreover, when we plot the stated beliefs against the corresponding Bayesian posteriors, separating them by the two groups, the stated beliefs *outside* of the 5% range of the sample proportion are closer to the Bayesian benchmark than to the 50:50. This suggests that the improved performance under *Proportion* is not solely driven by a naive reporting of the observed proportion information. Please see Appendix B.4 for more details.

variation does not respond to increasing the informativeness of the five reports.

Table 2.2: Effect of Information Strength on Under-inference by Report

	(1)	(2)	(3)	(4)	(5)	(6)
	Majority	Proportion	Difference	Count	Sequence	All
$\ln \left(\frac{p(O S_{\gamma_R})}{p(G S_{\gamma_R})} \right)$	0.535*** (0.0565)	0.679*** (0.0325)	0.311*** (0.0183)	0.356*** (0.0169)	0.364*** (0.0158)	0.367*** (0.0159)
Constant	0.0967 (0.194)	-0.228* (0.124)	-0.174* (0.104)	-0.186** (0.0841)	-0.0339 (0.0707)	-0.0934 (0.0706)
Observations	97	390	387	856	1475	3205

Note: We calculate the ratio of stated posteriors and then take the natural log to form the explained variable, $\ln \left(\frac{\pi(O|S_{\gamma_R})}{\pi(G|S_{\gamma_R})} \right)$. For the explanatory variable, we calculate the ratio of Bayesian posteriors and then take the natural log, $\ln \left(\frac{p(O|S_{\gamma_R})}{p(G|S_{\gamma_R})} \right)$. The observations with $\pi(G|S_{\gamma_R}) = 0$ or 1 are dropped. Columns (1) - (5) represent the regression estimations under each of the five reports, respectively. Column (6) indicates the regression results with all the data pooled together. Two categorical variables, gender and grades, are added as controls in all the regressions. Standard errors are clustered at the subject level and presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We summarize these results as follows:

Result 6. *Subjects' belief updating is the closest to Bayesian benchmark when using Proportion, despite Proportion being less informative compared to Difference, Count, and Sequence. Moreover, among the sample features that are equally informative (Difference, Count, and Sequence), subjects' belief updating is closer to the Bayesian benchmark under Count and Sequence than under Difference.*

Our findings suggest that, different from the Bayesian benchmark, subjects do not exhibit equal proficiency in utilizing the various sample features in the realized signals for belief updating. Additionally, the performances do not respond to the informativeness of sample features in two ways: (1) subjects' performances do not monotonically improve with the informativeness of the features provided in the reports: they are better at using *Proportion* compared to other features with higher instrumental value, namely

Count/Sequence/Difference. (2) some non-instrumental feature helps: when comparing *Difference*, *Count*, and *Sequence*, subjects are better at using *Count* and *Sequence*, even though these additional features do not add more instrumental value for Bayesian inference compared to *Difference*. Our results also shed light on the “Strength-Weigh bias” by suggesting that subjects exhibit better performance in utilizing the “Strength” (sample proportion) when used independently, rather than when combined with “Weigh” (sample size).

Furthermore, the varying difficulties associated with retrieving proportion information from the received reports could explain the differences in belief updating performances across the reports. On the one hand, when facing *Count* and *Sequence*, subjects may need to conduct additional mental calculation to extract the proportion information. This computational burden could tax subjects’ belief updating behaviors, resulting in a compression towards 50:50 and reduced sensitivity to changes in the likelihood ratio. On the other hand, retrieving the proportion information under *Majority* and *Difference* requires additional inference about all possible proportions that could yield the same information. This additional step of inference may result in less effective utilization of the information when updating beliefs.

Last but not least, the observation that the deviations under *Count* and *Sequence* are smaller compared to those under *Difference* and *Majority* suggests that the complexity associated with making inferences may be greater than that of performing calculations. However, it is important to note that these arguments assume that subjects perceive *Proportion* as the most useful feature for belief updating and would like to extract it from received reports. We provide further support for this assumption in the next section.

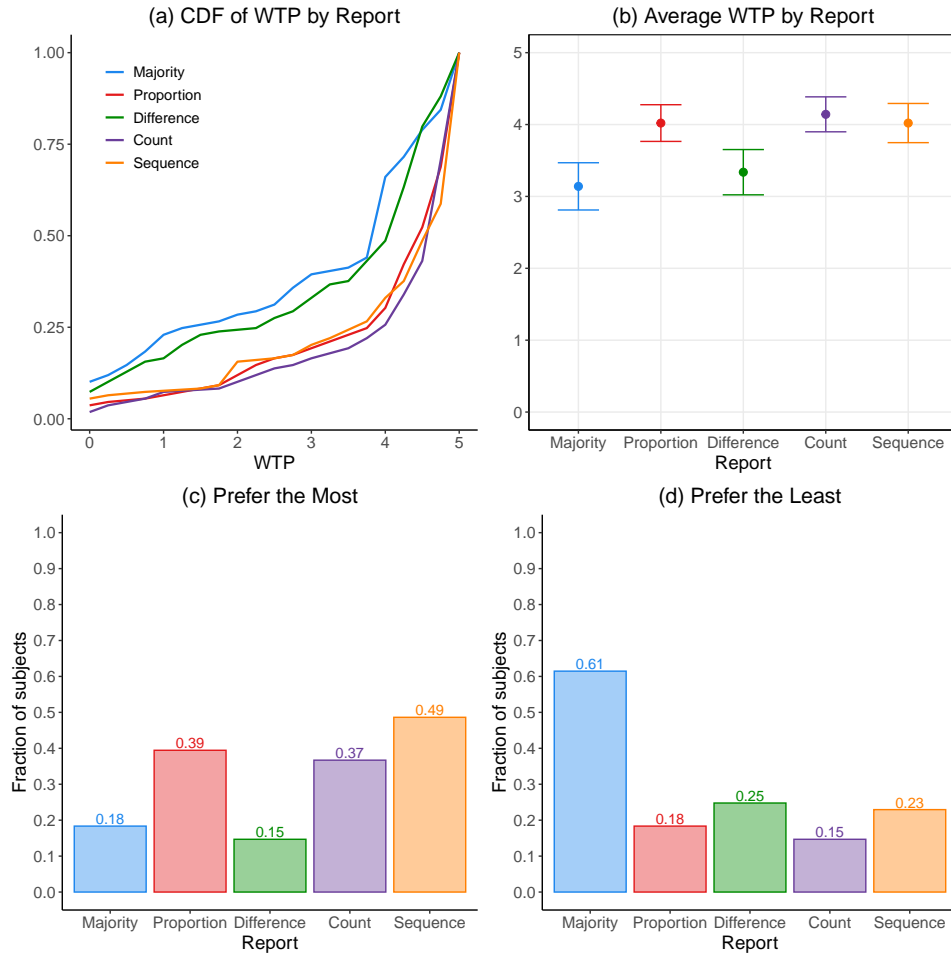
2.4.2 Preferences across Sample Features

In this section, we explore subjects' perceived usefulness of the five reports by assessing their elicited willingness to pay (WTP), and compare it with theoretical predictions of instrumental value.

Panel (a) of Figure 2.9 depicts the distribution of WTP for each report. The distributions of *Proportion*, *Count* or *Sequence* first order stochastically dominate those of *Majority* or *Difference*. Panel (b) of Figure 2.9 shows the average WTP for each report. There is no significant difference in the average WTP among *Proportion*, *Count* and *Sequence*. The average willingness to pay for *Difference* is also lower than that for *Proportion* by \$0.68. Between *Proportion* and *Difference*, approximately 65% of subjects express a preference for the former over the latter. Our results indicate that subjects prefer *Proportion*, *Count* and *Sequence* the most, while preferring *Majority* the least, and *Difference* is somewhere in between.³⁰

³⁰Pairwise Wilcoxon rank test on ranking with multiple testing correction (Benjamini-Hochberg adjustment) suggests that the gap of WTP between *Proportion/Count/Sequence* and *Difference* is significant at 99% confidence level, and the difference between *Difference* and *Majority* is significant at 90% of the confidence level.

Figure 2.9: Preference over Reports



Note: Panel (a) plots the cumulative density function of the reported willingness-to-pay, which is separated by reports. Panel (b) plots the average willingness-to-pay of each report. In Panel (b), 95% confidence intervals are included. Panel (c) plots the fraction of subjects who rank the report as the most preferred and separated by reports, and Panel (d) plots the fraction of subjects who rank the report as the least preferred (tied results are included).

To test the extent to which the gap in WTP is driven by different monetary scales subjects use for evaluation, Panels (c) and (d) of Figure 2.9 plot the fraction of subjects who consider each report as the most preferred and the least preferred, respectively. We find that the gap observed in average WTP is not solely due to different scales that subjects use to rank reports. The ordinal ranking demonstrates a consistent pattern: the majority of subjects rank *Proportion*, *Count*, and *Sequence* as the most preferred reports,

while ranking *Majority* and *Difference* as the least preferred reports.

In addition, there exists notable heterogeneity in the perceived usefulness of reports containing information about sample proportion, namely *Proportion*, *Count* and *Sequence*. Some subjects prioritize receiving the sample proportion only, while others recognize the value of incorporating additional features. Among the subjects, 39% rank *Proportion* as the most preferred report, while 37% and 49% rank *Count* and *Sequence* as the most preferred, respectively. Subjects who rank *Proportion* highest are willing to pay an average of \$1.17 more to avoid receiving additional features beyond sample proportion. On the other hand, those who rank *Count* or *Sequence* as the most preferred report appreciate the values of the extra features alongside sample proportion, as indicated by their willingness to pay an average of \$0.66 more to receive *Count* or *Sequence*, compared to *Proportion*.

In sum,

Result 7. *The preference for sample features deviates from instrumental value in two ways:*

1. *On average, subjects consider Proportion, Count and Sequence as equally useful, despite the features in the latter two being more informative than Proportion;*
2. *Subjects, on average, value Count and Sequence more than Difference, even though all three are equally informative for Bayesian inference.*

Our findings suggest that subjects' perceived usefulness of sample features does not align with their instrumental value. On average, the subjects have a strong preference for reports that contain the feature of sample proportion compared to those that do not. However, they fail to fully recognize the usefulness of other features such as sample difference and sample size, even though incorporating the latter two with *Proportion* makes the information more useful for Bayesian inference.

These findings suggest that subjects, on average, have a stronger preference for sample features that contain *Proportion* compared to those that do not. Features that contain *Proportion* (*Count* and *Sequence*), require subjects to conduct some calculations to get the proportion information. Features that do not contain *Proportion* (*Difference* and *Majority*), require additional inference about all the potential sample proportions that could lead to the same difference or majority information. It is noteworthy that there are differences in the degree of aversions towards these two types of additional efforts. Subjects demonstrate a stronger aversion (higher *WTP*) to avoid the need to make additional inferences compared to the need to perform additional calculations. This suggests that subjects may perceive the former as more difficult than the latter.

The observed heterogeneity in the perceived usefulness of reports containing the sample proportion indicates a potential variation in the relationship between individual preferences and performances. Some subjects exhibit a “Strength-Weight preference” by preferring *Proportion* the most, while others prioritize reports that have sample size along with *Proportion*. These findings suggest that there might be some heterogeneity in the association between preferences and performances, which we will discuss in detail in the next section.

2.4.3 Association between Preferences and Performances

In this section, we aim to examine the association between subjects’ perceived usefulness and their actual use of the five reports. We investigate whether subjects who underestimate the usefulness of certain features also tend to use them suboptimally. By analyzing this association, we can gain valuable insights into the nature of deviations from the Bayesian benchmark, distinguishing between intentional deviation and inatten-

tive heuristics.

On the one hand, if subjects' preferences align with their performance, it would suggest that subjects have a sophisticated understanding of the usefulness of each report for belief updating. Consequently, the observed non-standard belief updating would likely be an intentional deviation from the Bayesian approach. On the other hand, if subjects' preferences are inconsistent with their performance, it would indicate that subjects fail to accurately predict their performance. Other behavioral traits might affect how they value information as well. In such cases, the non-standard belief updating is more likely to be a result of inattentive heuristics.

To achieve this goal, we measure each subject's performance across the five reports by calculating their average absolute deviation for each report. We use each subject's WTP values for the five reports to measure subject-level preference, and use the ten pairwise comparisons to calibrate the complete relationship among the five reports. We employ regression estimation to formalize the relation between preference and performance. The dependent variable is the difference in the average absolute deviation between Report X and Report Y , for each pair of reports. We construct a categorical variable to capture the relative comparison between WTP_X and WTP_Y , which serves as the explanatory variable. We also use the indicator variable on whether the average absolute deviation under Report X is smaller than that under Report Y as an alternative dependent variable. It helps determine whether subjects are more likely to perform better (indicated by a smaller deviation) under one report compared to the other.

Table 2.3 demonstrates the main regression results. Compared to the case of indifference ($WTP_X = WTP_Y$), subjects deviate 3.28 less under the more-preferred report than under the less-preferred one. Going by one category of the pairwise comparison outcomes between X and Y (e.g., from indifference to preferring X over Y) is associated with an increase of 66% ($e^{0.508} - 1 \approx 0.66$), in the likelihood of deviating less in X compared to

Y . Both results are statistically significant at the 95% confidence level. These findings indicate that, on average, subjects are consistent between preferences and performances: they perform better under the report they prefer.

Table 2.3: Association between Preference and Performance

	(1)	(2)
	$AD_X - AD_Y$	$\mathbb{1}\{\text{Perform Better in X}\}$
$WTP_X > WTP_Y$	-3.279 ** (1.506)	0.508 ** (0.221)
$WTP_Y > WTP_X$	-0.496 (1.467)	-0.225 (0.218)
(Intercept)	2.348 * (1.405)	-0.145 (0.182)
N	1090	1090
(Pseudo) R^2	0.024	0.034

Note: In Column (1), the dependent variable is the difference in the average absolute deviation from Bayesian posterior between Reports X and Y in a given pair. We construct a categorical variable that takes the value of 1, 0, or -1 if $WTP_X > WTP_Y$, $WTP_X = WTP_Y$, or $WTP_X < WTP_Y$, respectively, to be the independent variable. In Column (2), we use Logit model and the indicator variable on whether the average absolute deviation under Report X is smaller than Report Y as an alternative dependent variable to capture whether subjects are more likely to perform better (smaller deviation) under one report versus the other. Standard errors are clustered at the subject level and presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We use the ordinal rankings of both preference and performance to explore the heterogeneity of the preference-performance relationship. To be more specific, we rank the five reports based on the number of sample features they contain or the level of informativeness of those features. According to this criterion, the ranking of reports is as follows: 1st *Sequence*, 2nd *Count*, 3rd *Difference*, 4th *Proportion*, and 5th *Majority*.

For each pair of reports, we refer to the one ranked lower on this list as Report X and the alternative as Report Y . Within each pair of reports, we define a preference-performance relation as “Perform better under Preferred” if a subject exhibits a smaller average absolute deviation (AAD) and states a larger WTP for one report compared to the alternative in that pair. In addition, we consider a preference-performance relation

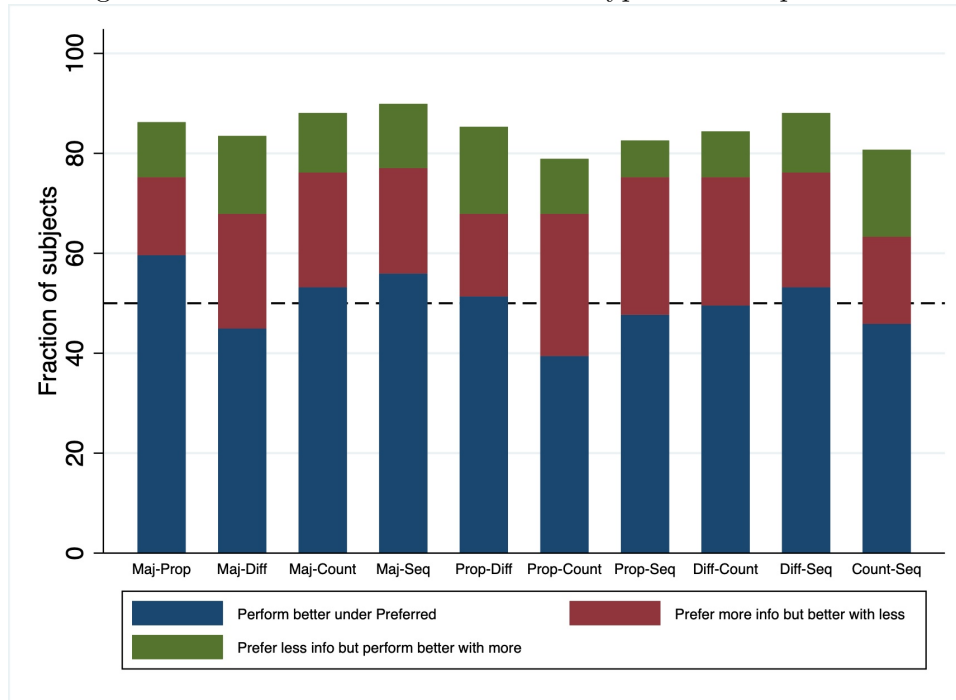
type as “Prefer more but better with less” if a subject has a smaller AAD but states a lower WTP on Report Y compared to Report X . On the other hand, a preference-performance relation is categorized as “Prefer less but better with more” if a subject has a larger AAD but states a larger WTP on Report Y compared to Report X . Table 2.4 illustrates the definition of association types.

Table 2.4: Definition of Association Type

Association Type	Reports X and Y : AAD and WTP
Perform Better under Preferred	$AAD_{Y(X)} < AAD_{X(Y)}$ and $WTP_{Y(X)} > WTP_{X(Y)}$
Prefer Less but Better with More	$AAD_Y < AAD_X$ and $WTP_Y < WTP_X$
Prefer More but Better with Less	$AAD_Y > AAD_X$ and $WTP_Y > WTP_X$

Note: The ten pairwise comparisons calibrate the association between preference and performance across the five reports. For each pair of reports, denoted as Report X and Report Y , we refer to Report X as the one with fewer features (regardless of informativeness), or less informative features, while Report Y is the one with more features (regardless of informativeness), or more informative features. The notations, $Y(X)$ and $X(Y)$, mean that the same relationship holds when replacing all the Y with X and all the X with Y .

Figure 2.10: Distribution of Association Types in 10 Report Pairs



Note: The fractions of subjects classified by the preference-performance types are plotted against the ten pairs of reports. In each pair, $X - Y$, the former name in short is Report X , the latter name in short is Report Y . X either contains fewer features (regardless of informativeness) or less informative sample features than Y . In each pair of reports, a preference-performance association is defined as “Perform Better under Preferred” if a subject has a smaller average absolute deviation (AAD) and states a larger WTP on one report than the alternative. A preference-performance association type is defined as “Prefer More but Better with Less” if a subject has a smaller AAD but states a lower WTP on Report Y than Report X . A preference-performance relation is defined as “Prefer Less but Better with More” if a subject has a larger AAD but states a larger WTP on Report Y than Report X . The dashed line represents 50% of subjects as the reference. Tied results are excluded.

Figure 2.10 demonstrates the distribution of the three types across the ten pairs. Consistent with our aggregate results, the majority of subjects fall into the “Perform better under Preferred” type, representing approximately 50% of subjects in each pair. This indicates that a significant portion of subjects demonstrate consistency between their preferences and performances.

Furthermore, there is notable heterogeneity among the inconsistent types, where subjects’ preference and performance do not align. The second largest type is the “Prefer

More but Better with Less” type, which comprises, on average, 22% of subjects. These individuals express a preference for the report with more or more informative features but actually perform better under the one with fewer or less informative features. Additionally, 13% of the subjects belong to the ”Prefer Less but Better with More” type, indicating that they prefer the report with fewer or less informative features but achieve better performance under the one with more or more informative features. This diversity in the inconsistent types highlights the complex interplay between preferences and performances among the subjects.

To summarize,

Result 8. *On average, subjects are self-consistent between their preferences and performances, performing better under the sample feature they prefer. However, there is also non-negligible heterogeneity in the inconsistent association of ordinal rankings between preference and performance:*

- *Substantial inconsistencies are observed among subjects;*
- *The most prominent type of inconsistency is “Prefer More but Better with Less”: subjects prefer the report that contains more features or more informative ones but actually perform better under the report that contains only the necessary features for their belief updating.*

Our results suggest that the non-standard use of sample features in belief updating is more likely to be intentional deviations rather than inattentive heuristics. In other words, subjects underestimate the usefulness of certain sample features, and fail to make optimal use of them when updating beliefs, despite these features being instrumentally more valuable for Bayesian inference compared to other features. For instance, our results shed light on biases such as the “Strength-Weight bias” or the “Sample-Size

neglect” documented in previous literature (Kahneman and Tversky, 1972). These biases involve the evaluation of the sample proportion (referred to as “Strength”) and sample size (referred to as “Weight”), respectively. Our findings indicate that these biases are primarily associated with subjects’ over-valuing the importance of “Strength”, while under-appreciating the importance of “Weight” when it comes to belief updating.

The majority type among subjects whose preferences are inconsistent with their performances is “Prefer More but Better with Less.” This finding suggests that these subjects might fail to consider the cost-benefit trade-offs associated with processing additional sample features that do not matter for their belief updating. Despite the fact that the theoretically defined informativeness increases with more information, these subjects tend to prioritize quantity over relevance. In doing so, they may fail to recognize that the additional information does not necessarily improve the accuracy of their belief updating. Furthermore, this preference for more or more informative features may come at a cost. The additional effort or cognitive resources required to process these features can pose a challenge, potentially hindering subjects from making optimal use of the available information when updating their beliefs.

There is also a non-negligible fraction of subjects who demonstrate a lesser sophistication in understanding how additional sample features can aid in belief updating. This suggests that these individuals may not fully recognize the value of incorporating supplementary information for accurate belief revision. Furthermore, it is worth considering that non-standard preferences for information, such as a preference for simplicity, could potentially influence how they evaluate the usefulness of information.

2.5 Conclusion

In this paper, we use a controlled laboratory experiment to study how individuals use and perceive the usefulness of different statistical characteristics of realized signals, namely sample features when updating beliefs. In terms of performance, a Bayesian agent would be equally good at processing each sample feature, as they use the Bayes' rule to do so. However, what we find is that subjects are not equally good at processing each sample feature. First of all, subjects under-use the information contained in each of the five sample features, while the magnitudes differ across sample features. We find that subjects are better at using *Proportion* than the other features: subjects' stated posteriors are closest to the Bayesian benchmark under *Proportion*, even though it is less informative compared to *Difference*, *Count*, and *Sequence*. Subjects deviate the most from the Bayesian benchmark under the least informative sample feature – *Majority*. These results provide direct evidence of “Strength-Weight Bias” – better at using sample proportion but worse at using sample size for belief updating.

In terms of preference, subjects' perceived usefulness of sample features also deviates from what instrumental value/informativeness would predict. Subjects value *Proportion* as equally useful as *Count* and *Sequence*, even though the latter two have higher instrumental value or more informative than the former one. Overall, subjects prefer the features that can back out *Proportion* with some computational operations over those that require additional inference or contingent reasoning on all the possible *Proportion* information. These results suggest that subjects have “Strength-Weight Preference” of the information – prefer using sample proportion rather than using sample size for belief updating.

Combining preference and performance, we show that, on average, subjects make better use of the sample features they prefer, while there exists notable heterogeneity in

the inconsistency between preference and performance. This indicates that the biased use of sample features in belief updating is more likely to be an intentional deviation rather than inattentive heuristics. Overall, our results indicate that the suboptimal use of some informative sample features can account for a substantial amount of deviation from Bayesian benchmark in belief updating, which is positively correlated with how individuals perceive the usefulness of different sample features.

Our results open interesting questions for further research. One natural next step is to explore the generality of our current finding with other information on sample features. In our experiment, under less informative sample features, we deliberately choose the information that maps to different information under more informative ones. This allows us to see how the instrumental value of information would interact with the way subjects use the information in each sample feature. Thus, the information provided by different sample features maps to different Bayesian benchmarks, even if they have some sample features in common. For example, the Bayesian posterior under Report *Proportion* saying that 80% of balls are green is different from those under the Report *Count* either saying four green balls and one orange ball, 12 green balls and three orange balls, or 20 green balls and five orange balls, separately. As there is no one-to-one mapping between the information of sample features with different instrumental values, our results could be driven by the difference in Bayesian posteriors rather than the difference in subjective belief updating processes. Thus, it would be interesting for future work to explore to what extent our findings are due to the different updating behaviors.

Secondly, it would be interesting to directly ask whether subjects process the information to get certain sample features and use those to make inferences when receiving certain information and what behavioral traits drive their valuation of sample features. A contemporary paper by Bordalo et al. (2023) demonstrates that the similarity between information and hypothesis is one of the reasons behind this. However, as shown in

Appendix B.4, our finding is not purely driven by reporting whatever they received. It would be a fruitful direction for future research.

Chapter 3

Dynamic Binary Method

Joint work with Xin Jiang

3.1 Introduction

Information on individual beliefs is central for researchers to better understand economic behavior (Manski, 2004). Without data on what people think and expect, it is challenging to differentiate between alternative choice models, understand the boundaries of rationality, or examine new equilibrium concepts. However, eliciting individual beliefs poses its own set of challenges. Existing methods primarily rely on individuals selecting a number from 0% to 100% to represent their probabilistic beliefs (Charness, Gneezy and Rasocho, 2021), which introduces numerous issues. For instance, individuals may possess imprecise rather than precise probabilistic beliefs about a particular event (Giustinelli, Manski and Molinari, 2022). They might have a general notion but struggle to provide the best point estimate. When asked to state a point belief, cognitive difficulties may arise, leading to conservative responses and systematic deviations from truthful reporting (Charness, Gneezy and Rasocho, 2021).

In this paper, we introduce a new elicitation method called the *Dynamic Binary Method* (DBM). Unlike *Classical Methods* (CM), which directly ask respondents to select a number from 0 to 100 as their probabilistic beliefs, and use proper scoring rules such as the Binary Scoring Rule (BSR) to incentivize truthful reporting, DBM differs in how beliefs are stated and whether they must be a single value. Inspired by the bisection process – the iterated partition of a choice set into two equally large subsets, with perceptions elicited through a series of binary choices (Baillon, 2008) – DBM allows respondents to exit at any step and state interval beliefs if they prefer. This method is designed to achieve two primary goals: (1) alleviate the challenge of forming a precise point estimate of beliefs or perceptions, and (2) quantify the self-perceived precision of those beliefs.

To elaborate further, starting with the full belief space, in each step s , DBM divides the belief space $[I_l^s, I_u^s]$ into two equally sized intervals: $[I_l^s, \frac{I_l^s + I_u^s}{2}]$ and $(\frac{I_l^s + I_u^s}{2}, I_u^s]$, where I_l^s and I_u^s denote the lower and upper bounds of the presented interval, respectively. The decision maker (DM) must then select either $[I_l^s, \frac{I_l^s + I_u^s}{2}]$ or $(\frac{I_l^s + I_u^s}{2}, I_u^s]$, or they can opt to exit with the interval $[I_l^s, I_u^s]$. If the DM chooses to exit, the computer randomly selects a number a_R from the stated belief range $a = [a_l, a_u]$, following a uniform distribution. The selected number a_R is then applied in a proper scoring rule, such as the BSR, to determine the DM's payoff.

For an expected utility maximizer, choosing the mean of their true belief, no matter whether their true belief is precise or distributed, is optimal in both DBM and CM. However, an expected utility maximizer who does not perfectly foresee the optimal choice but instead considers randomization over $[I_l^s, \frac{I_l^s + I_u^s}{2}]$, $(\frac{I_l^s + I_u^s}{2}, I_u^s]$, or $[I_l^s, I_u^s]$, may opt to exit early with an interval whose midpoint equals the mean of their true belief. Thus, the decision to exit early indicates whether the DM is myopic or not. DBM also facilitates relative judgment by asking which range is more likely, thereby sidestepping the challenge

of finding the best point estimate. If it is main driving force behind biases in perception, for example, the compressed relationship between respondents' probabilistic estimates and "true" probabilities, it would have the potential to mitigate the difficulty of forming precise point estimates.

To assess the empirical validity of DBM, we conduct both within-subject and between-subject experiments using a diverse range of perception tasks from previous literature. Specifically, for the between-subject design, we utilize four task categories from controlled laboratory experiments: simple prior tasks (Danz, Vesterlund and Wilson, 2022), compound prior tasks (Liang, 2022), belief updating tasks (Danz, Vesterlund and Wilson, 2022), and estimation tasks (Dewan and Neligh, 2020; Falk and Zimmermann, 2018) with artificial settings such as balls and urns, counting peas in a bowl, or dots in a graph. Additionally, we incorporate four task categories from field or lab-in-the-field experiments: perception on economic or financial variables (Enke and Graeber, 2023), the labor market (Wiswall and Zafar, 2015*a*), and education (Wiswall and Zafar, 2015*b*), all of which have real-life settings.

To address the challenge of not knowing participants' true beliefs, we carefully design the questions to ensure that each task has an objective truth. Furthermore, we intentionally select parameters for each question to ensure that the objective truths span the entire belief space, including centered, extreme, and intermediate values. For the within-subject design, we allow each participant to complete a set of perception tasks using both DBM and CM in a randomly determined order. This approach aims to assess the extent to which the elicited beliefs in tasks using DBM can predict stated point beliefs in tasks using CM at the subject level.

First of all, we find that DBM does not perform significantly differently from CM at the aggregate level, regardless of whether the perception questions use artificial/laboratory settings or real-life settings. This finding is robust across different measures, including

the absolute deviation of the midpoint of elicited beliefs from the objective truth or the expected absolute deviation of elicited beliefs from the objective truth. This suggests that the challenge of forming a point estimate of beliefs/perceptions may not be the primary driver of biased perception elicited using CM. But DBM outperforms CM when the task has extreme values as the objective truth. This implies that some perception biases, such as central tendency, could result from the narrowed consideration set that respondents use to choose beliefs or perceptions from.

Furthermore, we find, from both between-subject and within-subject perspectives, that the length of stated beliefs in tasks using DBM is negatively correlated with their accuracy and can predict how well respondents perform in tasks using CM at the aggregate level: the longer the interval, the less accurate the stated belief in DBM and the less accurate the stated belief in CM. Moreover, within-subject results highlight participants' sophistication regarding the precision of their beliefs/perceptions: participants who stated point beliefs in DBM in more tasks demonstrate less deviation from the objective truth in their stated beliefs in CM. This pattern is particularly significant among participants who completed tasks with DBM first and subsequently used CM.¹

Note that this relationship is not strictly monotonic: stated beliefs reaching the point are not the most accurate and do not predict the most accurate beliefs stated in CM. Participants who always choose until reaching the point in all tasks using DBM are not the most accurate in tasks using CM. Moreover, our findings reject the hypothesis that participants have precise beliefs/perceptions but do not bother to choose until reaching the points for reasons such as complexity. If this were the case, we would expect no correlation between the length of their stated beliefs in DBM and the absolute deviation of their stated point beliefs from the objective truth in CM. This finding suggests that par-

¹We interpret this difference as a fatigue effect as in our Experiment 1, subjects are underpaid given the time they took to finish the experiment and the standard payment suggested by Prolific.

ticipants possess some level of awareness regarding how accurate their beliefs/perceptions would be when using DBM.

Lastly, we compare three methods of using the stated beliefs elicited with DBM to predict point beliefs elicited with CM. We find that predictions using a weighted average between subjective truth (the midpoint of stated beliefs in DBM) and the cognitive default (e.g., midpoint of the slider bar), with the relative weight on the default determined by the length of stated beliefs in DBM, are closest to the average stated beliefs in CM. This approach outperforms both using the midpoint of stated beliefs in DBM alone and using objective truth instead of subjective truth in the weighted average method. Our findings underscore the significance of incorporating the precision of stated beliefs and perceived truth to enhance predictions of economic behavior.

Relations to the existing literature. This paper makes several contributions to the existing literature. Firstly, our study aligns closely with previous research on perception/evaluation imprecision and the notion of cognitive uncertainty introduced by Enke and Graeber (2023). Most studies in this domain focus on capturing preference incompleteness, cognitive noise, or cognitive uncertainty using non-incentivized techniques. For instance, Enke and Graeber (2023) measure “cognitive uncertainty” by having participants first choose from a slider bar to state their beliefs/perceptions and then report a probabilistic value indicating the extent to which they are “certain” about their previous choice is the best on a second screen without incentivizing truth-telling. Similar technique is used in Giustinelli, Manski and Molinari (2022); Nielsen and Rigotti (2023) for the identification of belief imprecision by asking participants to report probability intervals after the question using a precise percent-chance format, with the question about belief range being unincentivized. Recently, Agranov and Ortoleva (2020) proposes an incentivized method to measure the extent to which people choose to randomize between

two risky options, focusing on eliciting the ranges of preference for randomization in the domain of choice under risk.

Our study contributes to the literature by proposing a new incentivized method for eliciting participants' imprecise beliefs in the domains of perception and inference.

Secondly, our study is situated within the growing empirical literature on preferences from randomization. Existing studies have documented randomization in various contexts, including objective lotteries (Agranov and Ortoleva, 2017; Dwenger, Kübler and Weizsäcker, 2018; Feldman and Rehbeck, 2022), ambiguity preferences (Cettolin and Riedl, 2019), time preferences (Agranov and Ortoleva, 2017), social preferences (Agranov and Ortoleva, 2017; Miao and Zhong, 2018), and even choices involving dominated options (Agranov, Healy and Nielsen, 2023; Rubinstein, 2002). The survey paper by Agranov and Ortoleva (2022) demonstrates high rates of preferences for randomization across these domains and shows their persistence even after explicit training.

Similar to these studies, we capture the prevalence of randomization using incentivized measures. Moreover, we extend this line of inquiry into the domain of belief formation and inference and document the prevalence of randomization over beliefs, thereby complementing existing literature in this area.

The rest of this paper is organized as follows. Section 3.2 delves into the theoretical benchmark of DBM and CM with BSR. Section 3.3 outlines the experimental design. Section 3.4 presents the results, and Section 3.5 concludes.

3.2 Theoretical Framework

Consider a decision maker (DM) with a probabilistic belief over a verifiable binary outcome $s \in \{A, B\}$, assuming they possess a true belief $p = Pr\{s = A\}$. Binarized scoring rule (BSR) uses two monetary prizes M_h and M_l for payment (where $M_h > M_l \geq$

0), and two i.i.d. draws $X_1, X_2 \sim U[0, 1]$ to determine the outcome (Hossain and Okui, 2013; Wilson and Vespa, 2018). Specifically, if $s = A$ is true, the DM gets the prize M_h so long as their stated belief a is greater than at least one of the two uniform draws X_1 and X_2 . If $s = B$ is false, the DM gets the prize M_h so long as their stated belief a is less than at least one of the two uniform draws X_1 and X_2 . Otherwise, the DM gets the prize M_l . Given the true belief p , the probability of winning the better prize M_h is

$$\pi(p, a) = p * (1 - (1 - a)^2) + (1 - p) * (1 - a^2) \quad (3.1)$$

Thus, BSR generates a reduced lottery:

$$\mathcal{L}(a|p) = \pi(p, a) \circ M_h \oplus (1 - \pi(p, a)) \circ M_l \quad (3.2)$$

Without loss of generality, assume $M_l = 0$. Given the true belief p , finding the optimal stated belief $a \in [0, 1]$ that maximizes the expected utility in the BSR is equivalent to maximizing the likelihood of receiving the prize M_h .

Classical Methods (CM) refer to implementation methods that elicit the DM's stated point belief a by directly asking the DM to report any value within the full choice space, such as any real number between 0 and 1. As this approach is widely used in existing literature, we refer to them as Classical Methods (CM). The stated belief a is then used in Equation (3.1) to determine the lottery for their rewards, i.e., Equation (3.2), and the outcome is realized accordingly.

Dynamic Binary Method (DBM) is based on the bisection method. The choice interval is repeatedly partitioned into two equally lengthy sub-intervals, for which the

DM's beliefs are elicited through a series of binary choices. Starting with the full choice space, for example, $[0, 1]$, at each step, the method divides the choice interval $[I_l, I_u]$, where I_l and I_u denote the lower and upper bounds separately, in two halves at the midpoint $\frac{I_l+I_u}{2}$: $[I_l, \frac{I_l+I_u}{2}]$ and $(\frac{I_l+I_u}{2}, I_u]$. Unlike the standard bisection method, which requires the DM to continue until a specific point is reached, DBM allows the DM to exit at any step and choose the current interval $[I_l, I_u]$ as their belief. Upon exiting the process, a computer a_R is randomly selected within the last range $[I_l, I_u]$, following a uniform distribution. This a_R is then used as the stated belief in the BSR to determine the lottery for their rewards, i.e., Equation (3.2), and the outcome is realized accordingly. Therefore, the DM can either choose until the point where $a = I_l = I_u$, or select an interval $[I_l, I_u]$ as their stated belief, with a_R uniformly distributed within this interval.

3.2.1 Incentive Compatibility with CM for EU Maximizer

When the true belief p is precise, i.e., a singleton, and the CM is employed to elicit the stated belief, with BSR, the best response is to choose the point where $a^*(p) = p$ because $\mathcal{L}(a^*(p)|p)$ stochastically dominates any other available lottery $\mathcal{L}(a|p)$. Conversely, when the true belief p follows a non-degenerate distribution $f(p)$ with $\mu_p = E(p)$ and $\sigma_p^2 = Var(p) > 0$, and the CM is used to elicit the stated belief, the objective becomes maximizing the expected likelihood of receiving the prize M_h :

$$\max_a E_p[p * (1 - (1 - a)^2) + (1 - p) * (1 - a^2)] \quad (3.3)$$

The distribution over p reflects the idea that the perception of $Pr\{s = A\}$ can be noisy, uncertain, or imprecise (Enke and Graeber, 2023; Frydman and Jin, 2022; Giustinelli, Manski and Molinari, 2022). The best response in this situation is to select the point $a^*(p)$ where $a^*(p) = E(p) = \mu_p$.

Proposition 1. *Given the true belief p , regardless of whether the true belief is a singleton or a distribution, when the CM is used to elicit belief as a singleton and BSR is used to determine payoff, an expected utility maximizer will choose the point $a^*(p)$ where $a^*(p) = E(p) = \mu_p$.*

3.2.2 Incentive Compatibility with DBM for EU Maximizer

Since the DBM allows the DM to either continue until reaching a single point or exit early with a random variable uniformly distributed over the last range they chose, i.e., $a \sim Uniform[a_l, a_u]$, the optimization problem becomes:

$$\max_a E_p\{p * E_a[(1 - (1 - a)^2)|p] + (1 - p) * E_a[(1 - a^2)|p]\} \quad (3.4)$$

which is equivalent to

$$\max_a \{-Var(a) - [E(a) - E(p)]^2 + E(1 - p) + [E(p)]^2\} \quad (3.5)$$

where $Var(a)$ and $E(a)$ denote the variance and the mean of stated belief a , respectively. To maximize the expected utility, it is optimal to choose until the point a^* where $a^* = E(p)$ and $Var(a) = 0$.

In sum, given the true belief p , to maximize expected utility, it is optimal to continue until reaching the point $a^*(p) = E(p) = \mu_p$. This holds true regardless of whether the true belief follows a non-degenerate distribution or whether the DM is forced to choose a single point as their belief.

Proposition 2. *Given the true belief p , when the DBM is used to elicit belief without forcing the DM to choose a single point as their belief and BSR is used to determine payoff, an expected utility maximizer will choose until reaching the point $a^*(p) = E(p) = \mu_p$.*

3.2.3 Incentive Compatibility with DBM for Myopic EU Maximizer

If the DM is myopic – fails to perfectly foresee that the optimal choice is to choose until the point a^* where $a^* = E(p) = \mu_p$ and $Var(a) = 0$ in the DBM, they may compare among the three options in each step instead: choosing $Uniform(I_l, \frac{I_l+I_u}{2})$, choosing $Uniform(\frac{I_l+I_u}{2}, I_u)$, or choosing $Uniform(I_l, I_u)$. Whenever $E(p) < \frac{I_u+I_l}{2}$, choosing $Uniform[I_l, \frac{I_l+I_u}{2}]$ yields a higher likelihood of receiving M_h than choosing $Uniform(\frac{I_l+I_u}{2}, I_u]$ or exiting with $Uniform[I_l, I_u]$. Similarly, whenever $E(p) > \frac{I_u+I_l}{2}$, choosing $Uniform(\frac{I_l+I_u}{2}, I_u]$ yields a higher likelihood of receiving M_h than the other two options. Whenever $E(p) = \frac{I_u+I_l}{2}$, all three options yield the same likelihood of receiving M_h . Thus, the DM would be indifferent in choosing any of the three options.² Detailed proof can be found in Appendix C.2. In other words, whenever $E(p)$ is strictly within one of the two narrowed intervals, it is optimal to choose the one that contains $E(p)$. Otherwise, the myopic DM is indifferent between choosing $Uniform[I_l, \frac{I_l+I_u}{2}]$, choosing $Uniform(\frac{I_l+I_u}{2}, I_u]$, or exiting with $Uniform[I_l, I_u]$.

Proposition 3. *If the DM is myopic – fails to perfectly foresee that choosing until $a^* = E(p)$ is optimal in the DBM, then, in each step, they will be indifferent among $[I_l, \frac{I_l+I_u}{2}]$, $(\frac{I_l+I_u}{2}, I_u]$, or exiting with $[I_l, I_u]$, whenever $E(p) = \frac{I_l+I_u}{2}$. Otherwise, it is optimal to always choose the interval which strictly contains $E(p)$.*

In sum, the midpoint of the DM's stated belief, whether it is an interval or not, reveals the mean of their true belief for an expected utility maximizer. Additionally, without further behavioral assumptions, early exit in the DBM indicates whether the expected utility maximizer is myopic – fails to perfectly foresee that choosing until $a^* = E(p)$

²For continuous uniform distribution, whether $E(p) = \frac{I_l+I_u}{2}$ is contained in the left interval or right interval does not matter. For discrete uniform distribution, the myopic DM will be indifferent among $[I_l, \frac{I_l+I_u}{2}]$, $[\frac{I_l+I_u}{2} + 1, I_u]$, or exiting with $[I_l, I_u]$, whenever $E(p) = \frac{I_l+I_u}{2} + \frac{1}{2}$.

is optimal. This is orthogonal to whether their true belief is a precise singleton or an imprecise interval.

3.3 Experimental Design

In order to explore the empirical validity of DBM, we design the experiment with a collection of perception tasks that are used in the existing literature, and use the slider bar version of the CM as the benchmark, which allows both within-subject and between-subject investigations.

3.3.1 DBM and CM

We employ DBM to probe subjects' beliefs in a step-by-step manner. Initially, participants are queried about their assessment of probability relative to 50%. Subsequently, based on their response, they are prompted to determine whether the likelihood is below or above 25%, and this process continues iteratively. At each step, subjects are presented with two exclusive interval choices and the option to "Exit." Upon reaching the final step, participants must provide a point belief. Figure 3.1 shows the experimental interface of the DBM.

Figure 3.1: DBM: Experimental Interface

What do you think is the probability that the **Red Box** has been selected?

0 10 20 30 40 50 60 70 80 90 100

between 0% and 50% between 51% and 100%

Now, let's zoom in 0% ~ 50%.

0 10 20 30 40 50

between 0% and 25% between 26% and 50%

Now, let's zoom in 0% ~ 25%.

0 5 10 15 20 25

between 0% and 12% between 13% and 25%

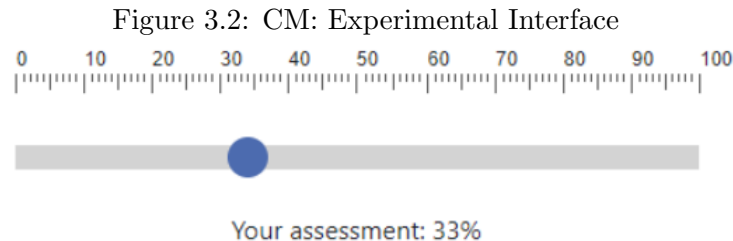
Exit

In instances where a subject provides a point belief, compensation is awarded based on the Binary Scoring Rule (BSR). Conversely, if an interval belief is reported, a random number within the specified interval is drawn from a uniform distribution. Subsequently, compensation is determined according to the BSR using the generated number.

We use the slider bar version of the CM to elicit subjects' probabilistic beliefs or perceptions, as it is widely used in many experimental studies.³ To ensure accuracy and relevance to the task's objective truth, we offer three distinct scales: the 100 scale, 4000 scale, and 250 scale. Notably, we deliberately avoid providing a default position on the slider bar to mitigate any potential anchoring effects. Additionally, above the slider bar,

³In the rest of the paper, we always refer CM as to the slider bar version of CM.

we include a ruler for subjects' reference, aiding in their precise assessment. Figure 3.2 shows the experimental interface of the CM.



3.3.2 Experiment 1: Within-subject Design

In Experiment 1, we use a within-subject design, allowing each subject to experience both the CM and DBM in a random order. To be more specific, the experiment consists of two blocks, with each block employing either CM or DBM to elicit subjects' beliefs or perceptions. Within each block, we use five different task categories regarding probabilistic beliefs or perceptions that are commonly used in existing literature. This design choice allows us to assess the generalizability of aggregate performances in the DBM compared to those in the CM.

Within each block, four of the five task categories use artificial settings such as balls and urns, peas, or dots, which are common in the laboratory experiments. These task categories include reporting prior belief (Danz, Vesterlund and Wilson, 2022), belief updating (Danz, Vesterlund and Wilson, 2022), forming compound prior belief (Liang, 2022), and estimating number of peas in a bowl (Falk and Zimmermann, 2018) or dots in a graph (Dewan and Neligh, 2020). The fifth task category involves subjects' perceptions or probabilistic beliefs about real economic variables, specifically the inflation rate and the S&P500 (Enke and Graeber, 2023). That is, there are two tasks within each task category: one from each block.

Each task has one objective truth, allowing us to measure the accuracy of subjects’ beliefs or perceptions objectively. Detailed questions used in the experiment can be found in Appendix C.4. To prevent anchoring, we carefully choose parameters so that the objective truths in all tasks are spread across the entire range between 0 and 1.⁴ To ensure comparability within the same task category, we deliberately choose parameters for tasks under the same category in the two blocks so that their objective truths are symmetric around 50%.⁵ Table 3.1 demonstrates the five task categories and the corresponding objective truths used in the two blocks.

Table 3.1: Task Categories and Objective Truths in Experiment 1

	Task Category	Block 1	Block 2
Laboratory/Artificial Settings	Prior Belief	20%	80%
	Belief Updating	33%	67%
	Compound Prior	60%	40%
	Estimation	(peas) 3000/4000	(dots) 120/250
Real-life Settings	Econ Variable	(inflation rate) 92%	(S&P 500) 8%

Note: Within each block, the tasks and parameters are fixed, but the order of tasks is randomly determined.

Within each block, the tasks and parameters are fixed, but the order of tasks is randomly determined. We implement two treatments, Treatment DBM-CM and Treatment CN-DBM, by alternating the order of CM and DBM used to elicit beliefs or perceptions in the tasks of each block. That is, in Treatment DBM-CM, DBM is used to elicit subjects’ beliefs or perceptions in Block 1, followed by CM in Block 2. Conversely, in Treatment CM-DBM, CM is used in Block 1, followed by DBM in Block 2. This design choice allows us to investigate the interaction between DBM, learning, and experience.

⁴For the estimation tasks involving counting peas in a bowl or dots in a graph, where the scales are 0 – 4000 and 0 – 250 respectively, we transform these into a 0 – 100 scale to avoid duplicated objective truths with other probabilistic tasks.

⁵The objective truths in the estimation tasks involving counting peas in a bowl or dots in a graph do not have this property as the scales are different.

3.3.3 Experiment 2: Between-subject Design

The design of Experiment 1 could make differences across task categories difficult to interpret, as both the tasks and their objective truths vary. Subjects might differ in their expertise across task categories, and the measured accuracy may be influenced by different objective truths used. Empirical evidence demonstrates that individuals' subjective beliefs tend to be center-biased (Danz, Vesterlund and Wilson, 2022) or compressed towards an "intermediate" value, such as midpoint of a slider bar (Enke and Graeber, 2023). Thus, even with the same subjective beliefs, performance may appear better when tasks have more centered objective truths (40% - 60%) compared to those with more extreme values (0% - 10%, or 90% - 100%) and those with intermediate values (10% - 40%, or 60% - 90%).

To address this concern, we design Experiment 2 with two main features: (1) within the same task category, we employ more objective truths that span the entire range between 0 and 1; and (2) we include three additional task categories with real-life settings alongside the existing belief updating tasks and perception tasks on inflation rate.

To be more specific, in Experiment 2, each subject needs to finish five tasks: two replicated from Experiment 1 (belief updating tasks and perception tasks about inflation rates) and three new perception tasks about real economic variables (income (Wiswall and Zafar, 2015*b*), unemployment rate, and education level).⁶ By replicating tasks from Experiment 1, we can test the robustness of the results. Combining Block 1 of Experiment 1 and Experiment 2 provides a balanced set of tasks between laboratory/artificial settings and real-life settings and mitigates the learning effects. Each subject see the five tasks in a random order.

Within each task category, every subject randomly receives one of three parameters

⁶We generate questions about the unemployment rate and education level with objective truths using a method similar to the tasks on inflation rates in Enke and Graeber (2023).

corresponding to one of three types of objective truths: centered truths (40% - 60%), extreme truths (0% - 10% or 90% - 100%), and intermediate truths (10% - 40% or 60% - 90%), each of which is equally likely to occur. Table 3.2 depicts the task categories and objective truths used in Experiment 2. This design choice allows us to distinguish the role of truth types from the impact of varied expertise across different task categories. We implement two treatments, Treatment CM and Treatment DBM, by using different methods to elicit beliefs or perceptions.

Table 3.2: Task Categories and Objective Truths in Experiment 2

	Task Category	Centered Truth	Intermediate Truth	Extreme Truth
Real-life Settings	Inflation Rate	56%	77%	92%
	Income	45%	30%	7%
	Unemployment Rate	56%	84%	98%
	Education Level	49	12	3
Laboratory/Artificial Settings	Belief Updating	47%	33%	6%

Note: The order of tasks is randomly determined.

3.3.4 Implementation and Recruitment Details

We recruited all subjects on Prolific, an online platform frequently used for research studies. To qualify for our study, subjects were required to have a minimum of 100 prior submissions on Prolific, with an approval rate of at least 98%. We implemented the experiment using the oTree platform (Chen, Schonger and Wickens, 2016). For Experiment 1, we recruited 102 subjects, with 51 subjects assigned to each order of methods. For Experiment 2, we recruited 149 subjects, with 72 subjects using CM and 77 subjects using DBM to elicit beliefs in the five tasks. Each participant also received a \$3 completion payment and took around 20 minutes to complete the study. In each experiment, subjects receive detailed instructions and are required to correctly answer comprehension questions before proceeding to the main parts of our study.

3.4 Results

We start by comparing the aggregate performance between DBM and CM using pooled data from Block 1 of Experiment 1 and Experiment 2, as shown in Section 3.4.1. We find that DBM does not perform significantly different from CM at the aggregate level. Next, we investigate circumstances where DBM might outperform CM in Section 3.4.2. This includes examining whether the objective truth has extreme, centered or intermediate values, and whether the task context involves laboratory/artificial or real-life settings. We use pooled data from Block 1 of Experiment 1 and Experiment 2 to study these factors. We find that DBM outperforms CM when the objective truth is extreme, while CM outperforms DBM with intermediate objective truths. However, with centered truths and across task types, DBM does not perform differently from CM.

Then we analyze to what extent the length of stated beliefs in DBM informs the accuracy of subjects' beliefs in Section 3.4.3. Using pooled data from Block 1 of Experiment 1 and Experiment 2, we document that for stated interval beliefs in DBM ($a_l \neq a_u$), the shorter the interval, the more accurate the stated belief. However, stated point beliefs in DBM, which constitute a significant fraction of all stated beliefs, are not the most accurate. We find similar results using data from both blocks of Experiment 1 for within-subject analysis. Finally, we compare multiple methods for using stated beliefs in DBM to predict stated point beliefs in CM in Section 3.4.4 and discuss how to effectively utilize the data collected with DBM.

3.4.1 DBM vs. CM: Aggregate Performance

To test the empirical performance of DBM, we compare the accuracy of beliefs or perceptions elicited in DBM with those in CM. This requires a measure of accuracy. We mainly focus on two measures, given their stated belief $a = [a_l, a_u]$ and the objective

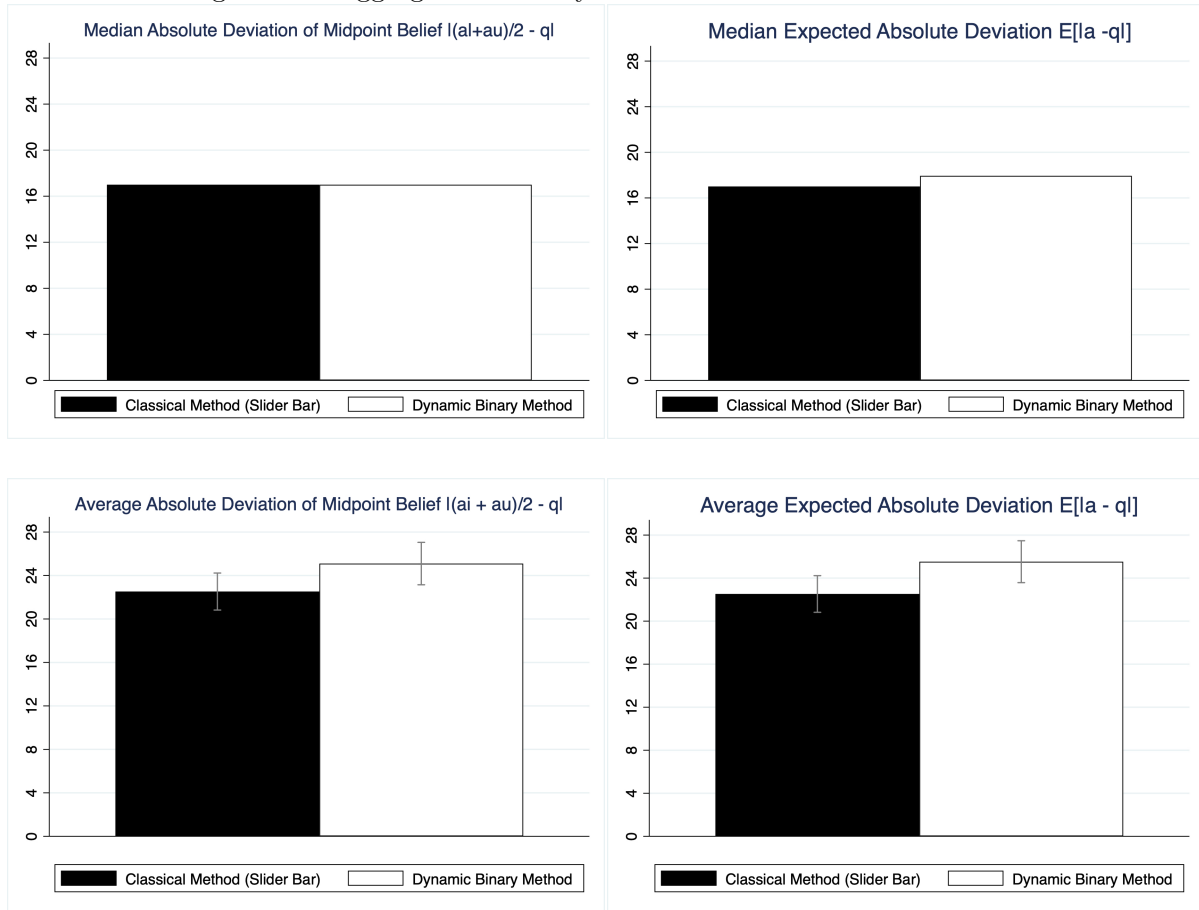
truth q :

1. Absolute difference between the midpoint of their stated beliefs and the objective truth (ADM): $|\frac{a_l + a_u}{2} - q|$; ;
2. Expected absolute difference between their stated beliefs and the objective truth (EAD): $E_a[|a - q|]$.

Note that the use of midpoint in the first measure is justified by the theoretical framework that, for an expected utility maximizer, whether myopic or not, the midpoint of their stated belief reveals the mean of their true belief. Moreover, if the stated belief is a singleton, that is, $a_l = a_u$, the two measures are equivalent to $|a - q|$ – the absolute difference between stated belief and the objective truth.

Figure 3.3 demonstrates the median and mean accuracy of stated beliefs elicited with DBM versus CM using the two measures mentioned above. The median accuracy of stated beliefs elicited with DBM is not significantly different from those with CM, and this finding is robust across the measures used. Specifically, we conduct quantile regression of the measured accuracy on the dummy variable indicating which elicitation method is used (DBM or CM), controlling for gender and self-reported familiarity with statistics. The estimated coefficient on the elicitation method dummy variable is not significantly different from zero even at the 90% level ($p = 0.240$ for ADM and $p = 0.148$ for EAD).

Figure 3.3: Aggregate Accuracy of Stated Beliefs in DBM vs. CM



Note: Each graph uses pooled data from Block 1 of Experiment 1 and Experiment 2. For the bottom two graphs of average accuracy, we plot the 95% confidence intervals.

The mean accuracy of stated beliefs elicited using DBM is slightly lower than those using CM. We use OLS regression of the measured accuracy on the dummy variable indicating which elicitation method is used (DBM or CM), controlling for gender and self-reported familiarity with statistics.⁷ The estimated coefficient on the elicitation method dummy variable is significantly different from zero at the 90% level: the average ADM using DBM is 2.88 larger than that using CM ($p = 0.098$), and the average EAD with DBM is 3.27 larger than that with CM ($p = 0.061$). This finding indicates that the

⁷All the regression models have gender and self-reported familiarity with statistics controlled.

aggregate performance of DBM is not significantly different from CM, although DBM exhibits slightly larger variance.⁸

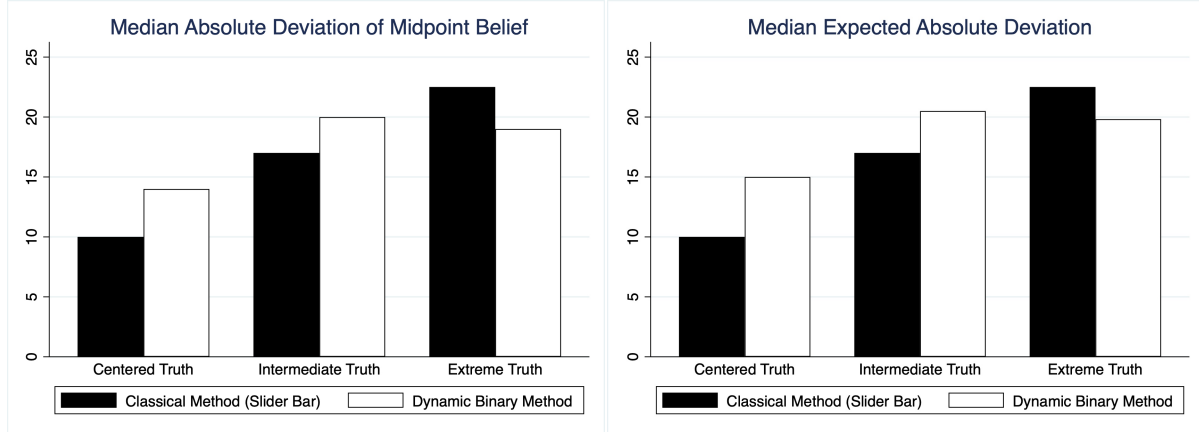
Result 9. *At the aggregate level, DBM does not perform significantly different from CM: the accuracy of stated beliefs elicited by the two methods are not significantly different.*

3.4.2 When does DBM Outperform CM?

Objective Truth Type. The null result at the aggregate level could be because DBM draws subjects' attention to non-centered values, thereby reducing subjects' tendency to choose numbers centered at the midpoint of the slider bar as their stated beliefs in each task. Figure 3.4 presents the median accuracy of beliefs elicited with DBM and CM using two measures, separated by the three types of objective truth: centered truths (40% - 60%), extreme truths (0% - 10% or 90% - 100%), and intermediate truths (10% - 40% or 60% - 90%).

⁸Similar to Enke and Graeber (2023)'s study and given that the directional results using average accuracy as the outcome variable are consistent with those using median accuracy but exhibit much larger variance, we primarily focus on median accuracy in the rest of the analysis to demonstrate the aggregate results of interest in the main draft.

Figure 3.4: Median Accuracy of Stated Beliefs in DBM vs. CM by Objective Truth Type



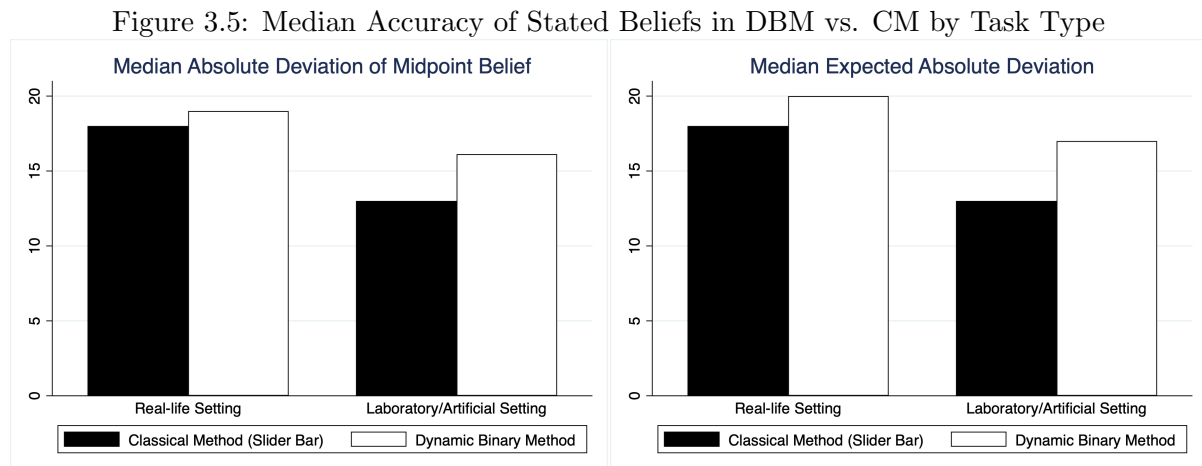
Note: Each graph uses pooled data from Block 1 of Experiment 1 and Experiment 2. Centered truths denote tasks with objective truths between 40% and 60%, extreme truths denote tasks with objective truths between 0% and 10% or between 90% and 100%, and intermediate truths denote tasks with objective truths between 10% and 40% or between 60% and 90%.

Consistent with Figure 3.4, the median accuracy of stated beliefs in DBM is significantly higher than in CM at the 95% confidence level when the objective truths are extreme (quantile regression, $p = 0.033$ for ADM and $p = 0.032$ for EAD). However, the median accuracy of stated beliefs in DBM is significantly lower than that in CM when the objective truths are intermediate (quantile regression, $p = 0.032$ for ADM and $p = 0.023$ for EAD). There are no significant differences for centered objective truths (quantile regression, $p = 0.134$ for ADM and EAD). This finding suggests that some deviations from the objective truths in CM could be attributed to the narrowed consideration set that subjects use to state beliefs or perceptions. Our new method, DBM, aids subjects by expanding the range of available numbers they consider.

Result 10. *DBM outperforms CM when the objective truth is extreme, while CM outperforms DBM with intermediate objective truths. However, there is no significant performance difference between DBM and CM when the objective truth is centered.*

Task Type. In addition to that, DBM may guide subjects to think through each task in

a step-by-step manner, which could help retrieve information and past experiences from memory, especially for tasks with real-life settings that do not provide all the information needed for answering the task question correctly. Figure 3.5 demonstrates the median accuracy of beliefs elicited with DBM and CM using two measures, separated by the two types of tasks: tasks with real-life settings which involves subjects' perceptions of inflation rates, unemployment rates, income distribution, and education levels by state; and tasks with laboratory/artificial settings which includes those on prior beliefs, belief updating, counting, and compound priors.



Note: Each graph uses pooled data from Block 1 of Experiment 1 and Experiment 2. Tasks with laboratory/artificial settings include those on prior beliefs, belief updating, counting, and compound priors. Tasks with real-life settings involve subjects' perceptions of inflation rates, unemployment rates, income distribution, and education levels by state.

As shown in Figure 3.5, the median accuracy of beliefs elicited using DBM is not significantly different from CM in tasks with real-life settings (quantile regression, $p = 0.489$ for ADM and $p = 0.484$ for EAD).⁹ In tasks with laboratory/artificial settings, the median accuracy using DBM is slightly lower than CM, but the significance of this result depends on the measure of accuracy (quantile regression, $p = 0.126$ for ADM and

⁹As the results with ADM are similar to those with EAD, we primarily use EAD as the measure of accuracy in the remainder of the analysis.

$p = 0.038$ for EAD). This indicates that DBM does not perform differently from CM across task types.

Result 11. *DBM does not perform differently from CM regardless of whether the task utilizes a laboratory/artificial setting or a real-life setting.*

3.4.3 Is the Interval Length Informative About Accuracy?

This section explores the relationship between the length and the accuracy of stated beliefs in the tasks using DBM. To ensure the lengths of stated beliefs are comparable across tasks, for all the analysis in this subsection, we use only tasks with a choice scale of 100, which rules out tasks with counting peas in a bowl and counting dots in a graph in Experiment 1. To achieve this goal, we start with pooled data from Block 1 of Experiment 1 and Experiment 2 to conduct a between-subject analysis, investigating how the median accuracy of stated beliefs in tasks using DBM varies with the number of steps taken. Additionally, we use the data from Blocks 1 and 2 in Experiment 1 to explore, from a within-subject perspective, to what extent the length of a subject’s stated belief in tasks using DBM can predict how well they perform in tasks using CM.

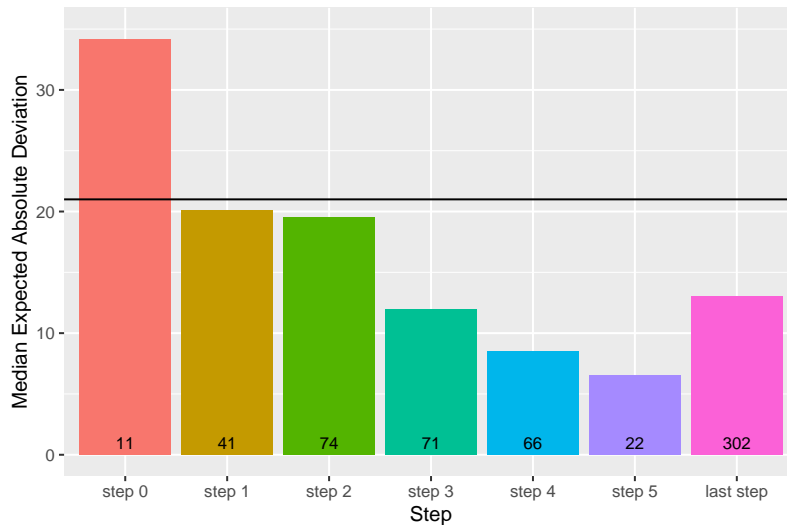
Our theoretical framework shows that an exit early before reaching the point indicates that the expected utility maximizer is myopic — failing to perfectly foresee that choosing until $a^* = E(p)$ is optimal, regardless of the precision of their true beliefs. Conversely, choosing until the end suggests that the expected utility maximizer is not. If this is the case, we would expect no correlation between the length of stated beliefs (number of steps taken) and their accuracy.

Figure 3.6 plots the median EAD of stated beliefs using DBM against the number of steps taken. Generally, the median EAD of stated beliefs decreases as the number of steps increases, with the correlation being significantly different from zero at the 95%

confidence level (quantile regression, $p = 0.018$). However, for the stated point beliefs in DBM, which constitute 51% of all the stated beliefs, the median EAD is slightly higher than those exiting right before the last step (i.e., Step 5). Similar patterns are observed when separated by task types and by objective truth types, as shown in Figure 3.7. This finding could result from overconfidence – where individuals overestimate the precision of their perceptions – or from risk aversion – where individuals dislike uncertainty in their reported beliefs. Distinguishing between potential mechanisms could be a fruitful direction for future research.

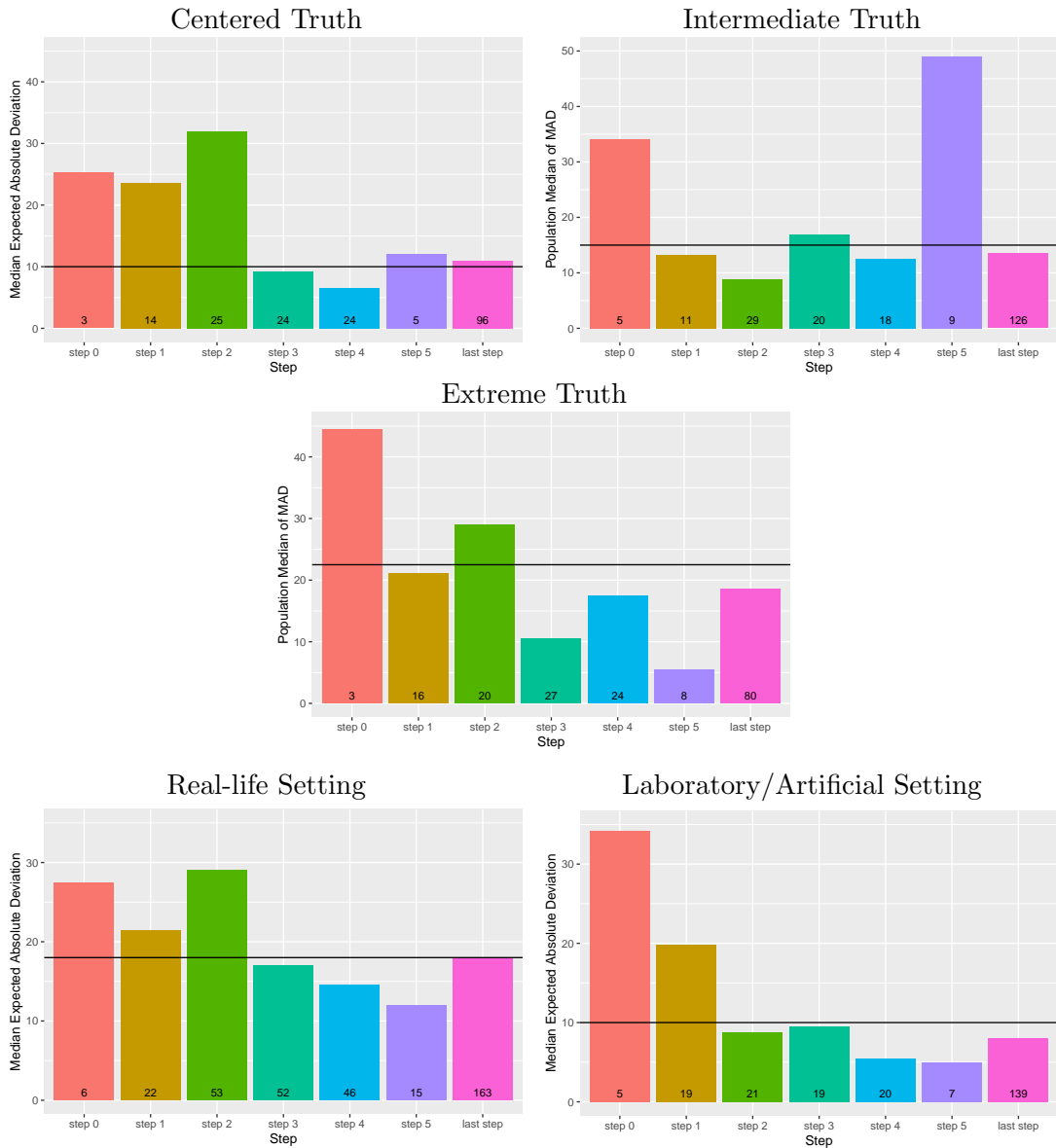
Result 12. *The length of stated beliefs in DBM is negatively correlated with their own accuracy at the aggregate level: the longer the interval, the less accurate the stated belief. However, this relationship is not strictly monotonic: stated beliefs reaching the point are not the most accurate ones.*

Figure 3.6: DBM: Median Accuracy of Stated Beliefs and Number of Steps Taken



Note: The number labeled inside each bar is the number of stated beliefs that exit in each step. The black horizontal line is located at the median expected absolute deviation with all the stated beliefs using CM pooled.

Figure 3.7: DBM: Accuracy and Number of Steps Taken by Objective Truth Type and Task Type



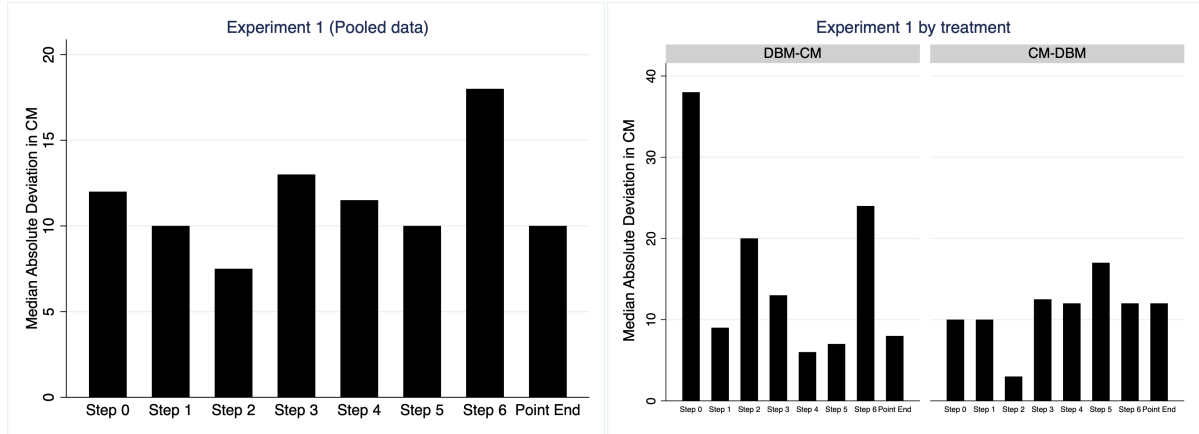
Note: The number at the bottom of the bar is the number of observations. The black horizontal line is the median absolute deviation in CM.

Moreover, we use data from Blocks 1 and 2 in Experiment 1 to study this question from a within-subject perspective. Specifically, since tasks from the same category between blocks have symmetric objective truths, we pair tasks from the same task category – one using DBM to elicit beliefs and the other using CM. Within each pair, we investigate the extent to which the length of stated belief in the task using DBM can predict the accuracy of stated belief in the the task using CM.

Figure 3.8 plot the median absolute deviation in tasks using CM against the number of steps taken in the paired tasks using DBM. When pooling Treatments DBM-CM and CM-DBM, there is no significant correlation between the median absolute deviation of stated beliefs with CM and the number of steps taken in their paired tasks with DBM. This mainly results from the null effect in Treatment CM-DBM.¹⁰ When separating the data by treatment, we find that in Treatment DBM-CM – using DBM in Block 1 and CM in Block 2 – the median absolute deviation of stated beliefs with CM significantly decreases as the length of stated beliefs increases (quantile regression, $p > 0.001$). Similar to the between-subject analysis discussed earlier, the relationship is not strictly monotonic: point or close-to-point beliefs in tasks using DBM do not predict the lowest median absolute deviation in paired tasks using CM.

¹⁰One reason the data in Treatment DBM-CM are much noisier is that subjects took longer than expected to complete Experiment 1, resulting in a base payment that was considered as lower than the recommended hourly rate by Prolific. Since DBM requires more time for subjects to think through and submit their beliefs, the quality of choices decreases due to fatigue when DBM is used in Block 2.

Figure 3.8: Median Accuracy of Stated Beliefs in CM and Number of Steps Taken in DBM

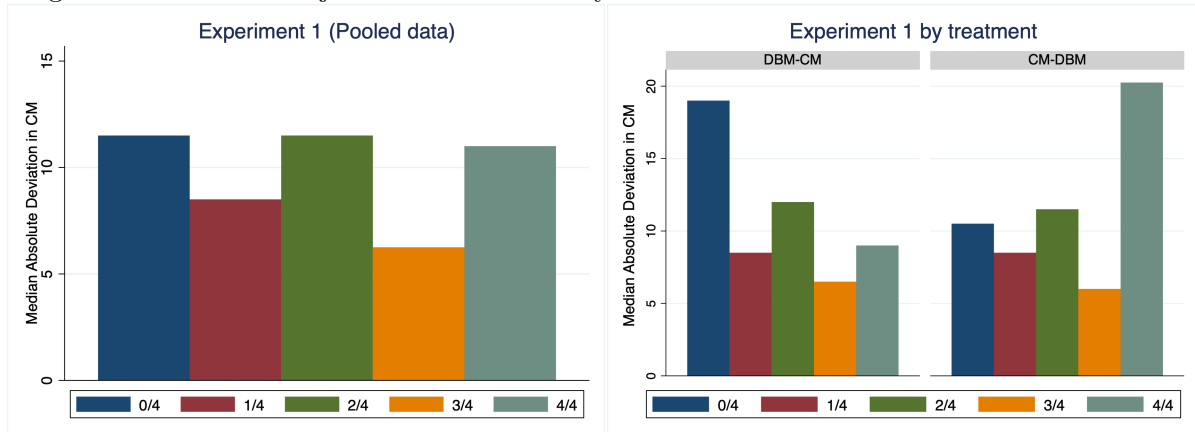


Note: X-axis denotes the number of steps taken by the stated beliefs in DBM. Each graph uses pooled data from Experiment 1. The left panel pooled data from Treatments DBM-CM and CM-DBM together, while the right panel separates results by treatment.

A similar pattern is observed in Figure 3.9, where we calculate, for each subject, the median accuracy of stated beliefs among the four paired tasks using CM, separated by the number of point beliefs they stated among the four tasks using DBM. Using pooled data, among those who state interval beliefs in at least one of the four tasks using DBM, the median subject who is more likely to state point beliefs in tasks using DBM is also more accurate in tasks using CM (quantile regression, $p = 0.015$). However, for subjects who state point beliefs in all the four tasks using DBM, the median subject is not the most accurate one in tasks using CM. This pattern is even stronger in Treatment DBM-CM (quantile regression, $p = 0.03$).

Result 13. *The length of stated beliefs in tasks using DBM is negatively correlated with the accuracy of stated beliefs in tasks using CM: the longer the interval in DBM, the less accurate the stated belief in CM. However, this relationship is not strictly monotonic, as stated beliefs reaching the point in DBM do not predict the most accurate beliefs stated in CM.*

Figure 3.9: Within Subject: Median Accuracy in CM and Fraction of Point Beliefs in DBM



Note: X-axis denotes the fraction of point beliefs in DBM. Each graph uses pooled data from Experiment 1. The left panel pooled data from Treatments DBM-CM and CM-DBM together, while the right panel separates results by treatment.

3.4.4 Predicting Point Beliefs Elicited with CM

In this section, we explore multiple methods for using the stated beliefs elicited with DBM to predict point beliefs elicited with CM. To achieve this, we primarily focus on data from Experiment 1 with a scale of 100. We use the paired stated beliefs in tasks using DBM to predict the corresponding beliefs in tasks using CM. To ensure comparability, we symmetrize the stated belief and objective truth for one task in each pair. There are several methods to utilize the data elicited with DBM:

1. Midpoint Prediction: according to our theoretical framework, the midpoint of stated beliefs using DBM serves as a natural predictor for stated point beliefs in CM:

$$\hat{a}_1 = \frac{a_l + a_u}{2}$$

where a_l and a_u denote the lower and upper bound of stated beliefs in DBM. For an expected utility maximizer, they will select the mean of their true belief as stated belief

in CM and will choose the belief whose midpoint is equal to the mean of their true belief in DBM.

2. Cognitive Default, Cognitive Noise, and Objective Truth: Existing studies indicate that individuals' stated beliefs in CM often compress towards a cognitive default (e.g., the center of the slider bar) (Danz, Vesterlund and Wilson, 2022; Enke and Graeber, 2023). This phenomenon can be modeled as a weighted average between the utility-maximizing decision $a^*(p)$ and the cognitive default d . The relative weight on d is determined by the cognitive noise or uncertainty λ (Enke and Graeber, 2023):

$$\hat{a}_2 = (1 - \lambda) * a^*(p) + \lambda * d$$

The greater the cognitive noise, the stronger the tendency to state the default d (e.g., center of the slider bar) in CM. To construct the predicted beliefs in CM, we use a straightforward method to determine λ : $\lambda = \frac{a_u - a_l}{100}$, where λ is the length of stated beliefs in the paired tasks using DBM relative to the scale. Using the objective truth $p = a^*(p)$ and the default $d = 50\%$, we can generate the predicted point beliefs in CM.

3. Cognitive Default, Cognitive Noise, and Subjective Truth: We propose a revised version of Method 2 by replacing the objective truth, which could be equally difficult for subjects with bounded rationality to perceive, with the subjective truth – the midpoint of stated beliefs in DBM:

$$\hat{a}_3 = (1 - \lambda) * \frac{a_l + a_u}{2} + \lambda * d$$

where $\lambda = \frac{a_u - a_l}{100}$ and the default $d = 50\%$.

Table 3.3 show the average predicted point beliefs using stated beliefs in DBM via the three methods mentioned above, separated by task category. Using t-tests to contrast with the average stated beliefs in CM, we find that the predicted beliefs using Method 3 – weighted average between cognitive default and subjective truth – are closest to the stated beliefs in CM. Specifically, using Method 3, the predicted beliefs in tasks with economic variables are not significantly different from the average stated beliefs in CM. Additionally, using Method 3, the predicted beliefs in tasks about Simple Prior and Compound Prior differ from those in CM at a 95% confidence level, which is a much smaller difference compared to the other two methods.

Table 3.3: Predicted Point Beliefs using Data in DBM and Stated Point Beliefs in CM

Task Category	Stated Beliefs in CM		Method 1		Method 2		Method 3
Simple Prior	30.3%	**	35.1%	***	22.2%	**	36.1%
Compound Prior	54.9%	***	61.9%	***	59.3%	**	61%
Posterior	43.1%	***	51.1%	***	35.2%	***	51%
Econ Variables	67.3%		70.8%	***	87.1%		69.9%

Note: Reported significance stars are based on a two-way t-test to determine whether the difference between the average stated beliefs in CM and the average predicted beliefs using each of the three methods is significantly different from zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Conversely, the predicted point beliefs using Method 2 – weighted average between cognitive default and subjective truth – are significantly different from those in CM at the 99% confidence level in each of the task categories. Method 1, which uses the midpoint, performs somewhere in between the other two methods. One plausible reason for the worst performance of Method 2 is that its predictions are constrained between the objective truth and the default, failing to capture stated beliefs in CM that fall outside this range.

3.5 Conclusion

In this paper, we propose a novel method, the *Dynamic Binary Method* (DBM), to elicit people’s beliefs or perceptions. Beliefs and perceptions are central to studying economic behavior, yet accurately eliciting them presents significant challenges. Existing elicitation methods that involve soliciting respondents’ best point estimates are susceptible to various biases, such as the cognitive difficulty of pinpointing imprecise beliefs. Unlike *Classical Methods* (CM), which require absolute judgments on the best point estimates of beliefs, DBM, inspired by the bisection method, prompts respondents to make a series of binary relative judgments. This approach allows respondents to state interval beliefs by exiting the process at any step before reaching a final point estimate.

To assess the empirical validity of DBM, we use a collection of perception tasks from existing literature, construct both between-subject and within-subject experiments, and use the slider bar version of CM to benchmark how well respondents would perform in each task.

The main finding is that, at the aggregate level, DBM does not perform significantly differently from CM, regardless of whether the perception question uses artificial/laboratory settings or real-life settings. This finding is robust to different measures we use, including the absolute deviation of the midpoint of elicited beliefs from the objective truth or the expected absolute deviation of elicited beliefs from the objective truth. However, DBM outperforms CM when the task has extreme values as the objective truth. This suggests that some biases in perception questions could result from the narrowed consideration set that respondents use to choose beliefs or perceptions from.

Furthermore, we find, from both between-subject and within-subject perspectives, that the length of stated beliefs in tasks using DBM is negatively correlated with their own accuracy and can predict how well respondents perform in tasks using CM at the

aggregate level: the longer the interval, the less accurate the stated belief in DBM and the less accurate the stated belief in CM. However, this relationship is not strictly monotonic: stated beliefs reaching the point are not the most accurate ones and do not predict the most accurate beliefs stated in CM.

Lastly, we compare three methods of using the stated beliefs elicited with DBM to predict point beliefs elicited with CM. We find that predictions using a weighted average between subjective truth (the midpoint of stated beliefs in DBM) and the cognitive default (e.g., midpoint of the slider bar), with the relative weight on the default determined by the length of stated beliefs in DBM, are closest to the average stated beliefs in CM. This approach outperforms both using the midpoint of stated beliefs in DBM alone and using objective truth instead of subjective truth in the weighted average method. Our findings underscore the significance of incorporating the precision of stated beliefs and perceived truth to enhance predictions of economic behavior.

Our results also raise several intriguing questions for future research. As explored in the literature review, there are non-incentivized methods for identifying preference incompleteness, taste imprecision, or the distribution of beliefs. It would be valuable to compare these methodologies to understand the degree to which they capture the same uncertainty in beliefs or perceptions and how this may differ between preference incompleteness and belief imprecision.

Additionally, it would be beneficial to gain a deeper understanding of the circumstances under which individuals possess precise versus imprecise beliefs. Such insights could aid in interpreting standard belief data and potentially enable the identification of imprecision even when individuals are unable to directly report it. Furthermore, it would be fascinating to explore any neurological or biological indicators of imprecision, which could provide a new dimension to understanding how individuals form and report their beliefs or perceptions.

Appendix A

Appendix for “What Drives Probability Matching?”

A.1 Theories of Mixing

As discussed in the paper, based on whether a model predicts that the mixing behavior varies with the correlation change and frame change in my environment, I can divide these models into three categories: models of *Correlation-Invariant Stochastic Choices*, models of *Correlation-Sensitive Stochastic Choices*, and *Framing Effects*. Here I describe in further detail other example theories in the first two classes.

A.1.1 Models of Correlation-Invariant Stochastic Choices

Models in this class share two features: (1) do not incorporate the state-wise outcome comparisons between options into the decision making process, which, by nature, has no room for mixing behavior to vary with different correlations between options; (2) fail to combine the evaluation of each option with the choice distribution $(\alpha, 1 - \alpha)$ in a non-linear way, which does not provide a channel for different correlations to interact with

individuals’ choice distributions differently.

A.1.1.1 Correlation-Orthogonal Preferences/Heuristics

Most preference-based and some heuristics-based stochastic choice models do not incorporate the state-by-state comparisons of outcomes between options in the decision making process. Thus, these preferences or heuristics are orthogonal to the correlation between options. So does the mixing behavior predicted by them. Here I describe two models in detail.

Probability Weighting The probability weighting model is an example of the preferences over reduced lotteries which allow for dominance violations. It assumes that the reduced lottery \mathcal{L} in Equation (1.1) is evaluated according to

$$\sum_x \omega(P(x))u(x) = \omega(\alpha * p + (1 - \alpha) * (1 - p))u(\$M) + \omega(\alpha * (1 - p) + (1 - \alpha) * p)u(\$0)$$

for some onto and increasing weighting function $\omega(\cdot) : [0, 1] \rightarrow [0, 1]$. With certain $\omega(\cdot)$, it can predict mixing in decision problems with the CPS. However, as the marginal distributions are fixed, the mixing behavior neither varies with the change in correlation between options in the Baseline Scenario, nor with the corresponding frame change in the Independence Scenario.

Expectation Matching Kogler and Kühberger (2007) propose that probability matching is the result of System 1 processing with the dual process theories. According to this perspective, the decision maker unconsciously states the expected outcomes as her predictions of which one is more likely to occur. For example, if Project A has a 75% chance to succeed and Project A has a 25% chance to succeed, the decision maker employs the

frequencies that Project A succeeds three out of four times and Project B succeeds once out of four times to allocate their tickets. Importantly, the interdependence of outcomes between options plays no role during this process. As a result, this theory predicts that mixing behavior does not vary with changes in the correlation between options.

A.1.1.2 Correlation-Dependent Preferences

While some theories do account for state-wise comparisons of outcomes between options, denoted as *Correlation-Dependent Preferences*, they still fail to predict varied mixing behavior in response to different correlations between options for various reasons. The common implication shared by these models is that once the marginal distributions ($A : p, B : 1 - p$) are fixed, the optimal choice distributions are determined.

Correlation-Orthogonal Preferences/Heuristics Most preference-based and some heuristics-based stochastic choice models posit that individuals evaluate each option based on its marginal distribution rather than considering how options are jointly determined. Thus, these preferences or heuristics are, by nature, orthogonal to the correlation between options, as is the mixing behavior predicted by them. Intuitively, for individuals who evaluate each option based on all of its absolute attributes and then choose the one with the higher evaluation (or mix if they are indifferent), their decision-making process is not influenced by the relative outcomes of each option compared to the alternative in each state. As a result, the sources of mixing are unrelated to how options are jointly determined. These theories predict that once the marginal distributions of options are fixed, mixing behavior does not vary.

Correlation-Dependent Preferences Some preference-based theories account for state-wise comparisons between options by proposing that the decision maker assigns weights to different states of the world based on the outcome differences between the

alternatives in each state. However, due to their inherent characteristics or without further assumption on the convexity of preference, the sources of mixing proposed by these theories are also orthogonal to how options are correlated in between, which yields identical predictions in decision problems across the three scenarios. Correlation-sensitive preferences (Lanzani, 2020) and the pairwise normalization model (Landry and Webb, 2021) are examples in this category.¹

To illustrate this, consider the correlation-sensitive preference (Lanzani, 2020), which nest regret-averse preference (Loomes and Sugden, 1982) and salience theory (Bordalo, Gennaioli and Shleifer, 2012). In these models, the decision maker follows several steps to evaluate the risky options: (1) for each possible joint realization of outcomes $(x, y) \in \{(A : \$M, B : \$M), (A : \$M, B : \$0), (A : \$0, B : \$M), (A : \$0, B : \$0)\}$, compares the two outcomes and gives a score, denoted as $\phi(x, y)$. This score reflects a combination of the preference for x over y and the level of attention allocated to that joint realization, with $\phi(x, x) = 0$; and then (2) aggregates all these individual comparisons according to the joint distribution π over possible realizations and the choice distribution $(\alpha, 1 - \alpha)$, yielding

$$\max_{\alpha} \alpha * \sum_{x_A, y_B} \pi(x_A, y_B) \phi(x_A, y_B) + (1 - \alpha) * \sum_{x_B, y_A} \pi(x_B, y_A) \phi(x_B, y_A) \quad (\text{A.1})$$

where $\pi(\cdot, \cdot)$ denotes the probability of possible joint outcomes. However, because the outcome evaluation $\pi(\cdot, \cdot)$ is skew symmetric (i.e., $\pi(x, y) = -\pi(y, x)$) — the distinct feature characterizing correlation-sensitive preference (Lanzani, 2020), and the two options

¹By letting each state of the world denote one attribute, the pairwise normalization model (Landry and Webb, 2021) assigns different weights, $\frac{x}{x+y}$, where the joint outcomes $(x, y) \in \{(A : \$M, B : \$M), (A : \$M, B : \$0), (A : \$0, B : \$M), (A : \$0, B : \$0)\}$, to different states of the world based on the comparisons of outcomes between options in each state. It predicts identical behaviors since it doesn't assume convex preference – allowing mixture to be strictly preferred in some circumstances other than indifference between two options. Please refer to Appendix A.1 for details.

have identical support of outcomes, either receiving $\$M$ or $\$0$, the decision maker faces identical decision problems across all the scenarios. Intuitively, in this context, given the same absolute distance between outcomes, the decision maker assigns identical weights but with different signs. Thus, the evaluation of each option does not change with variations in the correlation between options, as the score on State ω_4 cancels out with one of the scores on States ω_1 , ω_2 , and ω_3 in Table 1.1. Therefore, in the Baseline Scenario, the decision maker faces the same decision problem:

$$\max_{\alpha} \alpha * \phi(\$M, \$0) + 50\% \phi(\$0, \$M) \quad (\text{A.2})$$

Thus, these models predict identical behaviors when moving from the decision problems with the CPS to those with the APS in the Baseline Scenario: either choosing Project A with $\alpha = 100\%$ whenever $\phi(\$M, \$0) > 0$, or mixing at a constant rate when $\phi(\$M, \$0) = 0$. Once the correlation between options is fixed, the decision maker's objective utility function does not change across choice frames.

Similarly, regardless of how the decision maker would perceive the distribution of the correlation between options, they face the same problem as in Equation (A.2) in the Unknown Scenario:

$$\begin{aligned} \max_{\alpha} & q * \{ \alpha * [50\% \phi(\$M, \$0) - 50\% \phi(\$0, \$M)] + 50\% \phi(\$0, \$M) \} \\ & + (1 - q) * \{ \alpha * [50\% \phi(\$M, \$0) - 50\% \phi(\$0, \$M)] + [75\% \phi(\$0, \$M) + 25\% \phi(\$M, \$0)] \} \\ & = \alpha * \phi(\$M, \$0) + 50\% \phi(\$0, \$M) \end{aligned} \quad (\text{A.3})$$

where q and $1 - q$ represent that the decision maker believes there is q percent chance that the two options are positively correlated and q chance that they are negatively cor-

related, respectively. In sum, these models predict that mixing behavior does not vary across problems in each scenario. Moreover, the magnitudes of mixing are identical across all three scenarios.

Landry and Webb (2021) propose a model of choice-set dependence in which the decision maker evaluates an option through a series of pairwise attribute comparisons. The value attached to each attribute comparison is normalized by the magnitude of the attributes under consideration. By letting each state of the world denote an attribute, the utility index on Option A from the choice set $\{A, B\}$ is

$$V(A, \{A, B\}) = \sum_{s_i \in S} \frac{x_{s_i}^A}{x_{s_i}^A + x_{s_i}^B}$$

where $x_{s_i}^A$ ($x_{s_i}^B$) denotes the outcome of Option A (Option B) in state s_i . Without additional assumption, the decision problem becomes:

$$\max_{\alpha} \alpha * V(A, \{A, B\}) + (1 - \alpha) * V(B, \{A, B\}) = \alpha * \sum_{s_i \in S} \frac{x_{s_i}^A}{x_{s_i}^A + x_{s_i}^B} + (1 - \alpha) * \sum_{s_i \in S} \frac{x_{s_i}^B}{x_{s_i}^A + x_{s_i}^B}$$

With the CPS as in Table 1.1, it turns out to be:

$$\max_{\alpha} \alpha * 3 + (1 - \alpha) = 2\alpha + 1 \tag{A.4}$$

which implies that it is optimal to choose Option A with 100%. Moreover, with the APS as in Table 1.2, it becomes:

$$\max_{\alpha} \alpha * (0.5 + 2) + (1 - \alpha) * 0.5 = 2\alpha + 0.5 \tag{A.5}$$

Which is maximized at $\alpha = 100\%$. More importantly, the maximization problem in

Equation (A.5) is a monotonic transformation of the problem in Equation (A.4). Thus, without additionally assuming convex preferences – mixing between options can be better in some circumstances, this model also predicts that mixing behavior is identical regardless of changes in the correlation between options.

A.1.2 Models of Correlation-Sensitive Stochastic Choices

For the models that predict varied mixing behavior depending on the correlation between options, I describe irrational diversification (Baltussen and Post, 2011; Rubinstein, 2002) and the evolutionary models proposed by Brennan and Lo (2012) in detail.

Irrational Diversification This theory assumes that the decision maker maximizes expected utility but incorrectly believes that she will be paid for all tickets, rather than one randomly-selected choice. With some concave utility function, i.e., risk aversion, the decision maker would mistakenly believe that mixing could allow them to hedge against the risk when facing the CPS. However, when facing the APS, such opportunity does not exist and thus, the decision maker tends to choose the dominant option with 100%.

To illustrate this, let’s revisit the prevailing example in Tables 1.1 and 1.2. Suppose the decision maker’s utility function is $u(x) = \sqrt{x}$. With the incorrect belief that they will get paid with all the tickets rather than a randomly selected one, the expected utility in the CPS as in Table 1.1 becomes:

$$\max_{\alpha} 75\% * \sqrt{\alpha * \$M} + 25\% * \sqrt{(1 - \alpha) * \$M}$$

To maximize expected utility, it is optimal to choose $\alpha = \frac{9}{10}$. On the contrary, when facing the APS as in Table 1.2, the decision problem becomes:

$$\max_{\alpha} 25\% * \sqrt{\$M} + 50\% \sqrt{\alpha * \$M}$$

To maximize expected utility, it is optimal to choose $\alpha = 100\%$. Therefore, this model predicts that mixing behavior varies with the correlation change between options: the decision maker is more likely to choose the dominant option with 100% when facing the APS, compared to the CPS.

Evolutionary Foundation The evolutionary model proposed by Brennan and Lo (2012) considers probability matching as an evolutionarily stable strategy. That is, the decision maker makes a binary decision between the two options, and then receive feedback from it. With the feedback received, the decision maker updates their belief on the joint distribution of outcomes between options and thus their choice distribution, i.e., α on Option A and $(1 - \alpha)$ on Option B. Thus, different correlations between options, which implies different joint distributions of outcomes between options, would lead to different choice distribution being updated. As a result, this model also predict that the mixing behavior is responsive to changes in the correlation between options.

A.2 First-Order Stochastic Dominance

I demonstrate some properties of first-order stochastic dominance, which are independent of the parameters I choose for the experimental design. First-order stochastic dominance (FOSD) is defined as the following: Option A *FOSD* Option B if $\forall x \in \mathbb{R}, Pr_B(X \leq x) \geq Pr_A(X \leq x)$, and $\exists x \in \mathbb{R}, Pr_B(X \leq x) > Pr_A(X \leq x)$. State-wise dominance (SWD) is a special case of FOSD, which is defined as the following: Option A state-wise dominates Option B, if $\forall s \in S, x_s^A \geq x_s^B$, and $\exists s \in S, x_s^A > x_s^B$, where s represents each possible state of the world (Quiggin, 1990). All the option pairs I use are FOSD pairs. The pairs with the APS also satisfy SWD. There is a particular relationship between FOSD and SWD. For any pair of FOSD options, there exists an option pair with identical marginal distributions satisfying not only FOSD but also SWD. By varying the correlation between options while keeping the marginal distributions the same, I can transfer any FOSD pair into an SWD pair.

Proposition 4. *For any FOSD option pair: Options A and B, the following statements hold:*

- a. *If A FOSD B, there exists a permuted partition ρ such that for each state of the world $j \in \{1, \dots, N\}$, $x_{Aj} \geq x_{B\rho(j)}$.*
- b. *If $\forall j \in \{1, \dots, N\}, \rho(j) = j$ and $x_{Aj} \geq x_{B\rho(j)}$, then Option A both FOSD and SWD Option B.*

Proof: [Proof of Proposition 1(a)]

For any FOSD option pair: Option A and B, I subdivide the marginal distributions of two options into a set of N equiprobable partitions where N is the least common multiple of the actual states of the world used to define the two options. Each of N states occurs with probability $1/N$. Then, I rank the payoff associated with each of the

N states from the lowest to the highest, i.e., $x_{A1} \leq x_{A2} \cdots \leq x_{AN}$. Next define ρ such that $x_{B\rho(1)} \leq x_{B\rho(2)} \cdots \leq x_{B\rho(N)}$.

Now for any j , first order stochastic dominance implies

$$Pr\{x_B \leq x_{Aj}\} \geq Pr\{x_A \leq x_{Aj}\} = \frac{j}{N}$$

By the ordering of $x_{B\rho(j)}$, this implies

$$x_{Aj} \geq x_{B\rho(j)}$$

Proof: [Proof of Proposition 1(b)]




By the definition of state-wise dominance, A is state-wise dominant B . Therefore, A FOSD B .

The proposition indicates that every FOSD pair can be partitioned into the Alternative Frame. If all the partitions coincide with the actual states of the world, the FOSD pair will also be an SWD one, i.e., the APS. I can transfer any FOSD pair into an SWD pair by varying the correlation between options without changing the marginal distributions. These properties guide my choices of experimental design.

A.3 Parameters used in Experiments













Table A.1 and Table A.2 demonstrate the correlations and interfaces used in the tasks under Category ($A : 67\%; B : 33\%$) and Category ($A : 80\%; B : 20\%$) in the Baseline treatment separately:

Table A.1: Baseline: Correlations and Interfaces in Tasks under Category ($A : 67\%; B : 33\%$)

Task	$CORR(A, B)$	Interface
1	0.5	Front:  Back: 
2	0	Front:  Back: 
3	-0.5	Front:  Back: 
4	-1	Front:  Back: 

Note: Each number denotes a two-sided coin. Task 1 and Task 4 correspond to the APS and CPS, respectively.

Table A.2: Baseline: Correlations and Interfaces under Category ($A : 80\%; B : 20\%$)

Task	$CORR(A, B)$	Interface
1	0.25	Front:  Back: 
2	0	Front:  Back: 
3	-0.25	Front:  Back: 
4	-0.5	Front:  Back: 
5	-0.75	Front:  Back: 
6	-1	Front:  Back: 

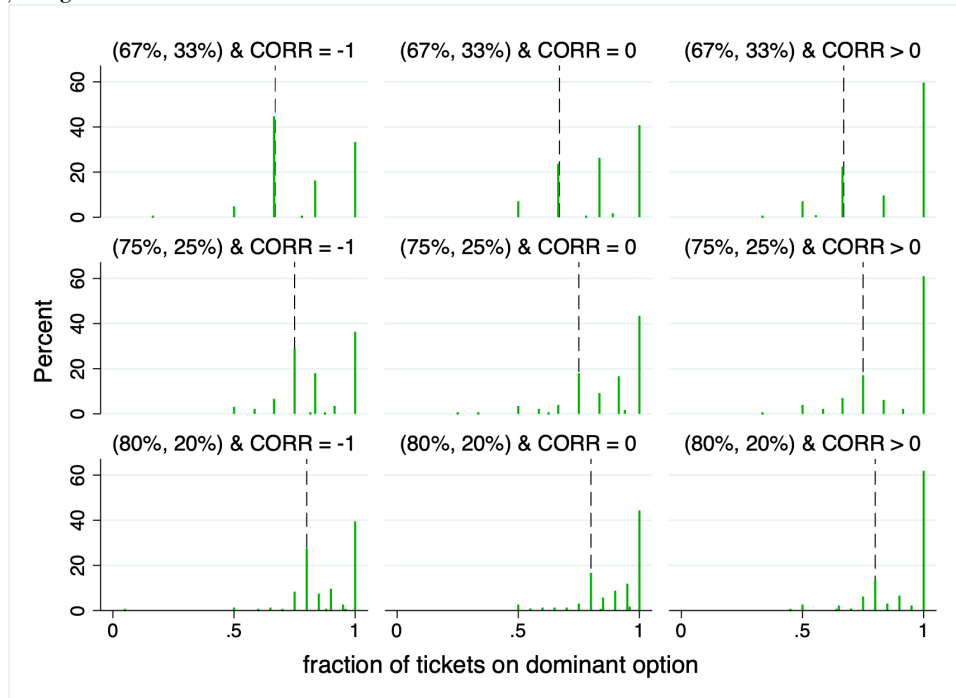
Note: Each number denotes a two-sided coin. Task 1 and Task 6 correspond to the APS and CPS, respectively.

Tasks in the Independence treatment share the same interfaces as those in the Baseline but the correlation between options is fixed at zero.

A.4 The Magnitude of Mixing

In Figure A.1, I show histograms of the fraction of tickets allocated on the dominant options by subjects in the tasks where the two options are perfectly negatively correlated, zero correlated, and positively correlated across the three probability categories in the Baseline treatment. The black dash line denotes the fraction of tickets on dominant options that matches exactly with the probability of occurrence.

Figure A.1: Baseline: Histograms of frequency of choices in the tasks with CORR = -1, 0, or ≥ 0 .



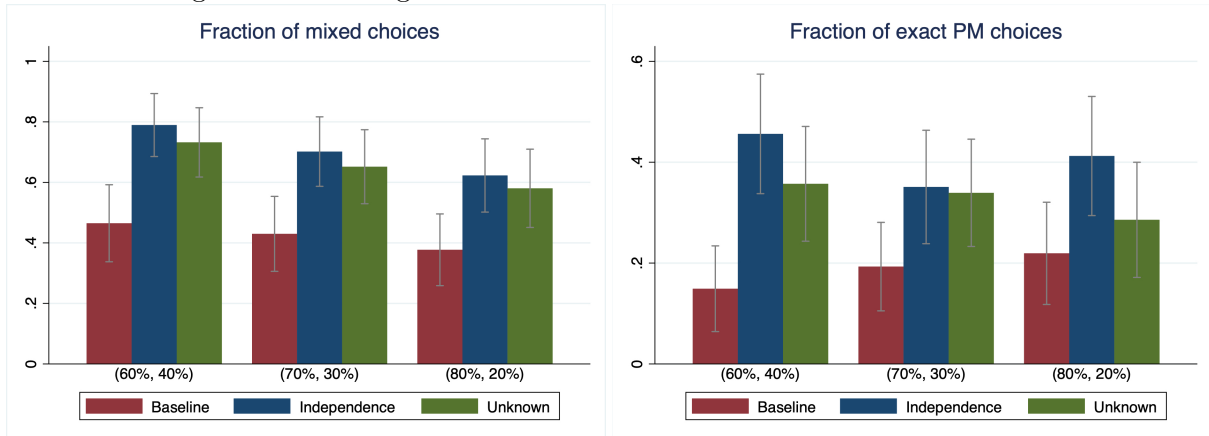
Notes: The black dash line denotes the fraction of tickets on dominant options that matches exactly with the probability of occurrence. The error bars depict 95% confidence intervals.

A.5 Block 3: Transfer of Learning

Design. I use Block 3 to explore the extent to which subjects learned in previous blocks can be transferred to a new setting, in which I use the same unified framework as the Unknown treatment. In Block 3, subjects face six distinct ticket-allocation tasks: ($A : 80\%$; $B : 20\%$), ($A : 20\%$; $B : 80\%$), ($A : 70\%$; $B : 30\%$), ($A : 30\%$; $B : 70\%$), ($A : 60\%$; $B : 40\%$), and ($A : 40\%$; $B : 60\%$). In each task, subjects can see a roll of ten two-sided coins, and the computer randomly selects a coin. Subjects need to predict the color on the randomly drawn coin by allocating ten tickets. All the other implementations and protocols are identical to the tasks in Blocks 1 and 2.

Results. I compare Block 3 across three treatments to study the extent to which learning in one environment is transferable to another. Figure A.2 plots the *likelihood of mixing* and *likelihood of exact PM* across the three treatments and separated by probability categories. In Table A.3, I regress the dependent variables of interest on the dummies of treatments, Baseline VS Independence and Independence VS Unknown, respectively. As shown in Figure A.2 and Table A.3, subjects are 28.3% (OLS, $p = 0.000$) less likely to make *mixed* choices and 22.4% (OLS, $p = 0.000$) less likely to make *exact PM* choices in the Baseline than in the Independence. However, the difference in Block 3 between the Independence and Unknown is not significant. These results emphasize the importance of correlation in decision-making from the prospect of learning: as subjects perceive the correlation in the Independence and Unknown treatments with frictions, the learning process is slow and limited.

Figure A.2: Mixing Behavior in Block 3 across the Three Treatments



Note: I plot the average mixing behavior in Block 3 across the three treatments and separated by probability categories. The error bars depict 95% confidence intervals.

Table A.3: Mixing Behavior in Block 3 across the Three Treatments

	(1)	(2)	(3)	(4)
	<i>mixed</i>	<i>exact PM</i>	<i>mixed</i>	<i>exact PM</i>
IvsB(Baseline = 1)	-0.283*** (0.0756)	-0.224*** (0.0675)		
IvsU (Independence = 1)			-0.000489 (0.0718)	-0.0465 (0.0706)
Constant	0.293 (0.242)	0.147 (0.113)	0.407 (0.287)	-0.0388 (0.116)
Observations	684	684	678	678

Note: Results from OLS regression. The dependent variable takes the value of 1 if the allocation choice in a task is classified as (1) *mixed* choice, or (2) *exact PM*, separately. IvsB is the dummy variable on whether the task is in the Baseline VS Independence. It takes the value of 1 if the task is in the Baseline. IvsU is the dummy variable on whether the task is in the Unknown VS Independence. It takes the value of 1 if the task is in the Independence. The regression also includes probability categories, dominant color, gender, school year and STEM as controls. Standard errors are clustered at the subject level and listed in parentheses. Full regression results can be found in Appendix A.6.1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.6 Additional Results

A.6.1 Full Regression Results

Table A.4, Table A.5, Table A.6, Table A.7, Table A.8, and Table A.9 illustrate the full regression results summarized on Table 1.14, Table 1.13, Table 1.11, Table 1.12, Table 1.15 , and Table A.3 of the main manuscript, respectively.

Table A.4: Baseline: Impacts of Correlation/Frame on Mixing Behavior

	Block 1&2 <i>mixed</i>	Block 1 <i>mixed</i>	Block 2 <i>mixed</i>	Block 1&2 <i>exact PM</i>	Block 1 <i>exact PM</i>	Block 2 <i>exact PM</i>
Correlation/Frame	-0.165*** (0.0346)	-0.108*** (0.0318)	-0.221*** (0.0420)	-0.108*** (0.0270)	-0.0843*** (0.0261)	-0.133*** (0.0369)
(75%, 25%)	-0.0168* (0.00846)	-0.0248* (0.0134)	-0.00880 (0.0103)	-0.0915*** (0.0233)	-0.109*** (0.0273)	-0.0738** (0.0294)
(80%, 20%)	-0.0272 (0.0185)	-0.0238 (0.0195)	-0.0306 (0.0225)	-0.128*** (0.0246)	-0.152*** (0.0321)	-0.105*** (0.0254)
Blue	-0.0152** (0.00729)	-0.0140 (0.0106)	-0.0164** (0.00735)	-0.0164 (0.0118)	-0.0304* (0.0164)	-0.00234 (0.0139)
STEM	0.0193 (0.108)	0.0560 (0.111)	-0.0174 (0.108)	-0.0229 (0.0804)	-0.0381 (0.0799)	-0.00758 (0.0860)
Female	0.341 (0.221)	0.339 (0.206)	0.343 (0.238)	0.112 (0.0763)	0.150** (0.0733)	0.0734 (0.0850)
Male	0.0693 (0.242)	0.0718 (0.231)	0.0668 (0.257)	0.0722 (0.108)	0.105 (0.0949)	0.0394 (0.128)
Sophomore	0.451** (0.214)	0.421* (0.224)	0.480** (0.216)	0.135 (0.114)	0.188 (0.125)	0.0831 (0.115)
Junior	0.276 (0.206)	0.279 (0.217)	0.274 (0.211)	0.174 (0.110)	0.203** (0.101)	0.145 (0.133)
Senior	0.269 (0.191)	0.250 (0.202)	0.288 (0.194)	0.174* (0.0913)	0.203** (0.0801)	0.145 (0.112)
Graduate	0.189 (0.213)	0.163 (0.224)	0.216 (0.216)	0.0520 (0.0863)	0.0714 (0.0811)	0.0326 (0.102)
Constant	0.0266 (0.276)	0.0753 (0.272)	-0.0221 (0.293)	0.0562 (0.0908)	0.0470 (0.0757)	0.0654 (0.117)
Observations	3420	1710	1710	3420	1710	1710

Note: Results from OLS regression. The dependent variables take the value of one if the allocation choice in a task is classified as *mixed* or *exact PM*, respectively. Correlation captures the correlation parameters, which takes values of -1, -0.5, 0, 0.5 under Category (67%, 33%); -1, -0.67, -0.33, 0, 0.33 under Category (75%, 25%); and -1, -0.75, -0.5, -0.25, 0, 0.25 under Category (80%, 20%). Each regression also includes categorical variables of probability categories, gender, and school year, as well as indicator variables of dominant color and STEM, as controls. (75%, 25%) and (80%, 20%) are the dummy variables which take the value of one if the task is in Category (75%, 25%) and if the task is in Category (80%, 20%), respectively. Blue takes the value of 1 if the dominant color is blue. STEM takes the value of one if the subject has a STEM major. Female takes the value of one if the subject's gender is female (the benchmark is "other"). Male takes the value of one if the subject's gender is male. Sophomore, Junior, Senior, and Graduate takes the value of one if the subject's school year falls into one of the categories. Standard errors are clustered at the subject level and listed in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: Baseline: Impacts of Different Marginal Correlation Changes on Mixing Behavior

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>mixed</i>	<i>exact PM</i>	<i>mixed</i>	<i>exact PM</i>	<i>mixed</i>	<i>exact PM</i>
Marginal Correlation Change at $CORR = -1$	-0.00702 (0.0352)	-0.347*** (0.0795)				
Marginal Correlation Change at $CORR = 0$			-0.126** (0.0599)	-0.0316 (0.0616)		
Marginal Correlation Change at $CORR \geq 0$					-0.304*** (0.0644)	-0.0376 (0.0312)
(75%, 25%)	-0.0215 (0.0184)	-0.150*** (0.0334)	-0.0211 (0.0182)	-0.0645* (0.0326)	-0.0181 (0.0138)	-0.0610** (0.0242)
(80%, 20%)	-0.0338 (0.0224)	-0.162*** (0.0333)	-0.0193 (0.0231)	-0.0925*** (0.0236)	-0.0307 (0.0192)	-0.0936*** (0.0213)
Blue	-0.0246** (0.00983)	-0.0158 (0.0170)	-0.0193* (0.0105)	-0.0158 (0.0179)	-0.0143 (0.0105)	-0.00877 (0.0151)
STEM	-0.0236 (0.119)	-0.0230 (0.0887)	0.0194 (0.113)	-0.0255 (0.0847)	0.0592 (0.106)	-0.00360 (0.0843)
Female	0.286 (0.257)	0.0389 (0.152)	0.350 (0.265)	0.121* (0.0640)	0.375** (0.184)	0.138** (0.0596)
Male	0.0334 (0.281)	-0.0344 (0.168)	0.0513 (0.285)	0.0903 (0.106)	0.0813 (0.210)	0.0997 (0.105)
Sophomore	0.527** (0.225)	0.261* (0.139)	0.452* (0.233)	0.117 (0.105)	0.363* (0.216)	0.0831 (0.0921)
Junior	0.421* (0.217)	0.298** (0.123)	0.234 (0.216)	0.160 (0.112)	0.130 (0.216)	0.143 (0.111)
Senior	0.364* (0.195)	0.268*** (0.0851)	0.268 (0.202)	0.184* (0.0956)	0.175 (0.199)	0.160* (0.0891)
Graduate	0.205 (0.217)	0.0846 (0.0917)	0.189 (0.228)	0.0509 (0.0843)	0.145 (0.221)	0.0486 (0.0804)
Constant	0.133 (0.308)	-0.122 (0.154)	0.0870 (0.322)	0.0243 (0.0840)	0.0788 (0.257)	0.0122 (0.0730)
Observations	1140	1140	1140	1140	1824	1824

Note: Results from OLS regression. The dependent variables take the value of 1 if the allocation choice in a task is classified as *mixed* choice, or as *exact PM* choice. The variable of marginal correlation change at $CORR = -1$ takes the value of 1 if the correlation parameter is -0.5 under Category (67%, 33%), -0.67 under Category (75%, 25%), or -0.75 under Category (80%, 20%), and takes the value of 0 if $CORR = -1$. The variable of marginal correlation change at $CORR = 0$ takes the value of 1 if $CORR = 0$, and takes the value of 0 if the correlation parameter is -0.5 under Category (67%, 33%), -0.33 under Category (75%, 25%), or -0.25 under Category (80%, 20%). The variable of marginal correlation change at $CORR \geq 0$ takes the value of 1 if $CORR \geq 0$ which includes the correlation parameter that is 0 under all categories, 0.5 under Category (67%, 33%), 0.33 under Category (75%, 25%), or 0.25 under Category (80%, 20%), and takes the value of 0 if the correlation parameter is -0.5 under Category (67%, 33%), -0.33 under Category (75%, 25%), or -0.25 under Category (80%, 20%). The regression also includes probability categories, dominant color, gender, school year and STEM as controls. (75%, 25%) and (80%, 20%) are the dummy variables which take the value of one if the task is in Category (75%, 25%) and if the task is in Category (80%, 20%), respectively. Blue takes the value of 1 if the dominant color is blue. STEM takes the value of one if the subject has a STEM major. Female takes the value of one if the subject's gender is female (the benchmark is "other"). Male takes the value of one if the subject's gender is male. Sophomore, Junior, Senior, and Graduate takes the value of one if the subject's school year falls into one of the categories. Standard errors are clustered at the subject level and listed in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6: Baseline VS Independence: Impacts of Correlation/Frame Change on Mixing Behavior

	Block 1&2 <i>mixed</i>	Block 1 <i>mixed</i>	Block 2 <i>mixed</i>	Block 1&2 <i>exact PM</i>	Block 1 <i>exact PM</i>	Block 2 <i>exact PM</i>
IvsB X CORR/Frame	-0.158*** (0.0366)	-0.0977*** (0.0334)	-0.218*** (0.0452)	-0.125*** (0.0286)	-0.108*** (0.0311)	-0.142*** (0.0380)
CORR/Frame	-0.00586 (0.0114)	-0.00990 (0.0106)	-0.00182 (0.0161)	0.0121 (0.0114)	0.0199 (0.0183)	0.00439 (0.0128)
IvsB(Baseline=1)	-0.175*** (0.0661)	-0.144** (0.0663)	-0.205*** (0.0721)	-0.0877* (0.0518)	-0.0564 (0.0515)	-0.119** (0.0565)
(75%, 25%)	-0.00268 (0.0122)	-0.0132 (0.0136)	0.00787 (0.0141)	-0.104*** (0.0188)	-0.125*** (0.0214)	-0.0827*** (0.0248)
(80%, 20%)	-0.0150 (0.0153)	-0.0176 (0.0158)	-0.0124 (0.0183)	-0.176*** (0.0221)	-0.193*** (0.0245)	-0.160*** (0.0260)
Blue	-0.0123** (0.00533)	-0.0140* (0.00778)	-0.0105** (0.00496)	-0.0111 (0.00791)	-0.0211* (0.0110)	-0.00117 (0.00988)
STEM	-0.126* (0.0689)	-0.115* (0.0690)	-0.138* (0.0734)	-0.0558 (0.0564)	-0.0689 (0.0554)	-0.0427 (0.0605)
Female	0.461*** (0.152)	0.519*** (0.131)	0.403** (0.179)	0.168*** (0.0486)	0.207*** (0.0428)	0.129** (0.0619)
Male	0.210 (0.164)	0.252* (0.144)	0.167 (0.191)	0.120* (0.0703)	0.133** (0.0620)	0.108 (0.0855)
Sophomore	0.231 (0.174)	0.177 (0.177)	0.284 (0.188)	0.116 (0.0804)	0.130 (0.0854)	0.102 (0.0882)
Junior	0.226 (0.168)	0.171 (0.172)	0.280 (0.180)	0.168** (0.0783)	0.162** (0.0795)	0.175* (0.0895)
Senior	0.155 (0.160)	0.101 (0.165)	0.210 (0.173)	0.109* (0.0652)	0.0947 (0.0650)	0.123 (0.0781)
Graduate	0.139 (0.169)	0.0774 (0.172)	0.200 (0.188)	0.0292 (0.0622)	0.0257 (0.0638)	0.0327 (0.0772)
Constant	0.241 (0.224)	0.261 (0.214)	0.221 (0.252)	0.165* (0.0868)	0.159* (0.0839)	0.170 (0.104)
Observations	6840	3420	3420	6840	3420	3420

Note: Results from OLS regression. The dependent variable takes the value of 1 if the allocation choice in a task is classified as *mixed*, or as *exact PM*, respectively. CORR/Frame is the variable of correlation parameters in the Baseline and Independence treatments. In the Baseline, it captures either the correlation change or the associated frame change. In the Independence, it denotes the frame change only. IvsB is the dummy variable on whether the task comes from the Baseline or Independence. It takes the value of 1 if the task is in the Baseline. The regression also includes the categorical variables of probability categories, gender, and school year, as well as the indicator variables of dominant color, and STEM, as controls. (75%, 25%) and (80%, 20%) are the dummy variables which take the value of one if the task is in Category (75%, 25%) and if the task is in Category (80%, 20%), respectively. Blue takes the value of 1 if the dominant color is blue. STEM takes the value of one if the subject has a STEM major. Female takes the value of one if the subject's gender is female (the benchmark is "other"). Male takes the value of one if the subject's gender is male. Sophomore, Junior, Senior, and Graduate takes the value of one if the subject's school year falls into one of the categories. Standard errors are clustered at the subject level and listed in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Independence VS Unknown: Impacts of Frame Change on Mixing Behavior

	Block 1&2 <i>mixed</i>	Block 1 <i>mixed</i>	Block 2 <i>mixed</i>	Block 1&2 <i>exact PM</i>	Block 1 <i>exact PM</i>	Block 2 <i>exact PM</i>
IvsU X Frame/Ex post <i>CORR</i>	0.00387 (0.0136)	0.00964 (0.0148)	-0.00190 (0.0192)	-0.0318** (0.0157)	-0.0648*** (0.0244)	0.00122 (0.0190)
Frame/Ex post <i>CORR</i>	-0.00509 (0.0114)	-0.00947 (0.0106)	-0.000710 (0.0161)	0.00962 (0.0114)	0.0177 (0.0183)	0.00159 (0.0129)
IvsU(Independence=1)	0.0469 (0.0700)	0.0664 (0.0710)	0.0274 (0.0763)	0.0566 (0.0567)	0.0920 (0.0572)	0.0211 (0.0613)
(75%, 25%)	0.00759 (0.0153)	0.00315 (0.0155)	0.0120 (0.0183)	-0.117*** (0.0231)	-0.128*** (0.0247)	-0.107*** (0.0292)
(80%, 20%)	-0.00648 (0.0177)	-0.0139 (0.0168)	0.000899 (0.0217)	-0.206*** (0.0259)	-0.220*** (0.0259)	-0.192*** (0.0313)
Blue	-0.00413 (0.00450)	-0.00590 (0.00675)	-0.00236 (0.00531)	-0.00855 (0.00735)	-0.0165 (0.0110)	-0.000590 (0.00952)
STEM	-0.115 (0.0745)	-0.106 (0.0738)	-0.123 (0.0815)	-0.0125 (0.0588)	-0.0165 (0.0592)	-0.00853 (0.0620)
Female	0.333 (0.231)	0.409* (0.218)	0.258 (0.251)	0.143 (0.114)	0.159 (0.127)	0.127 (0.108)
Male	0.0613 (0.239)	0.149 (0.227)	-0.0261 (0.259)	0.0432 (0.125)	0.0281 (0.137)	0.0582 (0.121)
Sophomore	0.0387 (0.183)	0.0478 (0.179)	0.0297 (0.202)	0.148* (0.0838)	0.137 (0.101)	0.158* (0.0812)
Junior	0.200 (0.181)	0.168 (0.176)	0.233 (0.198)	0.257*** (0.0873)	0.222** (0.0962)	0.293*** (0.0856)
Senior	0.0520 (0.175)	0.0542 (0.171)	0.0499 (0.192)	0.106 (0.0664)	0.0663 (0.0797)	0.145** (0.0603)
Graduate	0.181 (0.183)	0.164 (0.177)	0.197 (0.206)	0.170* (0.0929)	0.141 (0.0992)	0.200** (0.0953)
Constant	0.424 (0.289)	0.376 (0.276)	0.472 (0.315)	0.144 (0.144)	0.175 (0.162)	0.112 (0.136)
Observations	6780	3390	3390	6780	3390	3390

Note: Results from OLS regression. The dependent variable takes the value of 1 if the allocation choice in a task is classified as (1) *mixed* choice, (2) *exact PM*. Frame/Ex post *CORR* is the variable of correlation parameters in the Independence and Unknown treatments. In the Independence, it captures the frame change only. In the Unknown, it denotes no impact by design. IvsU is the dummy variable on whether the task is in the Unknown VS Independence. It takes the value of 1 if the task is in the Independence. The regression also includes probability categories, dominant color, gender, school year and STEM as controls. (75%, 25%) and (80%, 20%) are the dummy variables which take the value of one if the task is in Category (75%, 25%) and if the task is in Category (80%, 20%), respectively. Blue takes the value of 1 if the dominant color is blue. STEM takes the value of one if the subject has a STEM major. Female takes the value of one if the subject's gender is female (the benchmark is "other"). Male takes the value of one if the subject's gender is male. Sophomore, Junior, Senior, and Graduate takes the value of one if the subject's school year falls into one of the categories. Standard errors are clustered at the subject level and listed in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Three treatments: Mixing Behavior in Zero-correlation Tasks (Task 2)

	(1)	(2)	(3)	(4)
	<i>mixed</i>	<i>exact PM</i>	<i>mixed</i>	<i>exact PM</i>
IvsB(Baseline = 1)	-0.141** (0.0676)	-0.0990* (0.0541)		
IvsU(Independence = 1)			0.0292 (0.0683)	0.0607 (0.0567)
(75%, 25%)	-0.0110 (0.0170)	-0.101*** (0.0267)	-0.00885 (0.0194)	-0.126*** (0.0316)
(80%, 20%)	-0.0285 (0.0207)	-0.149*** (0.0279)	-0.0288 (0.0211)	-0.215*** (0.0326)
Blue	-0.0219* (0.0128)	-0.00146 (0.0179)	-0.0103 (0.0121)	-0.0177 (0.0167)
STEM	-0.110 (0.0694)	-0.0560 (0.0580)	-0.102 (0.0713)	-0.0165 (0.0562)
Female	0.472*** (0.164)	0.178*** (0.0520)	0.359 (0.232)	0.140 (0.149)
Male	0.202 (0.175)	0.133* (0.0728)	0.0637 (0.240)	0.0292 (0.159)
Sophomore	0.162 (0.181)	0.0386 (0.0847)	-0.00378 (0.175)	0.114 (0.0871)
Junior	0.116 (0.172)	0.0902 (0.0912)	0.167 (0.173)	0.238** (0.0926)
Senior	0.0799 (0.165)	0.0528 (0.0796)	0.0165 (0.168)	0.0753 (0.0737)
Graduate	0.0732 (0.174)	-0.0187 (0.0779)	0.136 (0.175)	0.184* (0.0972)
Constant	0.331 (0.237)	0.208** (0.101)	0.466 (0.286)	0.189 (0.172)
Observations	1368	1368	1356	1356

Note: Results from OLS regression with observations in Task 2 and pooling Blocks 1 and 2 together. The dependent variable takes the value of 1 if the allocation choice in a task is classified as (1) Mixed choice, or as (2) Exact PM separately. IvsB is the dummy variable on whether the task is in the Baseline VS Independence. It takes the value of 1 if the task is in the Baseline. IvsU is the dummy variable on whether the task is in the Unknown VS Independence. It takes the value of 1 if the task is in the Independence. The regression also includes probability categories, dominant color, gender, school year and STEM as controls. (75%, 25%) and (80%, 20%) are the dummy variables which take the value of one if the task is in Category (75%, 25%) and if the task is in Category (80%, 20%), respectively. Blue takes the value of 1 if the dominant color is blue. STEM takes the value of one if the subject has a STEM major. Female takes the value of one if the subject's gender is female (the benchmark is "other") and male takes the value of one if the subject's gender is male. Sophomore, Junior, Senior, and Graduate takes the value of one if the subject's school year falls into one of the categories. Standard errors are clustered at the subject level and listed in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Mixing Behavior in Block 3 across the Three Treatments

	(1)	(2)	(3)	(4)
	Mix	Exact PM	Mix	Exact PM
IvsB(Baseline = 1)	-0.283*** (0.0756)	-0.224*** (0.0675)		
IvsU (Independence = 1)			-0.000489 (0.0718)	-0.0465 (0.0706)
(70%, 30%)	-0.0614*** (0.0228)	-0.0307 (0.0253)	-0.0841*** (0.0236)	-0.0619* (0.0324)
(80%, 20%)	-0.127*** (0.0287)	0.0132 (0.0347)	-0.159*** (0.0319)	-0.0575 (0.0400)
Blue	0.0351** (0.0159)	0.0146 (0.0184)	0.0383*** (0.0132)	0.0206 (0.0229)
STEM	-0.0845 (0.0770)	-0.0755 (0.0691)	-0.111 (0.0717)	-0.0946 (0.0735)
Female	0.477** (0.196)	0.293*** (0.0655)	0.511** (0.243)	0.388*** (0.0640)
Male	0.267 (0.202)	0.224** (0.0909)	0.169 (0.254)	0.214** (0.0825)
Sophomore	0.173 (0.158)	0.154 (0.112)	-0.0656 (0.178)	0.246** (0.113)
Junior	0.0714 (0.148)	-0.0231 (0.0850)	0.149 (0.167)	0.293*** (0.0989)
Senior	0.0996 (0.140)	0.0536 (0.0906)	-0.0217 (0.167)	0.159* (0.0885)
Graduate	0.170 (0.159)	0.0301 (0.114)	0.123 (0.177)	0.255* (0.132)
Constant	0.293 (0.242)	0.147 (0.113)	0.407 (0.287)	-0.0388 (0.116)
Observations	684	684	678	678

Note: Results from OLS regression. The dependent variable takes the value of 1 if the allocation choice in a task is classified as (1) *mixed* choice, (2) *exact PM*. IvsB is the dummy variable on whether the task is in the Baseline VS Independence. It takes the value of 1 if the task is in the Baseline. IvsU is the dummy variable on whether the task is in the Unknown VS Independence. It takes the value of 1 if the task is in the Independence. The regression also includes probability categories, dominant color, gender, school year and STEM as controls. (75%, 25%) and (80%, 20%) are the dummy variables which take the value of one if the task is in Category (75%, 25%) and if the task is in Category (80%, 20%), respectively. Blue takes the value of 1 if the dominant color is blue. STEM takes the value of one if the subject has a STEM major. Female takes the value of one if the subject's gender is female (the benchmark is "other"). Male takes the value of one if the subject's gender is male. Sophomore, Junior, Senior, and Graduate takes the value of one if the subject's school year falls into one of the categories. Standard errors are clustered at the subject level and listed in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.6.2 Prevalence of Mixing

I classify subjects into three mutually exclusive types based on their choices: *Never Mix*, *Sometimes Mix*, and *Always Mix*. The *Never Mix* type refers to the subject who always allocates all the tickets on the dominant option in each task; the *Always Mix* type refers to the subject who always allocates at least one ticket on the dominated option in each task; the *Sometimes Mix* type refers to the subject who is in between.

Table A.10: Mixing Types across the Three Treatments

	Baseline		Independence		Unknown
<i>Always Mix</i>	17.54%	***	42.11%	**	44.64%
<i>Sometimes Mix</i>	64.91%	***	45.61%		46.43%
<i>Never Mix</i>	17.54%	***	12.28%	***	8.93%
<i>N</i> subjects	57		57		56

Note: I classify subjects into three mutually exclusive types based on their 66 choices in each treatment: *Never Mix*, *Sometimes Mix*, and *Always Mix*. The *Never Mix* type refers to the subject who always allocates all the tickets on the dominant option in each task; the *Always Mix* type refers to the subject who always allocates at least one ticket on the dominated option in each task; the *Sometimes Mix* type refers to the subject who is in between. I use the two-way t-test to test whether the difference is significant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As shown in Table A.10 where I classify subjects based on all the 66 choices, types in the Baseline are distributed differently from those in the Independence and Unknown treatments. In the Baseline, it is most prominent for subjects to *Sometimes Mix* (64.91%), which is significantly higher than those in the Independence and Unknown treatments (two-way t-test, $p = 0.000$). However in the Independence and Unknown treatments, the two most prominent types in the population are subjects who *Always Mix* (Independence: 42.11%; Unknown: 44.64%) and those who *Sometimes Mix* (Independence: 45.61%; Unknown: 46.43%). The fraction of *Always Mix* type in the Baseline (17.54%) is significantly lower than in the other two treatments (two-way t-test, $p = 0.000$). Moreover, the fraction of *Never Mix* type in the Baseline (17.54%) is significantly higher than in the other two treatments: (Independence: 12.28%; Unknown:

8.93%; two-way t-test, $p = 0.000$).

I also calculate the distributions of three types by blocks and find similar results. I classify subjects into the three types based on the 30 choices in Block 1, the 30 choices in Block 2, and the six choices in Block 3 separately, and separated by the three treatments. Conditional on Blocks 1 and 2, *Sometimes Mix* is the prominent type in the Baseline, while *Always Mix* is the dominant type in the Independence and Unknown. In each block, the fraction of *Always Mix* type is significantly lower in the Baseline than in the other two treatments (two-way t-test, $p = 0.000$). The type distribution in the Independence is similar to that in the Unknown. As illustrated in Table A.11, compared to Block 1, the fraction of *Never Mix* type increases when subjects gain some experience in Block 2.

Table A.11: Mixing Types across the Three Treatments: by Blocks

	Baseline		Independence		Unknown
Block 1					
<i>Always Mix</i>	24.56%	***	52.63%		53.57%
<i>Sometimes Mix</i>	56.14%	***	29.82%	***	35.71%
<i>Never Mix</i>	19.30%		17.54%	***	10.71%
Block 2					
<i>Always Mix</i>	21.05%	***	49.12%		51.79%
<i>Sometimes Mix</i>	50.88%	***	28.07%		26.79%
<i>Never Mix</i>	28.07%	***	22.81%		21.43%
Block 3					
<i>Always Mix</i>	31.58%	***	56.14%		57.14%
<i>Sometimes Mix</i>	19.30%	**	26.32%	***	17.86%
<i>Never Mix</i>	49.12%	***	17.54%	**	25.00%

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: I classify subjects into three types based on the thirty choices in Block 1, the thirty choices in Block 2, and the six choices in Block 3 separately, and separated by the three treatments. The *Never Mix* type refers to the subject who always allocates all the tickets on the dominant option in each task; the *Always Mix* type refers to the subject who always allocates at least one ticket on the dominated option in each task; the *Sometimes Mix* type refers to the subject who is in between. I use the two-way t-test to test whether the difference is significant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.6.3 Suggestive Strategies

In the exit survey, subjects are asked to choose one of the following statements as suggestive strategies:

- **Never Mix:** “Always allocate all your tickets on the color with larger likelihood. That is, if blue is more likely to happen, allocate all tickets on blue; otherwise, all on orange.”
- **Mix Evenly:** “Always mix between the 2 colors evenly.”
- **Exact PM:** “Always mix between the 2 colors in a similar way as to how they distribute. For example, if 6 blue and 3 orange, allocate 4 tickets on blue and 2 tickets on orange.”
- **Always Mix:** “Mixing is always better than allocating all tickets on 1 color.”
- **Sometimes Mix:** “It depends. For some questions, allocating all tickets on 1 color is better than mixing. For others, mixing is better.”

Table A.12 illustrates the fractions of subjects who chose each of the five strategies. Although it is not incentivized, the distribution of suggestive strategies are consistent with the aggregate behavior in the experiment: majority of subjects in the Baseline choose **Never Mix**, while the prominent strategy chose in the Independence and Unknown is **Exact PM**.

Table A.12: Distribution of Suggestive Strategies across the Three Treatments

	Baseline	Independence	Unknown
Never Mix	43.86%	29.82%	33.93%
Mix Evenly	3.51%	0%	1.79%
Exact PM	21.05%	35.09%	37.50%
Always Mix	1.75%	10.53%	8.93%
Sometimes Mix	29.82%	24.56%	17.86%

Note: Each cell represents the fraction of subjects in each treatment who chose the suggestive strategy. The fraction in bold denotes the most prominent strategy selected in each treatment.

A.7 Experimental Instructions

Following pages are the instructions subjects can see in the experiment.

Overview

- Welcome to our experiment on decision making. Thank you for participating!
- This experiment consists of 3 “blocks”. Block 2 will be explained once you complete Block 1. Similarly, Block 3 will be explained once you complete Block 2. I start by providing you with instructions. I will ask you questions to make sure that you understand the rules. You should be able to answer all these questions correctly.
- Please follow the instructions closely and carefully. You will not be allowed to start the study until answering each question correctly.
- At the end of the experiment, one of the decisions will be randomly selected as the **Decision-that-counts** for payment. Since all decisions are equally likely to be chosen, you should approach each decision as if it is the **Decision-that-counts**.
- In addition to being paid for one decision, you will also receive \$5 as participation payment.

Important Information

- You should think about each question **independently** of all other questions in this study. There is no point in strategizing across questions.
- You will note that I sometimes ask you similar-sounding questions. These questions might have similar answers, or very different ones. Please consider each individual question **carefully**.

Block 1 (*Baseline*)

This block consists of 30 rounds. In each round, you face a situation like the one described below.

You will see a roll of 2-sided coins, some of which have colors on the front and back sides. There are 2 types of colors, **blue** and **orange**. Colors are distributed among the front and back sides of coins based on the rules below:

- The **front side** of each coin either is colored with **blue** or has no color.
- The **back side** of each coin either is colored with **orange** or has no color.
- **Different colors are NOT necessarily exclusive.** That is, it is possible that 1 coin is colored with **blue** on the front side and **orange** on the back side.

The computer will randomly draw 1 coin and check the color on each side of the coin. You do **NOT** know which coin is drawn. Figure A.3 is an example of the environment you might face:

Figure A.3: Screenshot of Interface in Baseline

- You will see a roll of 9 2-sided coins, labeled 1-9.
- 6 out of 9 coins are colored **blue** on the **front sides**, while the rest of 3 coins are not. That is, 67% of coins contain **blue sides** and 33% of coins do not. 3 out of 9 coins are colored **orange** on the **back sides**, while the rest of 6 coins are not. That is, 33% of coins contain **orange sides** and 67% of coins do not.

Front sides of the 9 coins:



Back sides of the 9 coins:



Coin 1-6 contain **blue sides** and Coin 6-8 contain **orange sides**.

- You will have 6 tickets to bet on which color the Randomly Drawn Coin **contains**. The computer will randomly pick 1 out of 9 coins, and then 1 ticket is picked to apply on the Randomly Drawn Coin. You receive \$7 if the Randomly Drawn Coin contains the color you bet on that ticket, otherwise, \$0. This round can be chosen as the **Decision-that-counts**.

Please choose a ticket allocation. Remember the allocation must add up to 6 tickets.

How many tickets will you bet on **Blue side**?

How many tickets will you bet on **Orange side**?

Verify your choices:

Number of Tickets bet on Blue side:

Number of Tickets bet on Orange side:

Your Objective:

In each round of this block, you will have some tickets to bet on **which color** the Randomly Drawn Coin contains. On each ticket, you will choose one of the two bets:

- **Bet on Blue side:** the Randomly Drawn Coin contains a **blue side**.
- **Bet on Orange side:** the Randomly Drawn Coin contains a **orange side**.

You must verify your choices after that. The total numbers of coins, blue sides, orange sides, and tickets **VARY** from round to round.

With the example above:

You will have 6 tickets to bet on which color the Randomly Drawn Coin contains. You can choose any combination of “**Bet on Blue side**” and “**Bet on Orange side**” tickets, but the total number of tickets you choose has to be 6.

Your Payment:

If one round in this block is chosen as the **Decision-that-counts**, the computer will randomly draw 1 coin, and then randomly pick **1 ticket** to pay out. I will then check if the Randomly Drawn Coin **contains the color you bet on that ticket**.

If your bet on that ticket says **Bet on Blue side**, you will receive:

- \$7 if the Randomly Drawn Coin contains a **blue side**;
- \$0 if the Randomly Drawn Coin **does not**.

If your bet on that ticket says **Bet on Orange side**, you will receive:

- \$7 if the Randomly Drawn Coin contains an **orange side**;
- \$0 if the Randomly Drawn Coin **does not**.

Feedback after each round:

You will receive feedback after each round on:

- which ticket is picked;
- your bet on that ticket;
- which coin is randomly drawn;
- what the Randomly Drawn Coin contains;
- your payoff;
- what you could have earned by choosing the alternative on that ticket

At the end of the experiment, one of the decisions will be randomly selected as the **Decision-that-counts** for payment. Since all decisions are equally likely to be chosen, you should approach each decision as if it is the **Decision-that-counts**.

With the example above:

If you choose that 1 of your tickets say “Bet on Blue side” and 5 of your tickets say “Bet on Orange side”, then with 1/6 chance your bet is to “Bet on Blue side: the Randomly Drawn Coin contains a blue side” and with 5/6 chance your bet is to “Bet on Orange side: the Randomly Drawn Coin contains a orange side.”

If you choose that 5 of your tickets say “Bet on Blue side” and 1 of your tickets say “Bet on Orange side”, then with 5/6 chance your bet is to “Bet on Blue side: the Randomly Drawn Coin contains blue side” and with 1/6 chance your bet is to “Bet on Orange side: the Randomly Drawn Coin contains an orange side.”

Suppose Ticket 4 is drawn, If your bet on Ticket 4 says “Bet on Blue side”, you’d be paid \$7 if the Randomly Drawn Coin is from Coin 1-6, otherwise \$0. If your bet on

Ticket 4 says "Bet on Orange side", you'd be paid \$7 if the Randomly Drawn Coin is from Coin 6-8, otherwise \$0.

The **IMPORTANT** thing to remember is that to maximize your payment you should give us your BEST allocation of tickets.

{In Block 2, subjects see the same instruction and are informed that Block 2 is a repetition of Block 1. }

Block 1 (*Independence*)

This block consists of 30 rounds. In each round, you face a situation like the one described below.

You will see **2 rolls** of coins, **Roll Blue** and **Roll Orange**. Each roll has the same number of coins, some of which are colored. There are 2 types of colors, **blue** and **orange**. Colors are distributed among the coins based on the rules below:

- Each coin in **Roll Blue** either is colored with **blue** or has no color.
- Each coin in **Roll Orange** either is colored with **orange** or has no color.

The computer will randomly draw **2** coins: 1 from **Roll Blue** and 1 from **Roll Orange**. You do **NOT** know which coins are drawn. Figure A.4 is an example of the environment you might face:

Figure A.4: Screenshot of Interface in Independence Treatment

- There are **2 rolls** of coins, **Roll Blue** and **Roll Orange**. Each roll has 9 coins. Coins in **Roll Blue** are labeled b1-b9, and coins in **Roll Orange** are labeled o1-o9.
- In **Roll Blue**, 6 out of 9 coins are colored **blue**, while the rest of 3 coins are empty. That is, in **Roll Blue**, 67% of coins have colors on them, and 33% of coins have no color.

Roll Blue:



- In **Roll Orange**, 3 out of 9 coins are colored **orange**, while the rest of 6 coins are empty. That is, in **Roll Orange**, 33% of coins have colors on them, and 67% of coins have no color.

Roll Orange:



In **Roll Blue**, Coin b1-b6 have colors on them, while the rest are empty. In **Roll Orange**, Coin o6-o8 have colors on them, while the rest are empty.

- The computer will randomly pick 2 coins, one from each roll. You will have 6 tickets to bet on **from which roll** the Randomly Drawn Coin **contains color**. 1 ticket is picked. You receive \$7 if the Randomly Drawn Coin from the roll you bet on that ticket contains color, otherwise, \$0. This round can be chosen as the **Decision-that-counts**.

Please choose a ticket allocation. Remember the allocation must add up to 6 tickets.

How many tickets will you bet on **Roll Blue**?

How many tickets will you bet on **Roll Orange**?

Verify your choices:

Number of Tickets bet on Roll Blue:

Number of Tickets bet on Roll Orange:

Your Objective:

In each round of this block, you will have some tickets to bet on which coin contains color, that is, **from which roll** the Randomly Drawn Coin **contains color**. On each ticket, you will choose one of the two bets:

- **Bet on Roll Blue**: the Randomly Drawn Coin from **Roll Blue** contains color.
- **Bet on Roll Orange**: the Randomly Drawn Coin from **Roll Orange** contains color.

You must verify your choices after that.

The numbers of coins, colored coins and tickets **VARY** from round to round.

With the example above:

You will have 6 tickets to bet on from which roll the Randomly Drawn Coin contains color. You can choose any combination of “**Bet on Roll Blue**” and “**Bet on Roll Orange**” tickets, but the total number of tickets you choose has to be 6.

Your Payment:

If one round in this block is chosen as the **Decision-that-counts**, the computer will randomly draw **2 coins**, one from each roll. **1 ticket** is randomly picked to pay out. I will then check if the Randomly Drawn Coin from **the roll you bet** on that ticket **contains color**.

If your bet on that ticket says **Bet on Roll Blue**, I will check the Randomly Drawn Coin from **Roll Blue** and you will receive:

- \$7 if the Randomly Drawn Coin from **Roll Blue** contains color;
- \$0 if the Randomly Drawn Coin from **Roll Blue** **does not**.

If your bet on that ticket says **Bet on Roll Orange**, I will check the Randomly Drawn Coin from **Roll Orange** and you will receive:

- \$7 if the Randomly Drawn Coin from **Roll Orange** contains color;
- \$0 if the Randomly Drawn Coin from **Roll Orange** **does not**.

Feedback after each round:

You will receive feedback after each round on:

- which ticket is picked;
- your bet on that ticket;
- your payoff;

At the end of the experiment, one of the decisions will be randomly selected as the **Decision-that-counts** for payment. Since all decisions are equally likely to be chosen, you should approach each decision as if it is the **Decision-that-counts**.

With the example above:

If you choose that 1 of your tickets say “**Bet on Roll Blue**” and 5 of your tickets say “**Bet on Roll Orange**”, then with 1/6 chance your bet is to “**Bet on Roll Blue**: the Randomly Drawn Coin from **Roll Blue** contains color.” and with 5/6 chance your guess is to “**Bet on Roll Orange**: the Randomly Drawn Coin from **Roll Orange** contains color.”

If you choose that 5 of your tickets say “**Bet on Roll Blue**” and 1 of your tickets say “**Bet on Roll Orange**”, then with 5/6 chance your bet is to “**Bet on Roll Blue**: the Randomly Drawn Coin from **Roll Blue** contains color.” and with 1/6 chance your guess is to “**Bet on Roll Orange**: the Randomly Drawn Coin from **Roll Orange** contains color.”

Suppose Ticket 4 is drawn, If your bet on Ticket 4 says “**Bet on Roll Blue**”, you’d be paid \$7 if the Randomly Drawn Coin from **Roll Blue** is from Coin b1-b6, otherwise \$0. If your bet on Ticket 4 says “**Bet on Roll Orange**”, you’d be paid \$7 if the Randomly Drawn Coin from **Roll Orange** is from Coin o6-o8, otherwise \$0.

The **IMPORTANT** thing to remember is that to maximize your payment you should give us your BEST allocation of tickets.

{In Block 2, subjects see the same instruction and are informed that Block 2 is a repetition of Block 1. }

Block 1 (*Unknown*)

This block consists of 30 rounds. In each round, you face a situation like the one described below.

You will see a roll of 2-sided coins, some of which have colors on the front and back sides. There are 2 types of colors, **blue** and **orange**. Colors are distributed among the front and back sides of coins based on the rules below:

- The **front side** of each coin either is colored with **blue** or has no color.
- The **back side** of each coin either is colored with **orange** or has no color.
- **Different colors are NOT necessarily exclusive.** That is, it is possible that 1 coin is colored with **blue** on the front side and **orange** on the back side.

The computer will randomly draw 1 coin and check the color on each side of the coin. You do **NOT** know which coin is drawn. Figure A.5 is an example of the environment you might face:

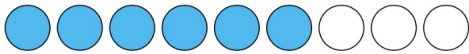
Figure A.5: Screenshot of Interface in Unknown Treatment

- You will see a roll of 9 2-sided coins, labeled 1-9:



- 6 out of 9 coins are colored **blue** on the **front sides**, while the rest of 3 coins are not. That is, 67% of coins contain **blue sides** and 33% of coins do not. 3 out of 9 coins are colored **orange** on the **back sides**, while the rest of 6 coins are not. That is, 33% of coins contain **orange sides** and 67% of coins do not.

Front sides of the 9 coins:



Back sides of the 9 coins:



- Different colors** are **NOT** necessarily **exclusive**. That is, it is possible that a coin is colored with **blue** on the front side and **orange** on the back side.
- You will have 6 tickets to bet on which color the Randomly Drawn Coin **contains**. The computer will randomly pick 1 out of 9 coins, and then 1 ticket is picked to apply on the Randomly Drawn Coin. You receive \$7 if the Randomly Drawn Coin contains the color you bet on that ticket, otherwise \$0. This round can be chosen as the **Decision-that-counts**.

Please choose a ticket allocation. Remember the allocation must add up to 6 tickets.

How many tickets will you bet on **Blue side**?

How many tickets will you bet on **Orange side**?

Verify your choices:

Number of Tickets bet on Blue side:

Number of Tickets bet on Orange side:

Your Objective:

In each round of this block, you will have some tickets to bet on **which color** the Randomly Drawn Coin contains. On each ticket, you will choose one of the two bets:

- **Bet on Blue side:** the Randomly Drawn Coin contains a **blue side**.
- **Bet on Orange side:** the Randomly Drawn Coin contains a **orange side**.

You must verify your choices after that. The total numbers of coins, blue sides, orange sides and tickets **VARY** from round to round.

With the example above:

You will have 6 tickets to bet on which color the Randomly Drawn Coin contains. You can choose any combination of “**Bet on Blue side**” and “**Bet on Orange side**” tickets, but the total number of tickets you choose has to be 6.

Your Payment:

If one round in this block is chosen as the **Decision-that-counts**, the computer will randomly draw 1 coin, and then randomly pick **1 ticket** to pay out. I will then check if the Randomly Drawn Coin **contains the color you bet on that ticket**.

If your bet on that ticket says **Bet on Blue side**, you will receive:

- \$7 if the Randomly Drawn Coin contains a **blue side**;
- \$0 if the Randomly Drawn Coin **does not**.

If your bet on that ticket says **Bet on Orange side**, you will receive:

- \$7 if the Randomly Drawn Coin contains an **orange side**;
- \$0 if the Randomly Drawn Coin **does not**.

Feedback after each round:

You will receive feedback after each round on:

- which ticket is picked;
- your bet on that ticket;
- your payoff;

At the end of the experiment, one of the decisions will be randomly selected as the **Decision-that-counts** for payment. Since all decisions are equally likely to be chosen, you should approach each decision as if it is the **Decision-that-counts**.

With the example above:

If you choose that 1 of your tickets say “Bet on Blue side” and 5 of your tickets say “Bet on Orange side”, then with 1/6 chance your bet is to “Bet on Blue side: the Randomly Drawn Coin contains a blue side” and with 5/6 chance your bet is to “Bet on Orange side: the Randomly Drawn Coin contains a orange side.”

If you choose that 5 of your tickets say “Bet on Blue side” and 1 of your tickets say “Bet on Orange side”, then with 5/6 chance your bet is to “Bet on Blue side: the Randomly Drawn Coin contains blue side” and with 1/6 chance your bet is to “Bet on Orange side: the Randomly Drawn Coin contains an orange side.”

Suppose Ticket 4 is drawn, If your bet on Ticket 4 says “Bet on Blue side”, you’d be paid \$7 if the Randomly Drawn Coin contains blue side , otherwise \$0. If your bet on Ticket 4 says “Bet on Orange side”, you’d be paid \$7 if the Randomly Drawn Coin contains orange side, otherwise \$0.

The **IMPORTANT** thing to remember is that to maximize your payment you should give us your BEST allocation of tickets.

{In Block 2, subjects see the same instruction and are informed that Block 2 is a repetition of Block 1. }

Block 3 (*Baseline/Independence/Unknown*)

This block consists of 6 rounds. In each round, you face a situation like the one described below.

You will see a roll of 10 **2-sided** coins, some of which have colors on the front and back sides. There are 2 types of colors, **blue** and **orange**. Colors are distributed among the front and back sides of coins based on the rules below:

- The **front side** of each coin either is colored with **blue** or has no color.
- The **back side** of each coin either is colored with **orange** or has no color.
- **Different colors** are **NOT** necessarily **exclusive**. That is, it is possible that a coin is colored with **blue** on the front side and **orange** on the back side.

The computer will randomly draw 1 coin. You do **NOT** know which coin is drawn. Figure A.6 is an example of the environment you might face:

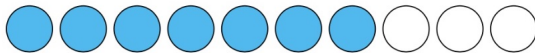
Figure A.6: Screenshot of Interface in Block 3

- You will see a roll of 10 2-sided coins, labeled 1-10:



- 7 out of 10 coins are colored blue on the front sides, while the rest of 3 coins have no color on the front sides. That is, 70% of coins contain blue sides and 30% of coins do not. 3 out of 10 coins are colored orange on the back sides, while the rest of 7 coins have no color on the back sides. That is, 30% of coins contain orange sides and 70% of coins do not.

Front sides of the 10 coins:



Back sides of the 10 coins:



- Different colors are NOT necessarily exclusive. That is, it is possible that a coin is colored with blue on the front side and orange on the back side.
- You will have 10 tickets to bet on which color the Randomly Drawn Coin contains. The computer will randomly pick 1 out of 10 coins, and then 1 ticket is picked to apply on the Randomly Drawn Coin. You receive \$7 if the Randomly Drawn Coin contains the color you bet on that ticket, otherwise, \$0. This round can be chosen as the **Decision-that-counts**.

Please choose a ticket allocation. Remember the allocation must add up to 10 tickets.

How many tickets will you bet on Blue?

How many tickets will you bet on Orange ?

Verify your choices:

Number of Tickets bet on Blue:

Number of Tickets bet on Orange:

Your Objective:

In each round of this block, you will have 10 tickets to bet on **which color of side** the Randomly Drawn Coin contains. On each ticket, you will choose one of the two bets:

- **Bet on Blue side:** the Randomly Drawn Coin contains a **blue side**.
- **Bet on Orange side:** the Randomly Drawn Coin contains a **orange side**.

You must verify your choices after that. The numbers of blue sides and orange sides **VARY** from round to round.

Your Payment:

If one round in this block is chosen as the **Decision-that-counts**, the computer will randomly draw **1 coin**, and then randomly pick **1 ticket** to pay out. I will then check if the Randomly Drawn Coin **contains the color you bet on that ticket**.

If your bet on that ticket says **Bet on Blue side**, you will receive:

- \$7 if the Randomly Drawn Coin contains a **blue side**;
- \$0 if the Randomly Drawn Coin **does not**.

If your bet on that ticket says **Bet on Orange side**, you will receive:

- \$7 if the Randomly Drawn Coin contains an **orange side**;
- \$0 if the Randomly Drawn Coin **does not**.

Feedback after each round:

You will receive feedback after each round on:

- which ticket is picked;
- your bet on that ticket;

- your payoff;

At the end of the experiment, one of the decisions will be randomly selected as the **Decision-that-counts** for payment. Since all decisions are equally likely to be chosen, you should approach each decision as if it is the **Decision-that-counts**. The color of coins may change from round to round. Your bonus payment is equal to the payoff of the selected round in dollars.

With the example above:

If you choose that 1 of your tickets say “Bet on Blue side” and 9 of your tickets say “Bet on Orange side”, then with 10% chance your bet is to “Bet on Blue side: the Randomly Drawn Coin contains blue side” and with 90% chance your guess is to “Bet on Orange side: the Randomly Drawn Coin contains orange side.”

If you choose that 9 of your tickets say “Bet on Blue side” and 1 of your tickets say “Bet on Orange side”, then with 90% chance your bet is to “Bet on Blue side: the Randomly Drawn Coin contains blue side” and with 10% chance your guess is to “Bet on Orange side: the Randomly Drawn Coin contains orange side.”

Suppose Ticket 4 is drawn, If your bet on Ticket 4 says “Bet on Blue side”, you’d be paid \$7 if the Randomly Drawn Coin contains a blue side, otherwise \$0. If your bet on Ticket 4 says “Bet on Orange side”, you’d be paid \$7 if the Randomly Drawn Coin contains an orange side, otherwise \$0.

The **IMPORTANT** thing to remember is that to maximize your payment you should give us your BEST allocation of tickets.

Appendix B

Appendix for “Preference for Sample Features and Belief Updating”

B.1 Incentive Compatibility of the Ranking-Card Method

Let $X_f = \{f_1, \dots, f_N\}$ be the set of forms and $X_m = \{m_1, \dots, m_K\}$ be the set of bundles “null information for + compensation”. Let $X = X_f \cup X_m$ be the choice set.

Assumption 1. X is well-ordered under \succsim .

Since X is a finite set, there is a utility function $u : X \rightarrow \mathbb{R}$ represents \succsim .

Let $R : X \rightarrow \mathbb{Z}$ be the *ranking* function that the agent assigns. Particularly, the agent sort the elements in X . We define $R(x)$ the number of elements *behind* x . For example, suppose an agent sort $X = \{a, b, c, d\}$ in the following order:

$$a, d, e \sim b, c$$

Note that e and b are at the same place. Then $R(a) = 4$, $R(d) = 3$, $R(e) = R(b) = 1$ and $R(c) = 0$.

In each trial, two elements in X will be chosen, and the one with higher ranking will be selected. Denote $C(\{x, y\})$ as the selected element given $x, y \in X$, then

$$C(\{x, y\}) = \begin{cases} \arg \max_{z \in \{x, y\}} R(z) & \text{if } R(x) \neq R(y) \\ x & \text{if } R(x) = R(y). \end{cases}$$

The selected element derives the agent's realized utility, $u(C(\{x, y\}))$.

We then give the main characterization of the utility function given the binary choice.

Proposition 5. *For any $x, y \in X$, $u(x) \geq u(y)$ if and only if $R(x) \geq R(y)$.*

Proof: The necessity part is trivial. We show the sufficiency part here. Assume $u(x) < u(y)$. Suppose $R(x) \geq R(y)$. Consider the case that x and y are both chosen. Then $C(\{x, y\}) = x$, and the implied utility specification is $u(x)$, which is strictly less than $u(y)$ and hence leads to a contradiction.

Since the ranking function characterizes the utility function, there is no incentive to state the preferences otherwise. Therefore, the ranking must be truthful.

B.2 Experimental Design Details

Table B.1: List of Reports

Report	Majority	Proportion	Difference	Count	Sequence (corresponding count)
1	Orange/Green				
2		0%/100%			
3		33%/67%			
4		20%/80%			
5		40%/60%			
6			± 1		
7			± 3		
8			± 5		
9			± 9		
10				3-0	
11				9-6	
12				2-1	
13				3-2	
14				10-5	
15				5-0	
16				9-0	
17				12-3	
18				4-1	
19					ooo (3-0)
20					ooooooooogggggg (9-6)
21					ogogooogooogogo (9-6)
22					ogo (2-1)
23					oog (2-1)
24					ooogg (3-2)
25					ogogo (3-2)
26					ogooogooogoo (10-5)
27					ooooooooogggggg (10-5)
28					ooooo (5-0)
29					ooooooooo (9-0)
30					oogooogooogoo (12-3)
31					oooooooooooggg (12-3)
32					oooog (4-1)
33					oogoo (4-1)

Note. The list shows the preassigned reports implemented in the experiment. In *Majority*, the subjects either read “more orange” or “more green”. In *Difference*, the listed reports represent the difference the subjects see; for instance, “ ± 3 ” means one color has 3 more balls than the other. In *Proportion*, they see the proportions of different-colored balls; for instance, “33%/67%” means 33% of balls are in one color and 67% are in the other color. In *Count*, the listed reports represent the counts the subjects see; for instance, “2-1” means 2 balls in one color and 1 ball in the other. In *Sequence*, the listed reports represent the specific sequence the subjects see; for instance, “ogo” means the subject sees a sequence of “orange-green-orange” balls. From the same report, the majority is randomly assigned. For instance, when a subject is assigned Report 20, she may be assigned “ogo” or “gog” with same probabilities.

B.3 Grether Model + Full Data

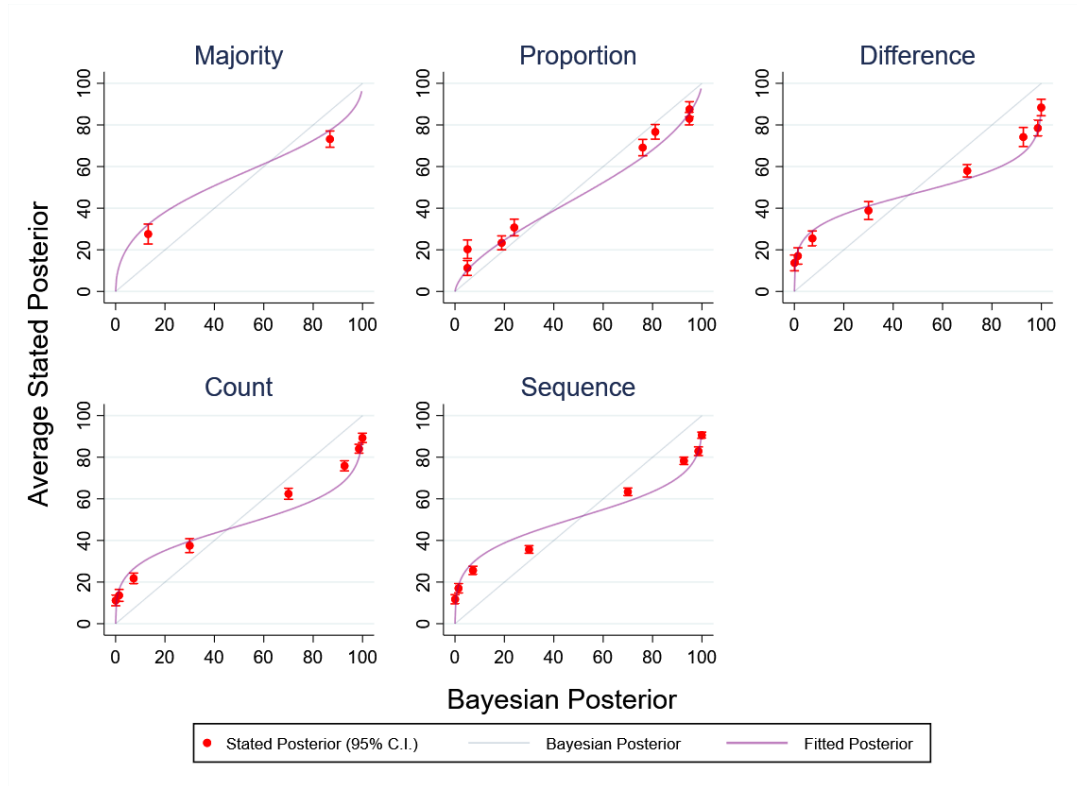


Figure B.1: Stated Belief and Bayesian Benchmark across Reports

Note: The stated posteriors are plotted against Bayesian posteriors by reports. On each point, we plot the 95% confidence interval. The blue lines represent the 45-degree line as the Bayesian benchmark. The fitted posterior is derived from Equation (2.7), where the coefficients are taken from Table 2.2. Include the linear approximation of the stated beliefs of 0% and 100%.

Table B.2: Estimated Responsiveness to changes in Likelihood Ratio with Interaction

	(1) All Five reports $\ln \left(\frac{\pi(\text{Box } O S_{7R})}{\pi(\text{Box } G S_{7R})} \right)$	(2) Without Majority $\ln \left(\frac{\pi(\text{Box } O S_{7R})}{\pi(\text{Box } G S_{7R})} \right)$	(3) Difference vs Count vs Sequence $\ln \left(\frac{\pi(\text{Box } O S_{7R})}{\pi(\text{Box } G S_{7R})} \right)$
$\ln \left(\frac{p(\text{Box } O S_{7R})}{p(\text{Box } G S_{7R})} \right)$	0.543*** (0.0536)	0.674*** (0.0325)	0.313*** (0.0180)
<i>Proportion</i>	-0.0527 (0.111)		
<i>Difference</i>	-0.131 (0.108)	-0.0785 (0.0733)	
<i>Count</i>	-0.0801 (0.101)	-0.0274 (0.0541)	0.0510 (0.0584)
<i>Sequence</i>	-0.0206 (0.0995)	0.0320 (0.0552)	0.110* (0.0559)
<i>Proportion</i> $\times \ln \left(\frac{p(\text{Box } O S_{7R})}{p(\text{Box } G S_{7R})} \right)$	0.131*** (0.0467)		
<i>Difference</i> $\times \ln \left(\frac{p(\text{Box } O S_{7R})}{p(\text{Box } G S_{7R})} \right)$	-0.230*** (0.0480)	-0.361*** (0.0246)	
<i>Count</i> $\times \ln \left(\frac{p(\text{Box } O S_{7R})}{p(\text{Box } G S_{7R})} \right)$	-0.187*** (0.0499)	-0.317*** (0.0249)	0.0431*** (0.0132)
<i>Sequence</i> $\times \ln \left(\frac{p(\text{Box } O S_{7R})}{p(\text{Box } G S_{7R})} \right)$	-0.180*** (0.0500)	-0.310*** (0.0227)	0.0498*** (0.0124)
Constant	-0.0396 (0.131)	-0.0963 (0.0983)	-0.140 (0.100)
N	3205	3108	2718

Notes: *** p -value < 0.01, ** p -value < 0.05 and * p -value < 0.1. Standard errors are clustered at the subject level with gender and grade as controls.

B.4 Report-Whatever-You-See Heuristics

It is possible that, instead of making better use of the proportion information, subjects might just naively report whatever they saw under Proportion. If the majority tends to do so and the rest performs in the identical way as under Count and Sequence, the naive resemblance could result in the finding that the stated beliefs are on average less compressed towards 50:50. We address this concern by classifying stated beliefs under Report Proportion into two types according to whether it is within $\pm 5\%$ of the proportion information provided. We find that the majority is out of the proportion $\pm 5\%$: 67% are out of proportion $\pm 5\%$, and 33% of stated beliefs are within proportion $\pm 5\%$. To further explore whether the out-of-proportion- $\pm 5\%$ type is more compressed towards 50:50 or closer to the Bayesian Benchmark, We plot the average stated posteriors against Bayesian posteriors under the Report Proportion and separate them by the two types in Figure B.2. For those out of proportion $\pm 5\%$, the stated beliefs are closer to the Bayesian benchmark than to 50:50. This result suggests that, instead of naively stating whatever subjects saw under Report Proportion, the majority indeed makes better use of the information under Proportion.

One possible explanation of the subjects’ better performance under Proportion is that the subjects are naively reporting the proportions they observe, and it naturally makes the estimated sensitivity close to one. We provide two pieces of evidence against this explanation. First, 67% of our subjects do *not* state their posterior beliefs close (plus or minus 0.05) to the actual proportion they see. Second, when we plot the stated posteriors against the Bayesian posteriors, the observations that are close to the presented proportions are showing more deviated (with respect to Bayesian) sensitivity than those are not close. Please see Figure B.2 for more details.

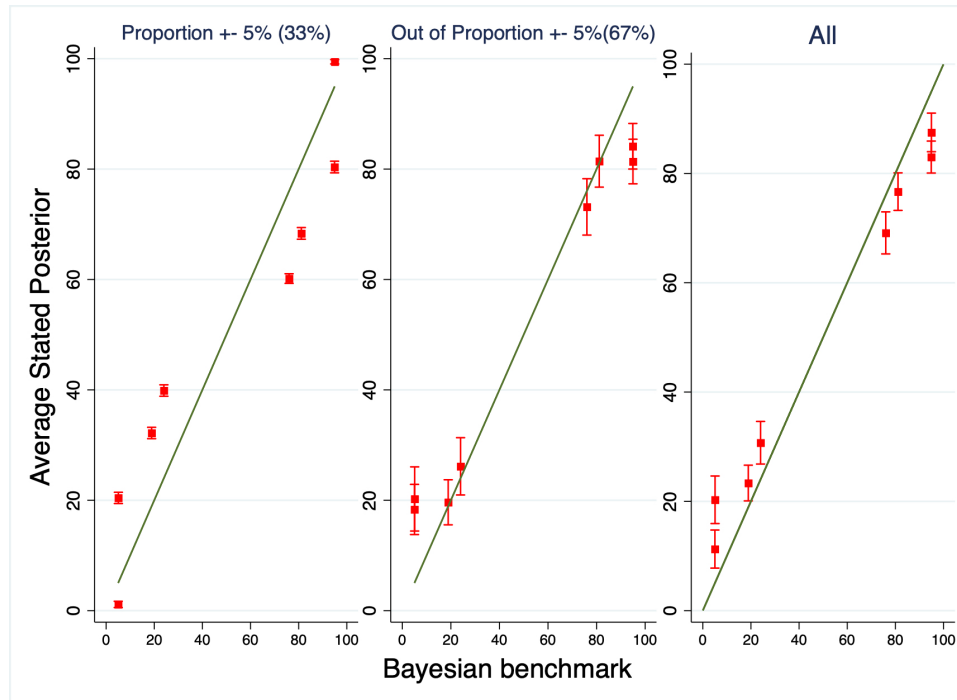


Figure B.2: Stated Beliefs under Report Proportion

Note: In the left and middle panels, we plot the average stated posteriors against Bayesian posteriors under Report Proportion and separate them by whether the stated belief is within proportion $\pm 5\%$. The percentage in the bracket is the fraction of stated beliefs which belong to the type. The right panel plots the pooled results. On each point, we plot the 95% confidence interval.

Appendix C

Appendix for “Dynamic Binary Method”

C.1 Binarized Scoring Rule and Incentive Compatibility

Consider a decision maker (DM) with a probabilistic belief over a verifiable binary outcome $s \in \{A, B\}$, assuming they possess a true belief $p = Pr\{s = A\}$. Binarized scoring rule (BSR) uses two monetary prizes M_h and M_l for payment (where $M_h > M_l \geq 0$), and two i.i.d. draws $X_1, X_2 \sim U[0, 1]$ to determine the outcome (Hossain and Okui, 2013; Wilson and Vespa, 2018). Specifically, if $s = A$ is true, the DM gets the prize M_h so long as their stated belief a is greater than at least one of the two uniform draws X_1 and X_2 . If $s = B$ is false, the DM gets the prize M_h so long as their stated belief a is less than at least one of the two uniform draws X_1 and X_2 . Otherwise, the DM gets the prize M_l .

Given the true belief p , the probability of winning the better prize M_h is given by

$$\pi(p, a) = p * (1 - (1 - a)^2) + (1 - p) * (1 - a^2) \quad (\text{C.1})$$

Thus, BSR generates a reduced lottery $\mathcal{L}(a|p) = \pi(p, a) \circ M_h \oplus (1 - \pi(p, a)) \circ M_l$. When the true belief p is a singleton and the choice space of a is continuous on $[0, 1]$ (as in the classical method, for example, slider bar), the best response is $a^*(p) = p$ as $\mathcal{L}(a^*(p)|p)$ stochastically dominates any other available lottery $\mathcal{L}(a|p)$.

Consider the situation where the true belief p follows a non-degenerate distribution $f(p)$, with $\mu_p = E(p)$ and $\sigma_p^2 = Var(p) > 0$. The DM can directly select any number between 0 and 1 as in classical methods. The distribution over p captures the idea that the perception of $Pr\{s = A\}$ can be noisy, uncertain, or imprecise (Enke and Graeber, 2023; Frydman and Jin, 2022; Giustinelli, Manski and Molinari, 2022). Without loss of generality, assume $M_l = 0$. Given the true belief p , finding the optimal stated belief $a \in [0, 1]$ that maximizes the expected utility in the BSR is equivalent to maximizing the likelihood of receiving the prize M_h . Unlike the case where the true belief p is a singleton, the optimization problem now involves maximizing the expected likelihood of receiving the prize M_h :

$$\max_a E_p[p * (1 - (1 - a)^2) + (1 - p) * (1 - a^2)] \quad (\text{C.2})$$

where a has a continuous choice space between 0 and 1, i.e., $a \in [0, 1]$. The best response in this situation is to select the point $a^*(p)$ where $a^*(p) = E(p) = \mu_p$.

As the DBM allows the DM not only to choose until a single point but also to choose a random variable with a uniform distribution over a range $[a_l, a_u]$, i.e., $a \sim Uniform[a_l, a_u]$, the optimization problem becomes:

$$\max_a E_p\{p * E_a[(1 - (1 - a)^2)|p] + (1 - p) * E_a[(1 - a^2)|p]\} \quad (\text{C.3})$$

which is equivalent to

$$\max_a \{-Var(a) - [E(a) - E(p)]^2 + E(1 - p) + [E(p)]^2\} \quad (\text{C.4})$$

where $Var(a)$ and $E(a)$ denote the variance and the mean of stated belief a , respectively. To maximize the expected utility, it is optimal to choose until the point a^* where $a^* = E(p)$ and $Var(a) = 0$.

In sum, given the true belief p , to maximize expected utility, it is optimal to select the point $a^*(p) = E(p)$. This holds true whether the true belief follows a non-degenerate distribution or if the DM is allowed to select a range or a mass point as their belief.

C.2 BDM with Myopic DM

For the DM who fails to foresee that the optimal choice is to choose until the point a^* where $a^* = E(p)$ and $Var(a) = 0$, they may compare among the three options in each step instead: $Uniform[I_l, \frac{I_l+I_u}{2}]$, $Uniform(\frac{I_l+I_u}{2}, I_u]$, or exiting with $Uniform[I_l, I_u]$, where I_l and I_u denote the upper and lower bounds of the interval in each step, respectively. Thus, the likelihood of receiving the prize M_h of choosing each option is:

- when $a = Uniform[I_l, \frac{I_l+I_u}{2}]$:

$$\begin{aligned}
& -Var(a) - [E(a) - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(\frac{I_l+I_u}{2} - I_l)^2}{12} - [\frac{(\frac{I_l+I_u}{2} + I_l)}{2} - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(I_u - I_l)^2}{12 * 4} - [\frac{3I_l + I_u}{4} - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(I_u - I_l)^2}{12 * 4} - (\frac{3I_l + I_u}{4})^2 + E(p) * \frac{3I_l + I_u}{2} + E(1-p)
\end{aligned} \tag{C.5}$$

- when $a = Uniform(\frac{I_l+I_u}{2}, I_u]$:

$$\begin{aligned}
& -Var(a) - [E(a) - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(I_u - \frac{I_l+I_u}{2})^2}{12} - [\frac{(\frac{I_l+I_u}{2} + I_u)}{2} - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(I_u - I_l)^2}{12 * 4} - [\frac{I_l + 3I_u}{4} - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(I_u - I_l)^2}{12 * 4} - (\frac{I_l + 3I_u}{4})^2 + E(p) * \frac{I_l + 3I_u}{2} + E(1-p)
\end{aligned} \tag{C.6}$$

- when exiting with $a = Uniform[I_l, I_u]$:

$$\begin{aligned}
& -Var(a) - [E(a) - E(p)]^2 + E(1-p) + [E(p)]^2 \\
&= -\frac{(I_u - I_l)^2}{12} - \left[\frac{I_l + I_u}{2} - E(p)\right]^2 + E(1-p) + [E(p)]^2 \quad (C.7) \\
&= -\frac{(I_u - I_l)^2}{12} - \left(\frac{I_l + I_u}{2}\right)^2 + E(p) * (I_l + I_u) + E(1-p)
\end{aligned}$$

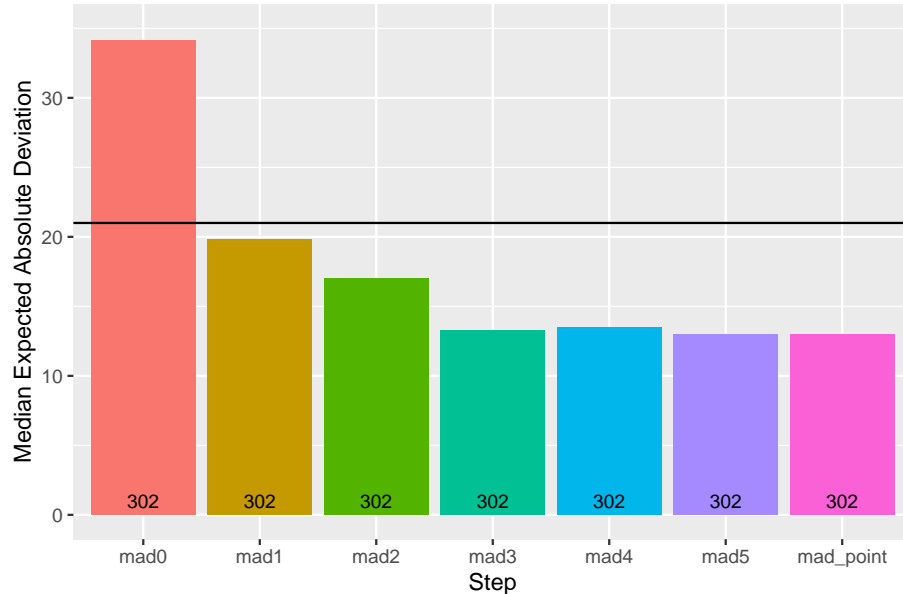
Whenever $E(p) < \frac{I_u + I_l}{2}$, choosing $Uniform[I_l, \frac{I_l + I_u}{2}]$ yields a higher likelihood of receiving M_h than choosing $Uniform(\frac{I_l + I_u}{2}, I_u]$ or exiting with $Uniform[I_l, I_u]$. Similarly, whenever $E(p) > \frac{I_u + I_l}{2}$, choosing $Uniform(\frac{I_l + I_u}{2}, I_u]$ yields a higher likelihood of receiving M_h than choosing $Uniform[I_l, \frac{I_l + I_u}{2}]$ or exiting with $Uniform[I_l, I_u]$. Whenever $E(p) = \frac{I_u + I_l}{2}$, all three options yield the same likelihood of receiving M_h . Thus, the DM would be indifferent in choosing any of the three options.

To sum up, whenever $E(p)$ is strictly within one of the two narrowed intervals, it is optimal to choose the one that contains $E(p)$. Otherwise, the myopic DM is indifferent between choosing $Uniform[I_l, \frac{I_l + I_u}{2}]$, choosing $Uniform(\frac{I_l + I_u}{2}, I_u]$, or exiting with $Uniform[I_l, I_u]$.

C.3 More Results

The observed right side of the U-shaped accuracy curve may be attributed to individuals who proceed to the last step being overconfident in their beliefs. To test this hypothesis, we examine the accuracy when individuals are compelled to halt their decision-making process earlier, specifically at the 3rd, 4th or the 5th steps. Figure C.1 shows their accuracy at each step. Contrary to expectations, our analysis reveals no significant difference in accuracy from the 3rd to the last step. If we assume that being forced to stop earlier wouldn’t alter their decisions, then their performance would not have improved even if they had stopped earlier. This suggests that individuals who reach the last step exhibit sophistication in their decision-making process. The observed heterogeneity between subjects who arrive at the last step and those who stop at the 3rd to 5th steps may stem from factors other than irrationality.

Figure C.1: Accuracy at Each Step for Subjects Reaching Point Beliefs



Note: The number at the bottom of the bar is the number of observations. The black horizontal line is the median absolute deviation in CM.

C.4 Questions used in Experiments 1 and 2

C.4.1 Experiment 1 Question Examples

1. **Inflation Rate** The computer randomly picked a year X between 1980 and 2018.

What do you think is the chance that the U.S. inflation rate in year X was lower than 7.4%?

In other words, imagine that, at the beginning of Year X, the set of products that is used to compute the inflation rate cost \$100. What do you think is the chance that, at the end of that same year, the same set of products cost less than \$107.4?

2. **S&P 500** The S&P 500 is an American stock market index that includes 500 of the largest companies based in the United States.

The computer randomly picked a year X between 1980 and 2018.

What do you think is the chance that the annual change rate of S&P 500 in Year X is less than -13%, i.e., the S&P 500 lost more than 13% of its value?

In other words, imagine that someone invested \$100 into the S&P 500 at the beginning of Year X. What do you think is the chance that, at the end of that same year, the value of the investment was less than \$87?

3. **Prior Probability** What do you think is the likelihood (percent chance) that the selected box is the Red box, the one with more red balls? (round to the nearest integer)
4. **Posterior Probability** To give you a hint of which box was selected, the computer drew a ball from the selected box.

The drawn ball is red. What do you think is the likelihood (percent chance) that

the selected box is the Red box, the one with more red balls? (round to the nearest integer)

5. **Compound Lottery** This is either a 30-70 race or a 90-10 race.

There is a 50% chance that this is a 30-70 race, otherwise this is a 90-10 race.

What do you think is the chance that the Red horse won?(round to the nearest integer)

6. **Count Peas** How many peas are there in the bowl in the picture on the left? (round to the nearest integer)
7. **Count Dots** How many dots are there in the picture on the left? (round to the nearest integer)

C.4.2 Experiment 2 Question Examples

1. **Income** In 2022, among all individuals aged 30, what is the percentage of those that are working full time that earn \$125,000 and above per year?
2. **Inflation Rate** Randomly pick a year from 1980 - 2018, what is the chance that the inflation rate in that year is lower than 2.6%?.

In other words, imagine that, at the beginning of Year X, the set of products that is used to compute the inflation rate cost \$100. What do you think is the chance that, at the end of that same year, the same set of products cost less than \$102.6?
3. **Education** How many states in 2019 have less than 29% of state-level population that have Bachelor’s degree or higher?
4. **Unemployment Rate** Randomly pick a year from 1980 - 2022, what is the chance that the unemployment rate in that year is lower than 5.5%?

5. **Posterior Probability** 50%, 50% priors, 15:1 red/blue balls in the Left urn, 1:17 red/blue in the Right urn, random draw one that is red, What's the probability it comes from the Left urn?

Bibliography

- Agranov, Marina, and Pietro Ortoleva.** 2017. “Stochastic choice and preferences for randomization.” *Journal of Political Economy*, 125(1): 40–68.
- Agranov, Marina, and Pietro Ortoleva.** 2020. “Ranges of preferences and randomization.” *Report, Princeton Univ.*[659].
- Agranov, Marina, and Pietro Ortoleva.** 2022. “Revealed preferences for randomization: An overview.” Vol. 112, 426–430, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.
- Agranov, Marina, Paul J Healy, and Kirby Nielsen.** 2023. “Stable Randomisation.” *The Economic Journal*, 133(655): 2553–2579.
- AlHewiti, Abdullah.** 2014. “Adherence to long-term therapies and beliefs about medications.” *International journal of family medicine*, 2014.
- Allen, Roy, and John Rehbeck.** 2023. “Revealed stochastic choice with attributes.” *Economic Theory*, 75(1): 91–112.
- Ambuehl, Sandro, and Shengwu Li.** 2018. “Belief updating and the demand for information.” *Games and Economic Behavior*, 109: 21–39.
- Baillon, Aurélien.** 2008. “Eliciting subjective probabilities through exchangeable events: An advantage and a limitation.” *Decision Analysis*, 5(2): 76–87.
- Baltussen, Guido, and Gerrit T Post.** 2011. “Irrational diversification: An examination of individual portfolio choice.” *Journal of Financial and Quantitative Analysis*, 1463–1491.
- Benjamin, Daniel J.** 2019. “Chapter 2 - Errors in probabilistic reasoning and judgment biases.” In *Handbook of Behavioral Economics - Foundations and Applications 2*. Vol. 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 69–186. North-Holland.
- Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer.** 2023. “How people use statistics.” National Bureau of Economic Research.

- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer.** 2012. “Salience theory of choice under risk.” *The Quarterly journal of economics*, 127(3): 1243–1285.
- Brennan, Thomas J, and Andrew W Lo.** 2012. “An evolutionary model of bounded rationality and intelligence.” *PloS one*, 7(11): e50310.
- Cabrales, Antonio, Olivier Gossner, and Roberto Serrano.** 2013. “Entropy and the Value of Information for Investors.” *American Economic Review*, 103(1): 360–77.
- Caplin, Andrew, and John Leahy.** 2001. “Psychological Expected Utility Theory and Anticipatory Feelings.” *The Quarterly Journal of Economics*, 116(1): 55–79.
- Cerreia-Vioglio, Simone, David Dillenberger, and Pietro Ortoleva.** 2015. “Cautious expected utility and the certainty effect.” *Econometrica*, 83(2): 693–728.
- Cerreia-Vioglio, Simone, David Dillenberger, Pietro Ortoleva, and Gil Riella.** 2019. “Deliberately stochastic.” *American Economic Review*, 109(7): 2425–45.
- Cettolin, Elena, and Arno Riedl.** 2019. “Revealed preferences under uncertainty: Incomplete preferences and preferences for randomization.” *Journal of Economic Theory*, 181: 547–585.
- Charness, Gary, Ryan Oprea, and Sevgi Yuksel.** 2021. “How do people choose between biased information sources? Evidence from a laboratory experiment.” *Journal of the European Economic Association*, 19(3): 1656–1691.
- Charness, Gary, Uri Gneezy, and Vlastimil Rasocho.** 2021. “Experimental methods: Eliciting beliefs.” *Journal of Economic Behavior & Organization*, 189: 234–256.
- Chen, Daniel L, Martin Schonger, and Chris Wickens.** 2016. “oTree—An open-source platform for laboratory, online, and field experiments.” *Journal of Behavioral and Experimental Finance*, 9: 88–97.
- Chew, Soo Hong, Larry G Epstein, and Uzi Segal.** 1991. “Mixture symmetry and quadratic utility.” *Econometrica: Journal of the Econometric Society*, 139–163.
- Danz, David, Lise Vesterlund, and Alistair J Wilson.** 2022. “Belief elicitation and behavioral incentive compatibility.” *American Economic Review*.
- Dertwinkel-Kalt, Markus, and Mats Köster.** 2015. “Violations of first-order stochastic dominance as salience effects.” *Journal of Behavioral and Experimental Economics*, 59: 42–46.
- Dewan, Ambuj, and Nathaniel Neligh.** 2020. “Estimating information cost functions in models of rational inattention.” *Journal of Economic Theory*, 187: 105011.

- Dustan, Andrew, Kristine Koutout, and Greg Leo.** 2022. “Reduction in belief elicitation.”
- Dwenger, Nadja, Dorothea Kübler, and Georg Weizsäcker.** 2018. “Flipping a coin: Evidence from university applications.” *Journal of Public Economics*, 167: 240–250.
- Enke, Benjamin, and Florian Zimmermann.** 2019. “Correlation neglect in belief formation.” *The Review of Economic Studies*, 86(1): 313–332.
- Enke, Benjamin, and Thomas Graeber.** 2023. “Cognitive uncertainty.” *The Quarterly Journal of Economics*, 138(4): 2021–2067.
- Epstein, Larry G, Yoram Halevy, et al.** 2019. *Hard-to-interpret signals*. University of Toronto, Department of Economics.
- Eyster, Erik, and Georg Weizsäcker.** 2011. “Correlation neglect in financial decision-making.” DIW Discussion Papers.
- Falk, Armin, and Florian Zimmermann.** 2018. “Information processing and commitment.” *The Economic Journal*, 128(613): 1983–2002.
- Feldman, Paul, and John Rehbeck.** 2022. “Revealing a preference for mixtures: An experimental study of risk.” *Quantitative Economics*, 13(2): 761–786.
- Frydman, Cary, and Lawrence J Jin.** 2022. “Efficient coding and risky choice.” *The Quarterly Journal of Economics*, 137(1): 161–213.
- Frydman, Cary, and Milica Milosavljevic Mormann.** 2018. “The role of salience in choice under risk: An experimental investigation.” *Available at SSRN 2778822*.
- Fudenberg, Drew, Ryota Iijima, and Tomasz Strzalecki.** 2015. “Stochastic choice and revealed perturbed utility.” *Econometrica*, 83(6): 2371–2409.
- Giustinelli, Pamela, Charles F Manski, and Francesca Molinari.** 2022. “Precise or imprecise probabilities? Evidence from survey response related to late-onset dementia.” *Journal of the European Economic Association*, 20(1): 187–221.
- Golman, Russell, and George Loewenstein.** 2018. “Information gaps: A theory of preferences regarding the presence and absence of information.” *Decision*, 5(3): 143–164.
- Golman, Russell, George Loewenstein, Andras Molnar, and Silvia Saccardo.** 2022. “The demand for, and avoidance of, information.” *Management Science*, 68(9): 6454–6476.

- Greiner, Ben.** 2015. "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association*, 1(1): 114–125.
- Grether, David M.** 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *The Quarterly Journal of Economics*, 95(3): 537–557.
- Griffin, Dale, and Amos Tversky.** 1992. "The weighing of evidence and the determinants of confidence." *Cognitive psychology*, 24(3): 411–435.
- Guan, Menglong, Ryan Oprea, and Sevgi Yuksel.** 2023. "Too Much Information." working paper.
- Gul, Faruk, and Wolfgang Pesendorfer.** 2006. "Random expected utility." *Econometrica*, 74(1): 121–146.
- Hossain, Tanjim, and Ryo Okui.** 2013. "The binarized scoring rule." *Review of Economic Studies*, 80(3): 984–1001.
- Hossain, Tanjim, and Ryo Okui.** 2020. "Belief formation under signal correlation." Available at SSRN 3218152.
- Kahneman, Daniel.** 1979. "Prospect theory: An analysis of decisions under risk." *Econometrica*, 47: 278.
- Kahneman, Daniel, and Amos Tversky.** 1972. "Subjective probability: A judgment of representativeness." *Cognitive psychology*, 3(3): 430–454.
- Kallir, Ido, and Doron Sonsino.** 2009. "The neglect of correlation in allocation decisions." *Southern Economic Journal*, 75(4): 1045–1066.
- Koehler, Derek J, and Greta James.** 2014. "Probability matching, fast and slow." In *Psychology of learning and motivation*. Vol. 61, 103–131. Elsevier.
- Kogler, Christoph, and Anton Kühberger.** 2007. "Dual process theories: A key for understanding the diversification bias?" *Journal of Risk and Uncertainty*, 34: 145–154.
- Kourouxous, Thomas, and Thomas Bauer.** 2019. "Violations of dominance in decision-making." *Business Research*, 12: 209–239.
- Kraemer, Carlo, and Martin Weber.** 2004. "How do people take into account weight, strength and quality of segregated vs. aggregated data? Experimental evidence." *Journal of Risk and Uncertainty*, 29: 113–142.
- Kreps, David M, and Evan L Porteus.** 1978. "Temporal resolution of uncertainty and dynamic choice theory." *Econometrica: journal of the Econometric Society*, 185–200.

- Landry, Peter, and Ryan Webb.** 2021. “Pairwise normalization: A neuroeconomic theory of multi-attribute choice.” *Journal of Economic Theory*, 193: 105221.
- Lanzani, Giacomo.** 2020. “Correlation made simple: Applications to salience and regret theory.” *The Quarterly Journal of Economics*.
- Leland, Jonathan W.** 1998. “Similarity judgments in choice under uncertainty: A reinterpretation of the predictions of regret theory.” *Management Science*, 44(5): 659–672.
- Leland, Jonathan W, Mark Schneider, and Nathaniel T Wilcox.** 2019. “Minimal frames and transparent frames for risk, time, and uncertainty.” *Management Science*, 65(9): 4318–4335.
- Liang, Yucheng.** 2021. “Learning from unknown information sources.” *Available at SSRN 3314789*.
- Liang, Yucheng.** 2022. “Learning from unknown information sources.” *Available at SSRN 3314789*.
- Liang, Yucheng.** 2023. “Boundedly Rational Information Demand.” working paper.
- Loewenfeld, Moritz, and Jiakun Zheng.** 2021. “Does correlation really matter in risk taking? An experimental investigation.”
- Loomes, Graham, and Robert Sugden.** 1982. “Regret theory: An alternative theory of rational choice under uncertainty.” *The economic journal*, 92(368): 805–824.
- Manski, Charles F.** 2004. “Measuring expectations.” *Econometrica*, 72(5): 1329–1376.
- Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa.** 2019. “Failures in contingent reasoning: The role of uncertainty.” *American Economic Review*, 109(10): 3437–74.
- Masatlioglu, Yusufcan, A Yesim Orhun, and Collin Raymond.** 2017. “Intrinsic information preferences and skewness.” *Ross School of Business Paper*.
- Miao, Bin, and Songfa Zhong.** 2018. “Probabilistic social preference: how Machina’s Mom randomizes her choice.” *Economic Theory*, 65: 1–24.
- Nielsen, Kirby.** 2020. “Preferences for the resolution of uncertainty and the timing of information.” *Journal of Economic Theory*, 189: 105090.
- Nielsen, Kirby, and Luca Rigotti.** 2023. “Revealed Incomplete Preferences.” *Available at SSRN 4622145*.
- Quiggin, John.** 1982. “A theory of anticipated utility.” *Journal of economic behavior & organization*, 3(4): 323–343.

- Quiggin, John.** 1990. “Stochastic dominance in regret theory.” *The Review of Economic Studies*, 57(3): 503–511.
- Ratcliff, Roger.** 1978. “A theory of memory retrieval.” *Psychological review*, 85(2): 59.
- Ravaioli, Silvio.** 2021. “Coarse and precise information in food labeling.” Working Paper.
- Rubinstein, Ariel.** 1988. “Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?).” *Journal of economic theory*, 46(1): 145–153.
- Rubinstein, Ariel.** 2002. “Irrational diversification in multiple decision problems.” *European Economic Review*, 46(8): 1369–1378.
- Schulze, Christin, Greta James, Derek J Koehler, and Ben R Newell.** 2019. “Probability matching does not decrease under cognitive load: A preregistered failure to replicate.” *Memory & cognition*, 47(3): 511–518.
- Shannon, C. E.** 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal*, 27(3): 379–423.
- Siegel, Sidney.** 1961. “Decision making and learning under varying conditions of reinforcement.” *Annals of the New York Academy of Sciences*.
- Tversky, Amos, and Daniel Kahneman.** 1986. “Rational Choice and the Framing of Decisions.” *Journal of Business*, 59(4 pt 2).
- Tversky, Amos, and Daniel Kahneman.** 1992. “Advances in prospect theory: Cumulative representation of uncertainty.” *Journal of Risk and uncertainty*, 5: 297–323.
- Vulkan, Nir.** 2000. “An economist’s perspective on probability matching.” *Journal of economic surveys*, 14(1): 101–118.
- Wilson, Alistair, and Emanuel Vespa.** 2018. “Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language.” Working paper.
- Wiswall, Matthew, and Basit Zafar.** 2015a. “Determinants of college major choice: Identification using an information experiment.” *The Review of Economic Studies*, 82(2): 791–824.
- Wiswall, Matthew, and Basit Zafar.** 2015b. “How do college students respond to public information about earnings?” *Journal of Human Capital*, 9(2): 117–169.
- Wolford, George, Sarah E Newman, Michael B Miller, and Gagan S Wig.** 2004. “Searching for patterns in random sequences.” *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 58(4): 221.