Hedging-Based Scoring Rules for Multiple-Choice Questions

Jingcheng Fu Xing Zhang Songfa Zhong *

September 29, 2023

Abstract

This paper proposes two novel scoring rules for multiple-choice questions based on the test-takers' propensity to hedge across possible answers. To examine these scoring rules, we randomly assign 2,986 participants in an IQ test into three conditions. In the control condition, participants choose one option, and receive one point for a correct response. In the treatment conditions, they can explicitly hedge by choosing k options: if the correct option is among the k chosen options, they receive $1/k$ point in the outcome-mixing treatment, and one point with probability $1/k$ in the probability-mixing treatment. We find that participants in both treatments hedge pervasively and score lower compared to those in the control. The observed differences depend on risk preferences of the participants. While scores in the three conditions exhibit similar psychometric quality, we observe a significant correlation between academic performance and IQ score measured in the probability-mixing condition, but not in the other two conditions. These observations contribute to the literature on the design of multiple-choice tests, preference for hedging, and the relationship between risk preferences and cognitive ability.

Keyword: convex preference, deliberate randomization, cognitive ability, multiple-choice test, experiment

JEL classification: C91, D81

*Fu: Residential College 4, National University of Singapore; Email: jingcheng.fu@gmail.com. Zhang: Graduate School of Business, Sungkyunkwan University; Email: zhangxingis@gmail.com. Zhong: Department of Economics, Hong Kong University of Science and Technology and Department of Economics, National University of Singapore; Email: zhongsongfa@gmail.com.

1 Introduction

Central to the choice under risk and uncertainty are preferences for hedging—"don't put all your eggs in one basket". Debreu (1959) provides an example with two uncertain options: one associates outcome b with Heads of a coin toss and outcome c with Tails, and the other associates c with Heads and b with Tails. Individuals with convex preference can hedge their risk through *outcome mixing*, and prefer $(b+c)/2$ with certainty to either uncertain option. He coins the notion of convex preference, a preference for mixing two uncertain consumption options to either uncertain consumption. Moreover, individuals can also hedge away the uncertainty about the states or their tastes through *probability mixing*, whereby mixing between two equally valued lotteries probabilistically is preferred to either lottery. Probability mixing has been widely discussed and commonly observed in decision making under risk and uncertainty [1]

Building on decision makers' general tendency to hedge, here we propose two alternative scoring rules for multiple-choice question (MCQ) tests, a widely used form of assessment in standardized tests. Under the standard scoring rule of MCQ, test-takers are allowed to choose only one option among n options, and receive one point for correct answer and zero otherwise. In the alternative scoring rules allowing preferences for hedging, test-takers can choose k out of n options ($1 \leq k \leq n$). In one scoring rule, the test-taker receives $1/k$ point if the correct option is among the k chosen options, and zero otherwise. We call this the *outcome-mixing* (OM, henceforth) scoring rule. In another scoring rule, the test-taker receives one point with the probability $1/k$ if the correct option is among the k options, and zero otherwise. We call this the *probability-mixing* (PM, henceforth) scoring rule.

Offering mixing options may affect scores in two different ways. First, it may lower the expected scores of test-takers under both scoring rules. Specifically, if test-takers are

[1] See Machina (1985), Chew et al. (1991), Cerreia-Vioglio (2009), Cerreia-Vioglio et al. (2015), Fudenberg et al. (2015), and Cerreia-Vioglio et al. (2019) in the domain of risk, and Raiffa (1961), Schmeidler (1989), Maccheroni et al. (2006), Dominiak and Schnedler (2011), Bade (2015), Saito (2015), Eichberger et al. (2016),Oechssler et al. (2019), Wakker and Yang (2019), Ke and Zhang (2020), Aoyama and Hanaki (2021), Baillon et al. (2022*a*), and Baillon et al. (2022*b*) in the domain of uncertainty. Dean and Ortoleva (2017) introduce a notion of preference for hedging for both risk and uncertainty.

certain about the correct answer, they should choose the correct answer instead of mixing across options; if test-takers are uncertain about which option is correct, they should choose the option with the highest perceived chance instead of mixing across options (except for the case with even chance) when they are to maximize the expected scores. Nevertheless, mixing can be preferred when test-takers have preferences for hedging. Namely, when risk-averse test-takers are uncertain about the correct answer, they may hedge by choosing multiple options to guarantee a fixed share of the point in OM. In this regard, more risk-averse test-takers are more likely to mix multiple options and score lower due to outcome mixing. Similarly, when ambiguity-averse test-takers entertain a set of priors about which is the correct answer, they may hedge away the ambiguity by randomizing among multiple options in PM.[2] Apart from preference explanations, test-takers with a false sense of diversification would also mix more and score lower in both treatments.

Second, the inclusion of mixing options may help improve the time allocation and hence possibly increase the scores of test-takers. It is commonly observed that decision makers tend to spend more time on the decisions that they are more indecisive or uncertain about which one to choose (Konovalov and Krajbich 2019, Agranov and Ortoleva 2017), but spending more time on a difficult choice does not necessarily lead to a better decision (Gill and Prowse 2023, Sunde et al. 2022). In the context of MCQ, test-takers often face some time constraint, but linger over difficult questions, leading to suboptimal time allocation. In this regard, the inclusion of mixing options may help improve time allocation and result in higher scores.

To assess these two alternative scoring rules, we conduct a between-subject experiment with 2,986 undergraduate subjects in an IQ test and randomly assign them into three conditions: the control condition using standard MCQ scoring rule, the OM treatment, and the PM treatment. We use 20 MCQ questions from an open-source IQ test developed and validated by Blum and Holling (2018). Similar to Raven's Progressive Matrices, each

[2]Probability mixing gives rise to a probability mixture of options with different likelihoods of being correct and hence violates the first order stochastic dominance when we consider risk instead of ambiguity. We provide more details in Section 2 on theoretical background.

question is an independent analogical reasoning task: subjects are asked to identify the missing item from a figural matrix out of eight options. After the IQ test, we also measure the degree of risk aversion using hypothetical binary choices and self-reported risk attitude question from the Global Preference Survey (Falk et al. 2018), and the tendency of false diversification among lotteries with the first order stochastic dominance (Rubinstein 2002). This enables us to examine not only the average treatment effect of the scoring rules, but also the differential treatment effects based on individual risk preferences.

We have four main observations. First, mixing option is pervasively chosen. 78 percent of the subjects in OM and 67 percent in PM choose multiple options at least once, and the average frequency of mixing among those subjects is 39 percent in OM and 28 percent in PM. Subjects are more likely to mix when facing harder questions or a higher time pressure. Second, subjects in both treatments score lower in the IQ test than those in the control condition. More specifically, compared to the Control, the scores are lowered by 0.11 and 0.07 standard deviation in OM and PM, respectively. Third, differential treatment effects are observed for risk preferences. More risk-averse subjects score lower in OM, but not in the Control and PM. While subjects with a stronger sense of false diversification score lower in all three conditions, the effect does not differ across conditions. Fourth, the three scoring rules have comparable psychometric quality, and the measured IQ score is significantly correlated with academic performance measured by grade point average for subjects in PM, but not in OM or the Control.

Our study contributes to several strands of literature. First, our study contributes to the design of multiple-choice tests. In the standard scoring rule, while subjects can only choose one option, they may guess randomly in their mind when they are uncertain about the correct option (Budescu and Bar-Hillel 1993, de Finetti 1965). Random guessing may produce noise in measuring performance, and leave subjects' partial knowledge in each question unobserved to the test-administrators. To reduce random guessing, a popular alternative scoring rule is negative marking—the introduction of a penalty for incorrect answers. With negative marking, test-takers can skip some questions to receive zero point,

which can be regarded as a way to hedge the risk of penalty for guessing the answer. While negative marking helps discourage random guessing, it can be biased against females, possibly due to loss aversion (Baldiga 2014, Coffman and Klinowski 2020). Relatedly, Zapechelnyuk (2015) provides justifications for a scoring rule that is similar to OM, and shows theoretically that it exhibits some desirable properties including simplicity, non-negative scores, discouragement of random guessing, and rewards for partial answers. In addition to OM, here we examine PM in which random guessing is explicitly allowed and thus partial knowledge is observable. We observe that these two scoring rules do not improve the time allocation and lower the scores of the subjects. We further show that these two alternative scoring rules exhibit similar psychometric quality to the standard scoring rule, and do not give rise to gender bias in test performance. In this regard, we envisage that OM and PM can be useful for considerations in designing MCQ tests.

Second, our study sheds light on measuring latent intelligence using MCQ. While IQ scores are commonly believed to measure some form of intelligence and are predictive of a wide range of life outcomes such as academic performance and job success, a substantial body of research casts doubt on the extent that IQ scores fully capture latent intelligence. For example, IQ tests may fail to measure the broad scope of intelligence such as creativity, and may be subjective to differential item functioning whereby test-takers with same latent intelligence from different groups, such as age, gender and culture, give different answers to the same question and receive differ scores from same tests (Embretson and Reise 2013). Moreover, it has been documented that scores in IQ tests and more generally scholastic tests can be systematically affected by the test-takers' non-cognitive traits including intrinsic motivation, attentiveness, and achievement motivation (Snyderman and Rothman 1987, Borghans et al. 2008, Duckworth et al. 2011, Gneezy et al. 2019). Adding to these studies, we demonstrate that risk preferences influence how test-takers respond to different scoring rules, and consequently IQ scores may reflect not only test-takers' latent intelligence but also their risk preferences.

Third, our observation reinforces the argument that cognitive and non-cognitive skills can be intertwined in part due to measurement. Numerous studies show that risk pref-

erences revealed from observable choices are correlated with cognitive ability (Frederick 2005, Burks et al. 2009, Dohmen et al. 2010, 2018, Benjamin et al. 2013, Rustichini 2015, Chapman et al. 2018, Falk et al. 2018, Lilleholt 2019). It has been widely recognized that cognitive ability can confound the measurement of risk preferences. For instance, individuals with low cognitive ability tend to choose randomly, giving rise to noises and biases depending on the elicitation methods. Consequently, this blurs the inference of the underlying preferences and contribute to a spurious correlation between cognitive ability and risk preferences (Andersson et al. 2016, Taylor 2016, Chapman et al. 2018, Andersson et al. 2020, Amador-Hidalgo et al. 2021). Yet, little has been done to examine whether risk preferences can conversely affect the measurement of cognitive ability. Our study fills this gap and shows that more risk-averse subjects mix more and score lower in OM, but not in PM and the Control, and subjects with a false sense of diversification score lowers in all three conditions.

Last, our study adds to the literature on preferences for hedging. In addition to OM as a more standard form of hedging, a recent literature investigates randomization as an alternative form of hedging (Agranov and Ortoleva 2022). It has been observed that individuals randomize deliberately in repeated decisions and choice lists (Agranov and Ortoleva 2017, Dwenger et al. 2018, Agranov et al. 2023, Feldman and Rehbeck 2022, Chew et al. 2022), making use of external randomization devices such as flipping a coin (Cettolin and Riedl 2019, Agranov and Ortoleva 2023, Levitt 2020, Zhang and Zhong 2020), and paying a small cost to delegate choice to external devices (Agranov and Ortoleva 2017, Cettolin and Riedl 2019). Complementing those earlier studies, the "correct" answers are observed and we can use participants' performance in the task to measure the impact of hedging on decision quality, which allows us to study the mechanisms underlying the two hedging behaviors. We find that both forms of hedging are prevalent and both are related to decision difficulty and time pressure. However, outcome mixing is significantly associated with risk aversion but probability mixing is not. Most importantly, here we show that the notion of preferences for hedging can be used to help design alternative scoring rules in MCQ.

The rest of the paper is organized as the follows. Section 2 provides the theoretical background, and section 3 describes the experimental design. The results are presented in Section 4, and concluding remarks in Section 5.

2 Theoretical Background

This section provides the theoretical background on how preferences for hedging can be used to design scoring rules. We focus on our experimental setting of MCQ with outcome and probability mixing. For simplicity, we consider that each MCQ has only two choices: option A and option B. Our experiment consists of three conditions. In the control condition, the choice set is $\{A, B\}$, in which individuals choose between the two options. If they choose the option that is the correct answer, they receive one point for this question. In the outcome-mixing condition, an outcome mixing option, denoted as $0.5A + 0.5B$, is included as the third option, and the choice set is $\{A, B, 0.5A + 0.5B\}$. If individuals choose $0.5A + 0.5B$, they receive 0.5 point with certainty. In the probability-mixing condition, a probability mixing option, denoted as $0.5A \oplus 0.5B$, is included as the third option, and the choice set is $\{A, B, 0.5A \oplus 0.5B\}$. If individuals choose $0.5A \oplus 0.5B$, a computer program will randomly choose either A or B as the chosen option with equal probability. If the randomly chosen option is the correct answer, individuals receive one point, and zero otherwise.

When individuals are uncertain about whether A or B is the correct answer, they see the options as lotteries. Option A is denoted as a lottery $F_A = \{1, p\}$, meaning that A is the correct answer with probability p and delivers one point, and delivers zero point otherwise. Correspondingly, option B is denoted as a lottery $F_B = \{1, 1 - p\}$, meaning that B is the correct answer with the complementary probability $1 - p$ and delivers one point, and delivers zero point otherwise. We assume that individuals believe that A is more likely to be the correct answer than B, that is, $p \geq 0.5$. The outcome mixing option is a degenerate lottery $F_{0.5A+0.5B} = \{0.5, 1\}$, which delivers 0.5 point with certainty. The probability mixing option is 50 percent chance of receiving $F_A = \{1, p\}$ and 50 percent chance of receiving $F_B = \{1, 1 - p\}$, which can be reduced to a simple

lottery $F_{0.5A\oplus 0.5B} = \{1, 0.5\}$, 50 percent chance of receiving one point.

Note that the expected score is p for choosing option A and 0.5 for choosing either outcome mixing or probability mixing. In this regard, the expected scores would be lower in the two treatments if participants choose to mix. Nevertheless, individuals may choose the mixing option due to their risk preferences or heuristics as detailed below.

Risk Aversion. Choosing the outcome mixing option can be formally rationalized by risk aversion. Consider individuals who are expected utility maximizers. They choose the outcome mixing option over option A if and only if $u(0.5) \geq p \cdot u(1)$, that is, $p \leq \frac{u(0.5)}{u(1)}$. Individuals would choose the outcome mixing option when they are risk averse, $\frac{u(0.5)}{u(1)} > 0.5$, and their assessed probability p is within the interval of $[0.5, \frac{u(0.5)}{u(1)}]$. Put it differently, when individuals are risk averse and not sufficiently confident about A being the correct answer, they choose the outcome mixing option. Moreover, when individuals are more risk averse, the interval $[0.5, \frac{u(0.5)}{u(1)}]$ is wider, so they are more likely to choose the outcome mixing option and score lower.

Preference for Randomization. Choosing the probability mixing option cannot be rationalized by most models of decision making under risk. More specifically, if individuals choose the probability mixing, they violate the first order stochastic dominance, as $F_A = \{1, p\} \succsim F_{0.5A\oplus 0.5B} = \{1, 0.5\}$.[3] In particular, while models with convex preference allow a preference for randomization between lotteries (Machina 1985, Chew et al. 1991, Cerreia-Vioglio et al. 2015, 2019), they satisfy the first order stochastic dominance and thus cannot account for choosing the probability mixing option.

Individuals may consider the choices as ambiguous lotteries when they are unsure about the exact probability about which choice is the correct answer. Namely, option A is a lottery $F_A = \{1, p\}$ with $p = [p_*, p^*]$, meaning that A is correct answer with probability p ranging from p_* to p^*. Correspondingly, option B is a lottery $F_B = \{1, 1-p\}$ with complementary probability $1 - p = [1 - p^*, 1 - p_*]$. In response to the Ellsberg paradox,

[3]Note that the reduction of compound lottery axiom is assumed. If two stage utility models are considered with the relation of the reduction of compound lottery axiom, choosing the probability mixing option violates the first order stochastic dominance in the stage 1 risk.

Raiffa (1961) suggests individuals can hedge away ambiguity through randomization. [4] In our context, the probability mixing option can be viewed as randomization through a coin flip, in which individuals choose option A with Heads, and choose option B with Tails. Probability mixing gives rise to an even-chance objective lottery associated with the coin flip, and thus eliminates the ambiguity about which answer is correct. In this regard, the probability mixing option is preferred to option A when ambiguity-averse individuals are unsure about the probability $p = [p_*, p^*]$.

We also would like to note that random guessing in the control condition can also be rationalized by preference for randomization when decision makers can randomize internally. For instance, in the notion of deliberate randomization (Machina 1985, Cerreia-Vioglio et al. 2019), when the decision maker is explicitly given a choice set of lotteries, she implicitly considers the convex hull of these lotteries. In our setting, the decision maker would consider all the probability mixtures of the multiple choices regardless of whether she is in the control condition or PM. Put it differently, if the decision maker relies on internal randomization to choose from the convex hull, the inclusion of an external randomization device in PM will not change her choice behavior or performance.

Naive Diversification. While the heuristic rule of diversification is useful, it can be wrongly applied to the various choice settings. For example, Benartzi and Thaler (2001) observe that investors often follow a naive diversification strategy as "1/n strategy", by dividing their contributions evenly across the funds offered in the plan. Rubinstein (2002) shows that subjects choose to mix among options even when mixing violates the first order stochastic dominance, and he interprets this as a false sense of diversification (for related discussions, see also Eliaz and Fréchette 2008, Agranov et al. 2023). In the MCQ, when individuals believe that A is more likely to be correct than B, they may falsely think that risk can be diversified by mixing the options. In this regard, subjects with a stronger false sense of diversification may choose to diversify naively in both treatments leading to lower scores in the IQ test. Moreover, they may also be more likely to guess randomly in the control condition and thus score lower as well.

[4] See Saito (2015), Ke and Zhang (2020), and Maccheroni et al. (2006) for detailed theoretical analysis on randomization under ambiguity.

Summary. While choosing either probability or outcome mixing option would lower the expected scores of the individuals, we show that the mixing options may be chosen due to either preferences or heuristics of the individuals.[5] More specifically, outcome mixing may be favorable to risk-averse individuals, and probability mixing may be preferred when individuals are to hedge away ambiguity through randomization. Individuals may also choose to mix in both conditions when they have a false sense of diversification.

3 Experimental Design

In our experiment, subjects are randomly assigned to conditions that differ in the availability of hedging options for taking an IQ test. We hypothesize that hedging behavior can stem from preferences under risk and uncertainty, as well as tendency of false diversification. To examine the underlying mechanisms of hedging behavior, after subjects complete the IQ test, we administer a series of questionnaires to measure the risk attitude, tendency of false diversification, and personality traits which may relate to hedging. Below we describe the tasks in detail.

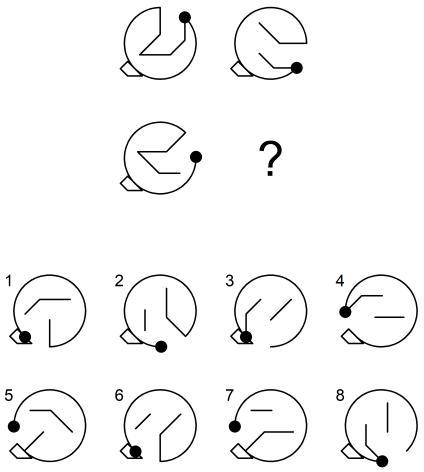
3.1 The IQ test

Our IQ test takes the form of multiple-choice questions with single correct options. Questions are generated by the Figural Analogy Test package, an intelligence measure that is similar to Raven's Progressive Matrices and available in the public domain (Blum and Holling 2018). In each question, there is a two-by-two matrix, in which three shapes are presented and the fourth shape is missing. Subjects are asked to find the missing

[5] In the IQ test, multiple MCQs may enable individuals to hedge across decisions. Samuelson (1963) recounts a story of his lunch colleague who would rather not risk \$100 for a lottery with 50 percent chance of winning \$200, but is willing to accept a sequence of 100 such bets. He shows that, under expected utility, if a lottery is rejected at any initial wealth level, a sequence of independent replicas of the lottery will be rejected. Chew and Epstein (1988) remark that the discrepancy between one decision and 100 decisions can be rationalized when non-expected utility models are considered. In this regard, if individuals choose the outcome mixing option in one MCQ, whether they would continue to do so in a number of MCQs depends on the utility of the individuals. In addition, when individuals face multiple decisions, they may exhibit narrow bracketing—one decision at a time without full regard to other decisions (Tversky and Kahneman 1981, Gneezy and Potters 1997, Thaler et al. 1997, Rabin and Weizsäcker 2009). In a similar vein, individuals may decide whether to choose the option to mix for each of the MCQs separately, rather than consider the possibility to hedge across all the MCQs.

shape out of eight options presented below the matrix, by applying the analogous changes among the shapes. Each question may involve a single change, or several changes (see Figure 1 for an example with two changes). The number of changes varies between one and four. A larger number of changes implies a higher level of difficulty, as each additional change demands more cognitive operations to solve the question. The test consists of 20 questions, which are evenly distributed across the four difficulty levels and randomly ordered.

Figure 1: Example Question in the Figural Analogy Test



Notes: Two changes are present in this example question: (i) a rotation of the main shape, and (ii) a subtraction of the lines inside the main shape. The correct answer is option 2.

3.2 Experimental conditions

Subjects are randomly assigned to one of the three treatments which differ by the scoring rule. In the control condition, subjects can only choose one option out of the 8 options, and receive the full score if the option they choose is correct. In two other treatments, subjects may choose a single option, which is marked the same way as in the control condition, or they can choose multiple options, say, k out of the 8 options.

In the *outcome-mixing* (OM) treatment, when multiple options are chosen, if the correct option is among the k chosen options, the subject will receive a fractional score, which is equal to the full score divided by the number of chosen options, i.e., $1/k$, and zero otherwise.

In the *probability-mixing* (PM) treatment, when multiple options are chosen, the computer will randomly select one option out of k options with equal probability, and assign the full score if that option is correct and zero otherwise.

Subjects are also randomly assigned with the time constraint of 20 or 40 minutes for completing the test. The variation in time constraints allows us to test the sensitivity of the effects of the scoring rules.

3.3 Test procedure

Prior to the introduction of the scoring rule, the nature of the test is explained, and subjects go through five practice questions, which take the same format for all subjects regardless of the treatment. Subjects are asked to identify the correct option, key in the option, and then click a button to check whether their answer is correct, and if it is not, they are informed of the correct option. They are free to spend as long as they intend on those questions. The practice duration and practice score (i.e., the count of correct answers from 0 to 5) are used as measures of test motivation and ability which are independent of the treatment assignment.

Subjects complete comprehension questions which cover both the length of the test and how the score is determined under a few hypothetical scenarios. All of the scenarios reinforce that only one option is correct, as the description states which option is the correct option. Two scenarios describing single choices are common to all conditions. For the OM and PM treatments there are two additional scenarios in which multiple options are chosen. The number of mistakes that a subject made in answering the comprehension questions common to all treatments offers a measure of the subject's understanding.

Subjects are told that the questions may vary in the level of difficulty, but they are not informed of the ordering or the difficulty composition of the questions. Subjects are not allowed to move on to the next question without answering the current question, or to revisit a question that has already been answered. Throughout the IQ test, a countdown timer is displayed at the top of the screen, which informs subjects the time they have

left. After the IQ test, we include a set of questionnaires to measure the non-cognitive skills as below.

3.4 Risk preferences and personality traits

3.4.1 Risk attitude

We measure risk attitude by questions taken from the Global Preference Survey (Falk et al. 2018), which consist of two parts. The first part elicits the certainty equivalent of a lottery. In a series of five questions, subjects make a hypothetical choice between a sure amount x and a lottery which offers 50:50 chance of receiving 300 RMB (1 RMB $\approx$ 0.14 USD at the time of the experiment). In the first question, x is equal to 160 RMB, and in the following questions, x depends on the previous choices: it increases if the lottery is chosen (all the way up to 310 RMB in the fifth question if the lottery is always chosen in the first four questions), and decreases if the sure amount is chosen (all the way down to 10 RMB). The second part asks the subject to report how willing or unwilling they are to take risks, on a scale between 0 and 10.

Following Falk et al. (2018), the answers from the two parts are standardized within our sample and a weighted average of standardized responses are taken as the measure of the subject's risk attitude. Falk et al. (2018) calculate the optimal weights by linking the survey responses to incentivized experimental measures of risk attitudes using data from a validation study. The weights are equal to 0.473 for the hypothetical-choice part, and 0.527 for the self-reported part.

3.4.2 False diversification

We measure the subject's proneness to false diversification, that is, the tendency to make diversified choice that violates first order stochastic dominance, based on two hypothetical questions from Rubinstein (2002). The first question is a one-stage decision problem, framed as choosing the gate of a mall to wait for a friend, hereby referred to as the Gate problem. Subjects are informed that the proportion of visitors entering by

each gate is 21 percent North, 27 percent East, 32 percent South and 20 percent West. They are asked to assign a probability to each gate to indicate their choice, with consistency imposed, such that the four probabilities add up to 100 percent. The dominant strategy is to choose South, that is, allocating 100 percent probability to South. The second question is framed as guessing for the colors of five cards randomly drawn from a deck, hereby referred to as the Card problem. Subjects are informed that there are 100 cards including 36 Green, 25 Blue, 22 Yellow, and 17 Brown, and are asked to make five guesses and imagine each correct guess is rewarded 10 RMB. The dominant strategy is to choose Green for all five guesses.

We classify the decisions by whether the subject takes the dominant strategy which maximizes the winning probability. Subjects who deviate from the dominant strategy are identified as having a tendency to take false diversification.

3.4.3 Personality Traits

Earlier studies documented that personality traits are associated with preferences under risk and uncertainty. For example, Rustichini et al. (2016) found that neuroticism and extraversion are linked to attitudes towards risk and ambiguity. Here we explore whether personality traits are related to hedging behavior. We measure the Big Five personality traits which include five basic personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism.[6] We use a 10-item abbreviated Big Five Inventory by Rammstedt and John (2007). We also measure subjects' tendency to pursue "the best choice" (Schwartz et al. 2002) using a 9-item maximizer scale by Diab et al. (2008), and the extent to which subjects experience regret after the decision has been made using a 5-item regret scale by Schwartz et al. (2002). Each item requires subjects to evaluate how well a sentence describe their personality and rate it on a 7-point Likert scale. Average scores for each dimension of the Big Five and for maximizer and

[6]Openness to experience captures the degree to which a person needs intellectual stimulation, change and variety. Conscientious measures the willingness to comply with conventional rules, norms and standards. Extraversion describes the degree of which a person needs attention and social interaction. Agreeableness describes the needs for pleasant and harmonious relations with others. Neuroticism describes how much a person experiences the world as threatening and beyond their control.

regret scales and converted into the proportion of maximum possible score.

3.5 Implementation

We pre-registered our experiment in the AEA RCT registry including the design and a simple power analysis. The sample size of 3,000 subjects is required for two reasons. First, we hypothesize that subjects respond differently to the treatments depending on preferences and personality, and only a large sample size can accommodate the analysis of heterogeneous treatment effects. Second, the effect sizes of the scoring rules are potentially small, as the inherent cognitive ability is likely to explain a large portion of individual difference. Given that IQ and achievement tests are widely used in various settings, we believe that a small effect size is of economic importance.

Our subjects are first-year undergraduate students from two universities, Zhejiang University of Science and Technology and Hangzhou Normal University, in Hangzhou, China. The students were required to attend a series of training sessions as part of their curriculum. In one of those training sessions, they were invited to complete the study as a psychometric test assessing their cognitive ability, risk attitude, and personality. At the end of the study, participants received private feedback on their performance on the IQ test as well as their scores in the risk attitude tasks and personality tests. To encourage completion and effort, 10 randomly drawn participants from each university were provided a flat fee (RMB100) plus a performance-based payment (RMB15 times the total score from the IQ test). The chance of receiving payment for the test is less than 1%, which means our study is closer to no incentive. The experimental instructions were in Chinese and the translation is provided in Appendix B. The experiment was programmed using oTree (Chen et al. 2016). Subjects were randomly assigned to the three conditions by the program. The randomization was at individual level, and communication during the session was not allowed.

While we recognize the importance of providing sufficient incentives in experiments, we do not provide monetary incentives for their response for various considerations. First, not providing monetary incentives aligns with common practices in IQ tests. Second,

providing incentive for IQ tests may create additional issue with crowding out effect which complicates the experiment. We note that the average score in our sample is substantially higher than observed in unincentivized settings: for a similar Figural Analogy Test, Blum et al. (2016) report an accuracy rate of 0.313 among over 400 university students and we observe an average of 0.473 pooling participants in all conditions. Moreover, all the results presented in the next section are robust to excluding the 11% of participants have spent less than five minutes on the IQ test. Those participants are not sufficiently motivated by the tasks, but they are evenly distributed across conditions ($p = 0.237$, χ^2 test).

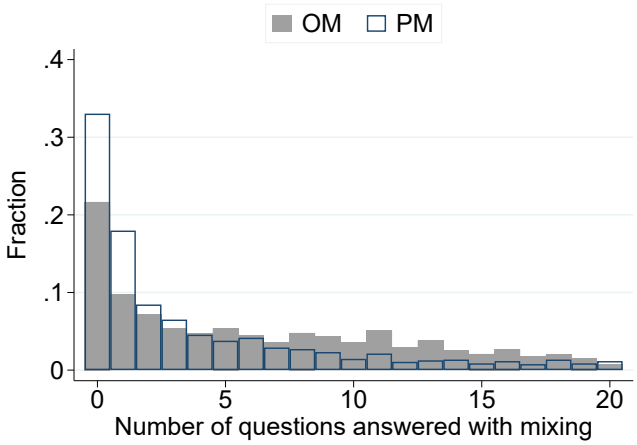
4 Results

4.1 Mixing behavior

We first examine the prevalence of mixing behavior across two treatment conditions. Figure 2 displays the frequency of mixing, i.e., the number of questions on which more than one options are chosen, at subject level. As can be seen, in both treatments, the large majority of the subjects exhibit preferences for hedging: 78.3 percent of the subjects in OM and 67.0 percent of those in PM have accessed the options to mix at least in one of the 20 questions. Mixing is not only widespread but also frequent for subjects who do mix: those in OM mix on 7.8 questions (39.0 percent of all questions), and those in PM mix on 5.6 questions (28.1 percent). The difference between OM and PM is significant (Wilcoxon rank-sum test, $p < 0.001$). Conditional on that the subject chooses to mix, the average number of options chosen is 3.1 in OM and 3.5 in PM; for questions answered with mixing, the majority of answers mix between two or three options (74% in OM and 71% in PM). To sum, we have the following first observation.

OBSERVATION 1. *Mixing behavior is prevalent and frequent in both OM and PM treatments.*

Figure 2: Frequency of mixing by condition



We investigate the mixing behavior in regressions.[7] We regress the proportion of questions answered with mixing on the dummy indicators for PM treatment and the 20-minute time constraint.[8] In addition to subjects' gender, age, and university, we control for subjects' practice score, time spent on the practice stage, and mistakes in the control questions to capture subjects' benchmark cognitive ability and motivation.

As can be seen in Table 1, subjects in the OM treatment choose to mix 12.0 percentage points more often than those in the PM treatment, and the difference is not sensitive to the inclusion of controls. This result confirms that, compared to hedging by probability mixing, subjects are more likely to hedge by outcome mixing. In addition, subjects tend to mix more when facing a larger time pressure: those under the 20-minute time constraint mix significantly more compared to those under the 40-minute time constraint; the difference is 4.3 percentage points when all other controls are included. When we separately analyze the two time constraints, the differences between OM and PM are qualitatively similar (Table A2).

In terms of subjects' initial understanding of the questions before the treatment, those who have higher practice scores mix significantly less: every point increase in the five-

[7]Throughout the paper, the p-values reported are two sided. For the regression analysis, when the unit of observation is the individual subject, we use the robust standard errors; when the unit of observation is the question in the IQ test, we cluster the standard errors at subject level.

[8]If we use the average number of options chosen instead of the proportion of questions answered with mixing as a subject-level measure of mixing behavior, we get similar results.

Table 1: Determinants of hedging by mixing

	(1)	(2)	(3)
PM	-0.120^{***}	-0.117^{***}	-0.118^{***}
	(0.012)	(0.012)	(0.012)
Time constraint (20 min)	0.046^{***}	0.045^{***}	0.043^{***}
	(0.012)	(0.012)	(0.012)
Practice score		-0.040^{***}	-0.038^{***}
		(0.006)	(0.006)
Logged practice time		0.048^{***}	0.043^{***}
		(0.012)	(0.012)
Mistakes in control q.		0.003	0.006
		(0.009)	(0.009)
Female			0.051^{***}
			(0.013)
Age			0.008
			(0.007)
Second university			0.010
			(0.014)
Constant	0.288^{***}	0.121	0.016
	(0.011)	(0.076)	(0.149)
Observations	1989	1989	1986
R-squared	0.055	0.086	0.097

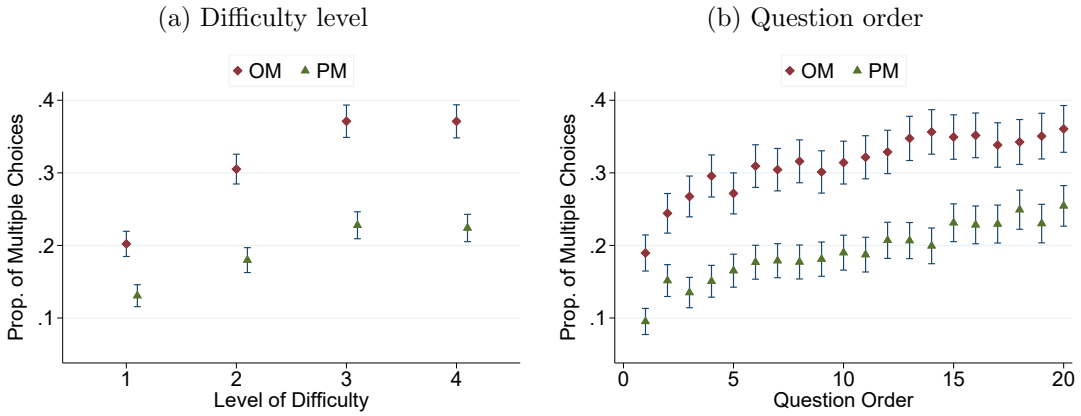
Notes: The dependent variable is the subject's proportion of questions answered with mixing. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Subjects in the Control are excluded from the analysis because they cannot choose to mix. Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

point practice is associated with a 3.8 percentage-point reduction of mixing. Those who work on the practice questions for a longer time mix more frequently. These results suggest that subjects with higher cognitive ability are less likely to mix.

The treatment difference and the effects of characteristics are corroborated when we look at the data at question level. This also allows us to examine the impact of question characteristics such as difficulty level and question order on mixing behavior (Table A1). Recall that questions fall into four difficulty levels determined by the number of changes among the shapes. As shown in Figure 3(a), subjects take more mixing on harder questions compared to easier questions; the patterns are similar for both treatments. Moreover, there is more mixing when subjects get to later questions when their test time is running out, as displayed in Figure 3(b). These results, along with the finding of nega-

tive association between practice score and mixing, are in line with the findings in the literature that a higher choice difficulty induces more mixing (e.g., Agranov and Ortoleva 2017).

Figure 3: Proportion of questions answered with mixing by question characteristics



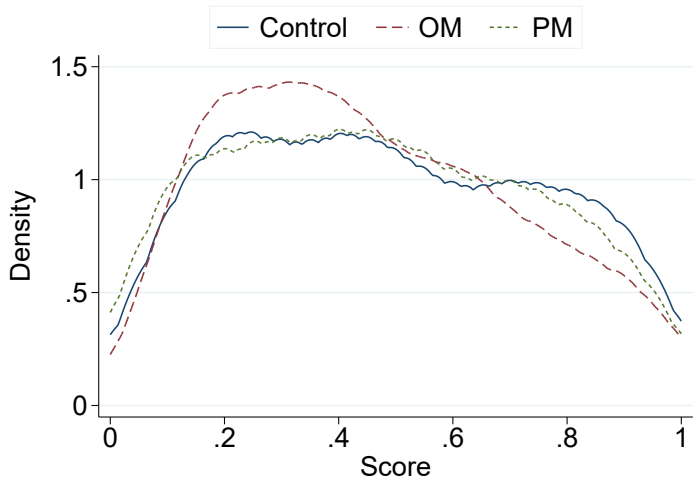
Notes: Brackets indicate 95% confidence intervals with the standard errors clustered at subject level.

4.2 Mixing and IQ score

Next we investigate the effect of mixing on IQ test score. Figure 4 displays the kernel density plot of the IQ score measured as proportion of the total score of 20. Subjects in the OM treatment have significantly lower scores than those in the Control as their scores are more concentrated at the lower range and less so at the higher range. The average score is 0.459 in the OM treatment and 0.488 in the Control ($p = 0.017$, Wilcoxon rank-sum test). Subjects in the PM treatment have an average score of 0.472 which is also slightly lower than those in the Control, but the difference is smaller and not statistically significant ($p = 0.183$).

Table 2 displays the difference in IQ scores due to the treatment. For easier interpretation, we present the results in z-scores. To calculate this, we subtract the overall sample mean from each individual's score and then divide by the sample standard deviation. Note that one SD is equivalent to 0.263 of the full score. Column (1) reports the raw estimates, column (2) includes ability controls and column (3) controls for both ability and demographic characteristics. Subjects in the OM treatment have a lower score compared to those in the Control; the difference is estimated to be 0.11 SD and is statistically

Figure 4: Distribution of IQ score by condition



significant. Those in the PM treatment also tend to have a lower score than those in the Control, and the difference is equal to 0.07 SD and is marginally significant when we control for the ability measures and when we further add the demographic variables. To sum, we have the second observation as follows.

OBSERVATION 2. *When subjects are allowed to mix, they score lower in the IQ test.*

We observe that the 20-minute time constraint which has a large and significant effect on the score, lowering the score by 0.33 SD. We further check the specification which interacts the time constraint with the treatment dummies; the interaction terms are small and insignificant, indicating that the treatment effects on scores are not sensitive to the time constraint (Table A2). One point increase in the practice score is associated with a 0.29 SD increase of the actual score. Subjects who spend longer time practicing have significantly higher score.

A robustness check is provided by the question-level analysis of the IQ scores. We observe that the scores are lower for harder questions, and the treatment difference is always in the same direction for all difficulty levels, although it is more pronounced for harder questions (Figure A1). There is a hump-shaped relationship between score and question order, and the treatment differences appear to be driven by the later questions (Figure A2). The regression analysis gives very similar results in terms of treatment

Table 2: Determinants of standardized IQ score

	(1)	(2)	(3)
OM	-0.112^{**}	-0.115^{***}	-0.113^{***}
	(0.044)	(0.040)	(0.039)
PM	-0.073	-0.072^{*}	-0.068^{*}
	(0.044)	(0.039)	(0.039)
Time constraint (20 min)	-0.385^{***}	-0.332^{***}	-0.327^{***}
	(0.036)	(0.032)	(0.032)
Practice score		0.302^{***}	0.294^{***}
		(0.013)	(0.013)
Logged practice time		0.077^{**}	0.099^{***}
		(0.033)	(0.033)
Mistakes in control q.		-0.186^{***}	-0.191^{***}
		(0.019)	(0.019)
Female			-0.144^{***}
			(0.035)
Age			-0.014
			(0.019)
Second university			-0.076^{**}
			(0.038)
Constant	0.254^{***}	-1.037^{***}	-0.886^{**}
	(0.038)	(0.197)	(0.387)
Observations	2986	2986	2980
R-squared	0.039	0.234	0.241

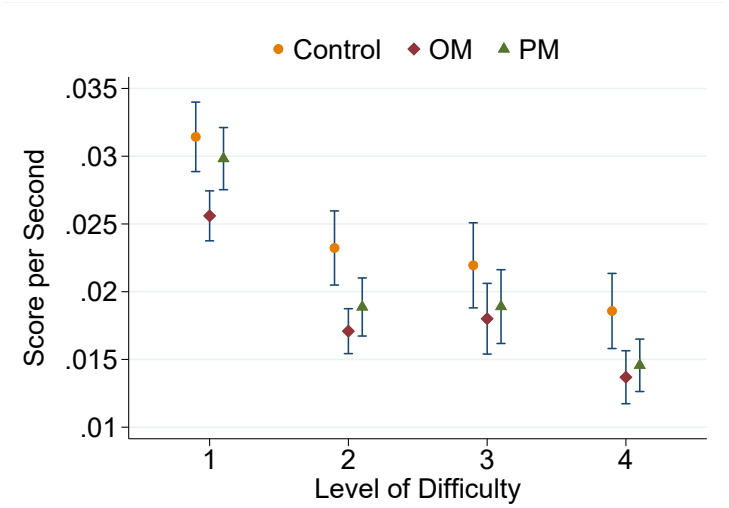
Notes: The dependent variable is IQ score standardized over the full sample. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

differences, effects of ability measures and demographic characteristics (Table A3), which suggests that controlling for the question characteristics does not affect our results.

Mixing and Decision Time. As we discussed earlier, offering mixing options may affect scores in two ways. First, mixing may decrease the expected scores of test-takers as the subjects are better off by choosing the answer that is most likely to be correct. Second, mixing may increase the scores as it might help improve the time allocation by preventing subjects from spending too much time on hard question. The observed effect on scores on average supports the former but not the latter. To further examine the effect on time allocation, we examine the return on time investment (ROTI) at question level, measured by the question score (maximum 1) divided by the time spent on the question (measured in seconds). Should mixing help improve time allocation and prevent

subjects from hanging up too long on some specific questions, ROTI would be higher for harder questions in PM and OM. Figure 5 displays ROTI by question difficulty level for each condition. We find that ROTI is lower in PM and OM, and the gaps between conditions does not shrink or reverse for harder questions. In this regard, we do not observe differences in time allocation across conditions. This observation is robust if we split the sample by the 20-minute and 40-minute time constraints (Figure A3), and corroborated by regression analysis controlling for question order and subject fixed effects (Table A4).

Figure 5: Return on time investment by condition



Notes: Brackets indicate 95% confidence intervals with the standard errors clustered at subject level.

4.3 Differential treatment effects

In our earlier discussion in Section 2, we predict that outcome mixing may be preferred by risk-averse individuals, and probability mixing may be preferred by the individuals who are to hedge away ambiguity through randomization. Individuals also tend to mix options when they have a false sense of diversification. To investigate the underlying mechanisms of mixing behavior, after the IQ test, we measured the participants' risk attitude and proneness to false diversification. In this section we test whether subjects of different risk attitude and proneness of false diversification are affected by the treatments differently, and we also explore whether personality traits predict how subjects are affected by the treatments.

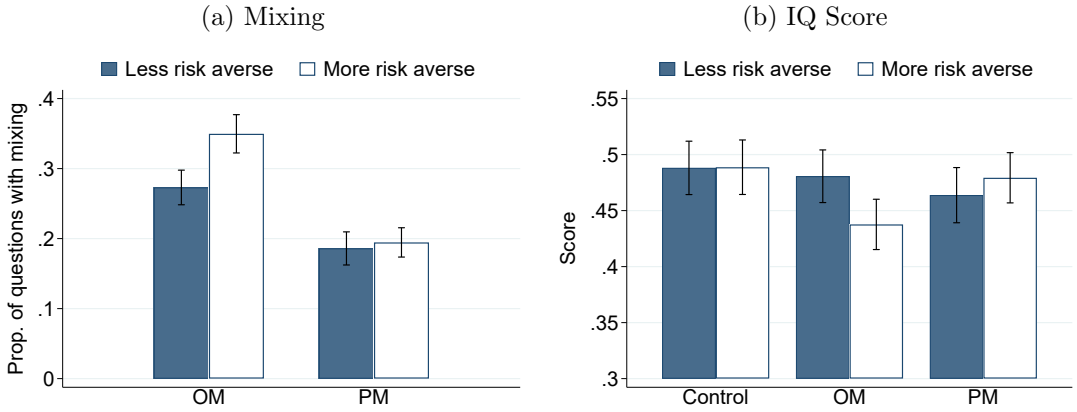
Before the analysis, we conduct a test following DellaVigna et al. (2022) to find out whether the treatment conditions have any spillover effects on the responses to the measures after the IQ test. The idea is that suppose there are spillover effects of the treatment manipulation, which give rise to different responses in the questionnaires, we expect the characteristics to be unbalanced across treatments. The summary statistics of the characteristics and the balance checks are reported in Table A5. In general we do not see any evidence of spillover effects, as risk preference, choices in the false diversification questions and responses to personality scales are evenly distributed across treatments. The only characteristic that appears to take slightly lower value in the OM treatment is the maximization scale, but it is only marginally significant ($p = 0.087$).

We first analyze the differential treatment effects with respect to risk attitudes, as we hypothesize in the theory section that risk-averse individuals are more likely to choose to mix and have lower scores in OM, compared to PM. To visualize the treatment effects by risk attitude, we split the sample by whether the subject is more risk averse than the median subject, and compare the treatment effects in the two subsamples. As shown in Figure 6, more risk averse subjects mix significantly more (panel (a)) and score significantly lower (panel (b)), compared those less risk averse subjects in the OM treatment. This is consistent with our prediction that the treatment difference in the amount of mixing is larger among the subjects who are more risk averse. By contrast, risk attitude has no effect on the mixing behavior for subjects in the PM treatment, and has no effect on the scores in both PM and the Control.

The results are corroborated by regression analysis reported in Table 3, which provides the raw estimates, the augmented estimates with ability controls, and the further augmented estimates with the false diversification, personality measures, and demographic characteristics. Columns (1)–(3) display the treatment difference on mixing by comparing the PM treatment to OM, columns (4)–(6) display the treatment difference on IQ score by separately comparing both OM and PM to the Control.

The predictions regarding mixing and IQ score are both confirmed by the regression results. Risk aversion significantly predicts mixing in the OM treatment but not in PM

Figure 6: Score and mixing by treatment and level of risk aversion



Notes: Subjects who are above-median risk averse are classified as "more risk averse", and the rest as "less risk averse". The brackets indicate 95% confidence intervals.

– subjects who are more risk-averse are estimated to mix 6.2 percentage points more frequently in OM than those who are less risk-averse and the difference is close to zero in PM. The interaction between risk aversion and PM is always highly significant and similar before and after controlling the ability and other characteristics such as false diversification and personality traits. The scores of risk averse subjects are lower if they are in OM, compared to if they are in the Control. The effect is 0.17 SD and marginally significant with the raw estimate, and increases to 0.22 SD and becomes highly significant when we include the ability controls and other characteristics in the estimation. Risk-averse subjects also score significantly lower in OM compared to PM (which is similar to the Control), and the difference is significant across specifications.

OBSERVATION 3A. *Risk averse individuals are more likely to mix and score lower in OM treatment but not in PM treatment.*

Next we examine whether our measure of false diversification affects mixing and IQ score. As explained in the theory section, false diversification can lead to mixing in both OM and PM, and result in lower IQ scores. Recall that we have two measures of the subjects' tendency to take false diversification: the Gate problem and the Card problem. Because the former features a one-shot choice whereas the latter a sequence of choices, we examine the two separately. Only 5.4 percent of the subjects maximize the winning probability for the Gate problem, and the number is 46.7 percent for the Card problem;

Table 3: Treatment effects conditional on risk attitude

Dependent Variable:	Proportion of mixing OM and PM only			IQ score Pooled sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Risk averse	0.075^{***}	0.071^{***}	0.062^{***}	0.024	0.063	0.100^{*}
	(0.018)	(0.018)	(0.018)	(0.063)	(0.056)	(0.057)
Risk averse × OM				-0.170^{*}	-0.198^{**}	-0.222^{***}
				(0.088)	(0.079)	(0.077)
Risk averse × PM	-0.066^{***}	-0.062^{***}	-0.065^{***}	0.036	−0.040	−0.051
	(0.024)	(0.024)	(0.024)	(0.089)	(0.079)	(0.078)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Ability controls	No	Yes	Yes	No	Yes	Yes
Other characteristics	No	No	Yes	No	No	Yes
OM vs. PM p-value†	0.006	0.009	0.007	0.018	0.042	0.026
Observations	1989	1989	1986	2986	2986	2980
R-squared	0.063	0.094	0.114	0.041	0.236	0.264

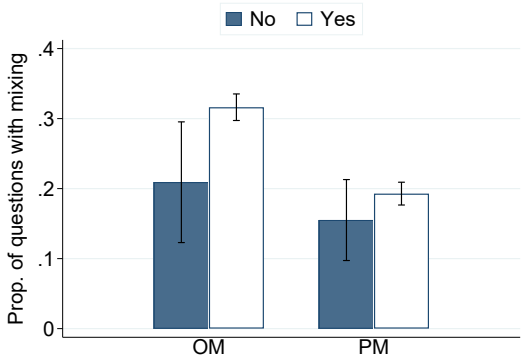
Notes: "Risk averse" is a dummy indicator for being more risk-averse than the median subject. "Ability controls" include practice score, logged practice time, number of mistakes in the comprehension questions. "Other characteristics" include gender, age, university, false diversification choices on Gate problem and Card problem, the Big Five personality traits, and the maximization and regret scales. †Tests of whether the effect of risk aversion differs in OM and PM, i.e., the p-value for the coefficient of "Risk averse × PM" in columns (1)–(3), and for a comparison between the coefficients of "Risk averse × OM" and "Risk averse × PM" in columns (4)–(6). Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

probability matching frequently characterizes other responses in both problems.

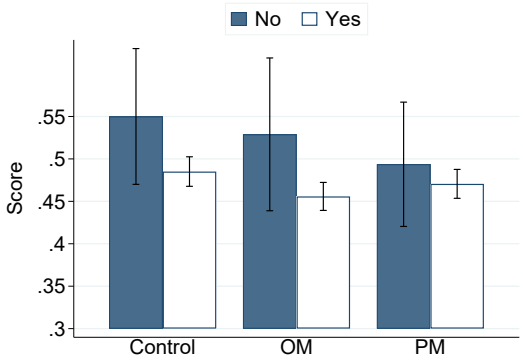
We split the sample by whether the subject's response to each of the false diversification questions maximizes the winning probability, and examine the subsamples for Gate problem and those for Card problem in Figure 7. Panels (a) and (c) suggest that responses to the false-diversification questions weakly predicts mixing: those who maximize the winning probability are less likely to mix than those who take false diversification, yet the difference tends to be small and is only significant in the OM treatment for the Gate problem. For IQ score, as shown in panels (b) and (d), subjects who are more likely to falsely diversify, especially in the Card problem, on average score lower in all three conditions. There are two possible reasons. First, subjects with a stronger sense of false diversification may inherently have lower cognitive ability and hence score lower. Second, subjects with a stronger sense of false diversification are more likely to guess randomly in both treatments and control, resulting in lower scores in all three conditions.

Figure 7: Score and mixing by treatment and false diversification

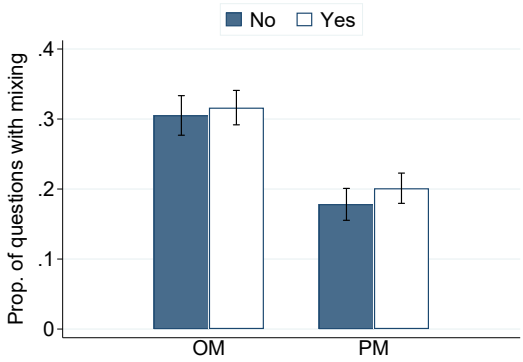
(a) Mixing: Gate problem



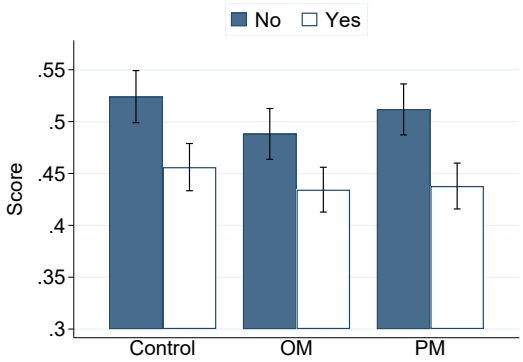
(b) IQ Score: Gate problem



(c) Mixing: Card problem



(d) IQ Score: Card problem



Notes: Subjects who do not maximize the winning probability are classified as showing a tendency to take false diversification (i.e., "Yes"). The brackets indicate 95% confidence intervals.

Table 4: Treatment effects conditional on false diversification

(a) Gate Problem

Dependent Variable:	Proportion of mixing OM and PM only			IQ score Pooled sample		
	(1)	(2)	(3)	(4)	(5)	(6)
FD	0.104**	0.092**	0.078*	−0.262*	−0.220	−0.176
	(0.043)	(0.042)	(0.041)	(0.151)	(0.138)	(0.133)
FD × OM				0.013	0.074	0.079
				(0.224)	(0.204)	(0.199)
FD × PM	−0.067	−0.062	−0.048	0.184	0.186	0.166
	(0.052)	(0.051)	(0.051)	(0.206)	(0.189)	(0.187)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Ability controls	No	Yes	Yes	No	Yes	Yes
Other characteristics	No	No	Yes	No	No	Yes
OM vs. PM p-value†	0.197	0.231	0.346	0.429	0.571	0.660
Observations	1989	1989	1986	2986	2986	2980
R-squared	0.058	0.089	0.110	0.041	0.235	0.259

(b) Card Problem

Dependent Variable:	Proportion of mixing OM and PM only			IQ score Pooled sample		
	(1)	(2)	(3)	(4)	(5)	(6)
FD	0.010	0.012	−0.000	−0.252***	−0.174***	−0.128**
	(0.019)	(0.018)	(0.018)	(0.063)	(0.056)	(0.057)
FD × OM				0.055	−0.029	−0.021
				(0.088)	(0.079)	(0.078)
FD × PM	0.012	0.000	0.001	−0.022	0.010	0.004
	(0.024)	(0.024)	(0.024)	(0.088)	(0.079)	(0.079)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Ability controls	No	Yes	Yes	No	Yes	Yes
Other characteristics	No	No	Yes	No	No	Yes
OM vs. PM p-value†	0.628	0.984	0.953	0.381	0.620	0.742
Observations	1989	1989	1986	2986	2986	2980
R-squared	0.056	0.087	0.108	0.054	0.242	0.263

Notes: "FD" is a dummy indicator for subjects who take false diversification (i.e. deviate from the choice which maximizes the expected value) in the Gate problem for panel (a), and the Card problem for panel (b). "Ability controls" include practice score, logged practice time, number of mistakes in the comprehension questions. "Other characteristics" include gender, age, university, risk aversion, the Big Five personality traits, and the maximization and regret scales. †Tests of whether the effect of FD differs in OM and PM (i.e., the p-value for the coefficient of "FD × PM" in columns (1)–(3), and for a comparison between the coefficients of "FD × OM" and "FD × PM" in columns (4)–(6)). Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

The corresponding regression analysis is reported in Table 4, including the raw estimates and the estimates with controls.[9] Consistent with the patterns displayed in Figure 7, there is only limited evidence that our measures of false diversification ("FD" in the table) predicts hedging behavior. Columns (1)–(3) of panel (a) shows that False diversification measured by the Gate problem is associated with a 10 percentage-point increase of mixing in the OM treatment, which drops to 8 percentage points when ability measures are controlled for, but is still marginally significant; false diversification has a much smaller and insignificant effect in PM, although the difference between the two treatments is not significant. Panel (b) shows that false diversification measured by the Card problem has little predictive power for mixing in either of the two treatments. The observed difference between the Gate problem and Card problem could be due to the fact that the Gate problem reflects mixing in one-shot decisions and the Card problem captures mixing in repeated decisions.

On the IQ score, the estimates confirm that subjects with a false diversification tendency perform worse than those without: for both the Gate and the Card problems, in the Control, the raw estimates point to a difference of over 0.25 SD in column (4); the difference drops considerably in size when we control for ability measures in columns (5) and (6) and is only significant for the Card problem, suggesting that part of the difference is attributed to negative correlations between false diversification and ability measures. Subjects showing a false diversification tendency do not score significantly lower in either OM or PM compared to their counterparts in the Control, and the same finding applies to all specifications.

OBSERVATION 3B. *Individuals who are prone to false diversification score lower in all three conditions.*

Finally, we conduct an additional analysis to examine whether subjects of different personality traits are affected by the treatments differently. We regress mixing and IQ

[9]For a robustness check, we also characterize the degree of false diversification by the departure from maximizing the winning probability normalized by its range, which gives a continuous measure taking the value of one if the choice minimizes the winning probability and zero the choice maximizes the winning probability. The results are similar.

score on each of the personality traits (the Big Five, maximization and regret), the treatment dummies and the interactions between the personality trait and treatment dummies, and control for ability measures, the demographic characteristics, and risk attitude and false diversification measures. The results are reported in Table A6. Other personality traits also generally have little predictive power for mixing; the only exception is that openness to experience negatively correlates with mixing in OM but not in PM. Only conscientiousness exhibits a correlation with IQ score which is significantly different across treatments: more conscientious subjects score significantly lower in the Control; in contrast, the correlation between conscientiousness and score is halved in OM and close to zero in PM.

4.4 Implications for MCQ design

This section reports two additional tests on properties of the scoring rules to investigate whether it is desirable to incorporate mixing into the MCQ design. First, we test whether allowing for mixing affects the gender difference in performance. Second, we summarize and compare the psychometric qualities of the three scoring rules.

It has been shown that the scoring rule with a penalty for incorrect answers can adversely affect gender equality (Baldiga 2014, Coffman and Klinowski 2020, Iriberri and Rey-Biel 2021, Karle et al. 2022). We analyze whether the gender differences in terms of mixing and IQ score change across treatment and report the results in Table 5. In OM, male subjects mix 9.0 percentage points less than females; the difference shrinks only slightly when we control for ability measures and characteristics including risk attitudes, indicating that this gender difference is not fully attributed to the difference in risk attitudes. In PM, there is no gender difference in mixing. In all treatments, male subjects score higher than females. In the Control, the raw difference is 0.19 SD and shrinks to 0.11 SD when we control for ability measures and characteristics. Neither the OM treatment nor the PM treatment significantly affect the gender difference in IQ score in all specifications. Overall, these results suggest that neither PM nor OM biases a particular gender.

Table 5: Treatment effects conditional on gender

Dependent Variable:	Proportion of mixing OM and PM only			IQ score Pooled sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Male	−0.090***	−0.085***	−0.074***	0.185***	0.158***	0.113*
	(0.018)	(0.018)	(0.019)	(0.064)	(0.057)	(0.060)
Male × OM				0.026	0.053	0.071
				(0.089)	(0.080)	(0.079)
Male × PM	0.061**	0.061**	0.062***	−0.056	−0.024	−0.005
	(0.024)	(0.024)	(0.024)	(0.091)	(0.081)	(0.080)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Ability controls	No	Yes	Yes	No	Yes	Yes
Other characteristics	No	No	Yes	No	No	Yes
OM vs. PM p-value†	0.013	0.011	0.010	0.360	0.334	0.331
Observations	1989	1989	1986	2986	2986	2980
R-squared	0.069	0.098	0.113	0.047	0.241	0.263

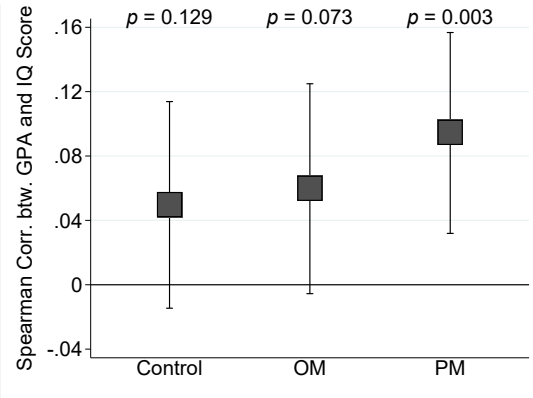
Notes: "Ability controls" include practice score, logged practice time, number of mistakes in the comprehension questions. "Other characteristics" include age, university, risk aversion, false diversification choices on Gate problem and Card problem, the Big Five personality traits, and the maximization and regret scales. †Tests of whether the coefficient of Male differs in OM and PM (i.e., the p-value for the coefficient of "Male × PM" in columns (1)–(3), and for a comparison between the coefficients of "Male × OM" and "Male × PM" in columns (4)–(6)).

Next we examine the psychometric quality of each scoring rule by summarizing the Cronbach's alpha and correlating the IQ scores with the subject's real academic performance. The Cronbach's alpha is a commonly used statistic for estimating psychometric test reliability. It captures the internal consistency of the question items used in measuring a latent variable, such as intelligence. It is a function of average inter-item covariances and the total variance, and ranges from 0 to 1. A high Cronbach's alpha indicates that the set of question items have large average covariances, and are probably measuring the same underlying construct. A reliability coefficient of 0.70 or higher is considered "acceptable" (Lance et al. 2006). We calculate the Cronbach's alpha using the scores of the 20 questions items. [10] We find that the Cronbach's alpha is 0.87 (95% CI [0.86, 0.88]) in the Control, 0.88 (95% CI [0.87, 0.90]) in OM, and 0.87 (95% CI [0.86, 0.88]) in PM, suggesting that all three scoring rules has achieved acceptable reliability.

[10]More specifically, Cronbach's alphas for each condition are calculated as follows: under a scoring rule, let $s_1, s_2, ..., s_{20}$ denote the scores for question item 1, 2, ..., and 20. Let σ_{ij} denote the inter-item covariance between s_i and s_j, and σ_i^2 $(= \sigma_{ii})$ denote the variance of the item s_i. Let $V = \sum_{i=1}^{20}\sum_{j=1}^{20}\sigma_{ij} = \sum_{i=1}^{20}\sigma_i^2 + \sum_{i=1}^{20}\sum_{j\neq i}^{20}\sigma_{ij}$, representing the total variance consisting of both item variances and inter-item covariance. Cronbach's alpha is calculated as $\alpha = \frac{20}{20-1}\cdot(1 - \frac{\sum_{i=1}^{20}\sigma_i^2}{V})$.

We then examine whether correlations between the IQ score and academic performance may differ across scoring rules. It is natural to expect the IQ score and academic performance to be positively correlated – students with higher cognitive ability are more likely to have better academic performance. This analysis is restricted to the 93.6 percent of subjects who gave consent for us to anonymously link the IQ test score to their academic performance measured by the grade point average (GPA).[11] We standardize the GPA within the faculty so that it measures the subject's relative standing within their faculty (a summary of raw GPA by faculty is provided in Figure A4).

Figure 8: Spearman Correlation between GPA and IQ score



Notes: The brackets indicate 95% confidence intervals.

The Spearman correlations between the GPA and IQ score are displayed in Figure 8. We find that the correlation is significant for subjects in the PM treatment (corr. $=$ 0.086, $p = 0.003$), and it is smaller and marginally significant in OM (corr. $= 0.060$, $p = 0.073$) and insignificant in the Control (corr. $= 0.050$, $p = 0.129$). The observed low correlations can be due to the possibility that our subjects are relatively homogeneous in their cognitive ability because they are from universities of similar ranks. While the confidence intervals largely overlap, the observed correlation appears to be smaller in the Control, which could be due to higher frequency of noisy responses.[12]

[11] The consent was collected before the session, and hence as expected, the subjects who drop out from the analysis are evenly distributed across the conditions ($p = 0.653$, chi-squared test). Our main experimental results are qualitatively similar and slightly strengthened when we exclude those subjects – see Table A7 and A8 for replications of Table 1 and 2.

[12] It is also possible that some common factors underlying academic performance may play a more important role in some scoring rules than others. IQ scores and academic performance are widely used as measures of cognitive ability. However, academic performance is also correlated with other non-cognitive abilities. For example, Borghans et al. (2016) show that while academic performance and IQ are correlated, academic performance is a better predictors of important life outcomes than IQ scores.

OBSERVATION 4. *The three scoring rules all exhibit desirable psychometric qualities; the IQ score significantly correlates with academic performance in PM but not the other two conditions.*

5 Concluding Remarks

This study proposes two scoring rules for MCQ based on preferences for hedging, and examine them in the setting of an IQ test using a randomised experiment. We show that the options of outcome and probability mixing are popularly chosen. Participants score lower in both OM and PM treatments, compared to the control condition. Moreover, risk-averse participants are more likely to choose the mixing option and also score lower in OM. Participants with a tendency for false diversification are more likely to mix in both treatments, and they score lower in all the three conditions. These observations shed light on the intertwinement between risk preferences and cognitive ability.

While it remains an open question how to make a best choice among different scoring rules in an MCQ test, our study brings the attention to two theoretically motivated scoring rules—OM and PM as potential alternatives to the standard MCQ scoring rule. Both OM and PM exhibit similarly desirable psychometric properties compared to the standard scoring rule. If we use the IQ scores to predict academic achievement, the score in PM has highest predictive power among the three. These observations are of interest for the interpretation of test scores and policy making in relation to the design of multiple-choice tests.

There are several questions we cannot fully address in this paper and may be avenues for future work. First, subjects in all three conditions may guess randomly in their mind, but we cannot observe such behavior. In this regard, PM is about preferences for hedging with an external and observable randomization device. A within-subject design can allow us to compare the choice behavior for each question across conditions; and a fuller elicitation of the subject's knowledge, e.g., by asking them to assign a probability to each option can be useful for detecting internal randomization. Second, we do not include

a condition with negative marking (Baldiga 2014, Coffman and Klinowski 2020), which is closely related to OM. In one form of negative marking, test-takers receive one point for right answers, a negative fractional point for wrong answers and zero for skipping the questions. Correspondingly, in OM, test-takers receive one point for right answers, zero point for wrong answers and $1/n$ for choosing all n options. Besides the gain-loss difference, test-takers in OM can decide the number of options to mix. Third, we do not measure ambiguity attitude in this study mainly due to time constraint – the IQ test itself already takes up to 20 or 40 minutes depending on the conditions, and the elicitation of ambiguity is generally time-consuming. Moreover, it can be difficult to measure attitudes toward ambiguity and randomization at the same time, as Baillon et al. (2022*b*) observe that measuring ambiguity-averse preference can be challenging when subjects can use random incentive system as a randomization device to hedge away ambiguity. It would be of interest for future work to systematically examine preferences for hedging under different frames, more refined rules and better measures of individual preferences for hedging.

Last, when decision makers face difficult questions, some research suggests that allowing for mixing can help decision making (Levitt 2020, Zhang and Zhong 2020). However, here we do not find evidence that mixing could improve the efficiency in time usage and total test scores in the MCQ test. We consider the following possible explanations. First, the negative impact of hedging on scores might countervail the benefit of using a randomization device as a tie-breaker. Second, the subjects face only 8 options in the control condition. When subjects are allowed to mix any options they wanted, the choice problem becomes which subset of the options to mix. As a result, the number of possible options explodes into over a hundred. This additional complexity can aggravate rather than alleviating the choice difficult problems. Future research can investigate the issue and identify when and how offering a randomization device can improve decision making.

References

Agranov, M., Healy, P. J. and Nielsen, K. (2023), 'Stable randomisation', *The Economic Journal* **133**(655), 2553–2579.

Agranov, M. and Ortoleva, P. (2017), 'Stochastic choice and preferences for randomization', *Journal of Political Economy* **125**(1), 40–68.

Agranov, M. and Ortoleva, P. (2022), Revealed preferences for randomization: An overview, *in* 'AEA Papers and Proceedings', Vol. 112, pp. 426–30.

Agranov, M. and Ortoleva, P. (2023), 'Ranges of randomization', *The Review of Economics and Statistics* pp. 1–44.

Amador-Hidalgo, L., Brañas-Garza, P., Espín, A. M., García-Muñoz, T. and Hernández-Román, A. (2021), 'Cognitive abilities and risk-taking: Errors, not preferences', *European Economic Review* **134**, 103694.

Andersson, O., Holm, H. J., Tyran, J.-R. and Wengström, E. (2016), 'Risk aversion relates to cognitive ability: Preferences or noise?', *Journal of the European Economic Association* **14**(5), 1129–1154.

Andersson, O., Holm, H. J., Tyran, J.-R. and Wengström, E. (2020), 'Robust inference in risk elicitation tasks', *Journal of Risk and Uncertainty* **61**, 195–209.

Aoyama, T. and Hanaki, N. (2021), 'Preference for randomization and validity of random incentive system under ambiguity: An experiment', *ISER DP* (1140).

Bade, S. (2015), 'Randomization devices and the elicitation of ambiguity-averse preferences', *Journal of Economic Theory* **159**, 221–235.

Baillon, A., Halevy, Y. and Li, C. (2022*a*), 'Experimental elicitation of ambiguity attitude using the random incentive system', *Experimental Economics* **25**, 1002–1023.

Baillon, A., Halevy, Y. and Li, C. (2022*b*), 'Randomize at your own risk: on the observability of ambiguity aversion', *Econometrica* **90**(3).

Baldiga, K. (2014), 'Gender differences in willingness to guess', *Management Science* **60**(2), 434–448.

Benartzi, S. and Thaler, R. H. (2001), 'Naive diversification strategies in defined contribution saving plans', *American Economic Review* **91**(1), 79–98.

Benjamin, D. J., Brown, S. A. and Shapiro, J. M. (2013), 'Who is 'behavioral'? cognitive ability and anomalous preferences', *Journal of the European Economic Association* **11**(6), 1231–1255.

Blum, D. and Holling, H. (2018), 'Automatic generation of figural analogies with the IMak package', *Frontiers in Psychology* **9**(AUG), 1–13.

Blum, D., Holling, H., Galibert, M. S. and Forthmann, B. (2016), 'Task difficulty prediction of figural analogies', *Intelligence* **56**, 72–81.

Borghans, L., Golsteyn, B. H., Heckman, J. J. and Humphries, J. E. (2016), 'What grades and achievement tests measure', *Proceedings of the National Academy of Sciences of the United States of America* **113**(47), 13354–13359.

Borghans, L., Meijers, H. and Ter Weel, B. (2008), 'The role of noncognitive skills in explaining cognitive test scores', *Economic Inquiry* **46**(1), 2–12.

Budescu, D. and Bar-Hillel, M. (1993), 'To guess or not to guess: A decision-theoretic view of formula scoring', *Journal of Educational Measurement* **30**(4), 277–291.

Burks, S. V., Carpenter, J. P., Goette, L. and Rustichini, A. (2009), 'Cognitive skills affect economic preferences, strategic behavior, and job attachment', *Proceedings of the National Academy of Sciences of the United States of America* **106**(19), 7745–7750.

Cerreia-Vioglio, S. (2009), Maxmin expected utility on a subjective state space: Convex preferences under risk. Working Paper.

Cerreia-Vioglio, S., Dillenberger, D. and Ortoleva, P. (2015), 'Cautious expected utility and the certainty effect', *Econometrica* **83**(2), 693–728.

Cerreia-Vioglio, S., Dillenberger, D., Ortoleva, P. and Riella, G. (2019), 'Deliberately stochastic', *American Economic Review* **109**(7), 2425–45.

Cettolin, E. and Riedl, A. (2019), 'Revealed preferences under uncertainty: Incomplete preferences and preferences for randomization', *Journal of Economic Theory* **181**, 547–585.

Chapman, J., Snowberg, E., Wang, S. and Camerer, C. (2018), Loss attitudes in the US population: Evidence from dynamically optimized sequential experimentation (DOSE). National Bureau of Economic Research Working Paper.

Chen, D. L., Schonger, M. and Wickens, C. (2016), 'oTree – An open-source platform for laboratory, online, and field experiments', *Journal of Behavioral and Experimental Finance* **9**, 88–97.

Chew, S. H. and Epstein, L. G. (1988), 'The law of large numbers and the attractiveness of compound gambles', *Journal of Risk and Uncertainty* **1**(1), 125–132.

Chew, S. H., Epstein, L. G. and Segal, U. (1991), 'Mixture symmetry and quadratic utility', *Econometrica: Journal of the Econometric Society* pp. 139–163.

Chew, S. H., Miao, B., Shen, Q. and Zhong, S. (2022), 'Multiple-switching behavior in choice-list elicitation of risk preference', *Journal of Economic Theory* p. 105510.

Coffman, K. B. and Klinowski, D. (2020), 'The impact of penalties for wrong answers on the gender gap in test scores', *Proceedings of the National Academy of Sciences of the United States of America* **117**(16), 8794–8803.

de Finetti, B. (1965), 'Methods for discriminating levels of partial knowledge concerning a test item', *British Journal of Mathematical and Statistical Psychology* **18**(1), 87–123.

Dean, M. and Ortoleva, P. (2017), 'Allais, ellsberg, and preferences for hedging', *Theoretical Economics* **12**(1), 377–424.

Debreu, G. (1959), *Theory of value: An axiomatic analysis of economic equilibrium*, number 17, Yale University Press.

DellaVigna, S., List, J. A., Malmendier, U. and Rao, G. (2022), 'Estimating social preferences and gift exchange at work', *American Economic Review* **112**(3), 1038–1074.

Diab, D. L., Gillespie, M. A. and Highhouse, S. (2008), 'Are maximizers really unhappy? The measurement of maximizing tendency', *Judgment and Decision Making* **3**(5), 364–370.

Dohmen, T., Falk, A., Huffman, D. and Sunde, U. (2010), 'Are risk aversion and impatience related to cognitive ability?', *American Economic Review* **100**(3), 1238–1260.

Dohmen, T., Falk, A., Huffman, D. and Sunde, U. (2018), 'On the relationship between cognitive ability and risk preference', *Journal of Economic Perspectives* **32**(2), 115–134.

Dominiak, A. and Schnedler, W. (2011), 'Attitudes toward uncertainty and randomization: an experimental study', *Economic Theory* **48**(2), 289–312.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R. and Stouthamer-Loeber, M. (2011), 'Role of test motivation in intelligence testing', *Proceedings of the National Academy of Sciences of the United States of America* **108**(19), 7716–7720.

Dwenger, N., Kübler, D. and Weizsäcker, G. (2018), 'Flipping a coin: Evidence from university applications', *Journal of Public Economics* **167**, 240–250.

Eichberger, J., Grant, S. and Kelsey, D. (2016), 'Randomization and dynamic consistency', *Economic Theory* **62**(3), 547–566.

Eliaz, K. and Fréchette, G. (2008), Don't put all your eggs in one basket!: an experimental study of false diversification. Working Paper.

Embretson, S. E. and Reise, S. P. (2013), *Item response theory*, Psychology Press.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D. and Sunde, U. (2018), 'Global evidence on economic preferences', *The Quarterly Journal of Economics* **133**(4), 1645–1692.

Feldman, P. and Rehbeck, J. (2022), 'Revealing a preference for mixtures: An experimental study of risk', *Quantitative Economics* **13**(2), 761–786.

Frederick, S. (2005), 'Cognitive reflection and decision making', *Journal of Economic Perspectives* **19**(4), 25–42.

Fudenberg, D., Iijima, R. and Strzalecki, T. (2015), 'Stochastic choice and revealed perturbed utility', *Econometrica* **83**(6), 2371–2409.

Gill, D. and Prowse, V. L. (2023), 'Strategic complexity and the value of thinking', *Economic Journal* **133**(650), 761–786.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S. and Xu, Y. (2019), 'Measuring Success in Education: The Role of Effort on the Test Itself', *American Economic Review: Insights* **1**(3), 291–308.

Gneezy, U. and Potters, J. (1997), 'An experiment on risk taking and evaluation periods', *The Quarterly Journal of Economics* **112**(2), 631–645.

Iriberri, N. and Rey-Biel, P. (2021), 'Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment', *European Economic Review* **131**, 103603.

Karle, H., Engelmann, D. and Peitz, M. (2022), 'Student performance and loss aversion', *Scandinavian Journal of Economics* **124**(2), 420–456.

Ke, S. and Zhang, Q. (2020), 'Randomization and ambiguity aversion', *Econometrica* **88**(3), 1159–1195.

Konovalov, A. and Krajbich, I. (2019), 'Revealed strength of preference: Inference from response times', *Judgment and Decision making* **14**(4).

Lance, C. E., Butts, M. M. and Michels, L. C. (2006), 'The sources of four commonly reported cutoff criteria: What did they really say?', *Organizational research methods* **9**(2), 202–220.

Levitt, S. D. (2020), 'Heads or Tails: The Impact of a Coin Toss on Major Life Decisions and Subsequent Happiness', *The Review of Economic Studies* pp. 1–28.

Lilleholt, L. (2019), 'Cognitive ability and risk aversion: A systematic review and meta analysis.', *Judgment & Decision Making* **14**(3).

Maccheroni, F., Marinacci, M. and Rustichini, A. (2006), 'Ambiguity aversion, robustness, and the variational representation of preferences', *Econometrica* **74**(6), 1447–1498.

Machina, M. J. (1985), 'Stochastic choice functions generated from deterministic preferences over lotteries', *The Economic Journal* **95**(379), 575–594.

Oechssler, J., Rau, H. and Roomets, A. (2019), 'Hedging, ambiguity, and the reversal of order axiom', *Games and Economic Behavior* **117**, 380–387.

Rabin, M. and Weizsäcker, G. (2009), 'Narrow bracketing and dominated choices', *American Economic Review* **99**(4), 1508–43.

Raiffa, H. (1961), 'Risk, ambiguity, and the savage axioms: comment', *The Quarterly Journal of Economics* **75**(4), 690–694.

Rammstedt, B. and John, O. P. (2007), 'Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German', *Journal of Research in Personality* **41**(1), 203–212.

Rubinstein, A. (2002), 'Irrational diversification in multiple decision problems', *European Economic Review* **46**(8), 1369–1378.

Rustichini, A. (2015), 'The role of intelligence in economic decision making', *Current Opinion in Behavioral Sciences* **5**, 32–36.

Rustichini, A., DeYoung, C. G., Anderson, J. E. and Burks, S. V. (2016), 'Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation', *Journal of Behavioral and Experimental Economics* **64**, 122–137.

Saito, K. (2015), 'Preferences for flexibility and randomization under uncertainty', *American Economic Review* **105**(3), 1246–71.

Samuelson, P. A. (1963), 'Risk and uncertainty: a fallacy of large numbers', *Scientia* **6**, 1–6.

Schmeidler, D. (1989), 'Subjective probability and expected utility without additivity', *Econometrica: Journal of the Econometric Society* pp. 571–587.

Schwartz, B., Ward, A., Lyubomirsky, S., Monterosso, J., White, K. and Lehman, D. R. (2002), 'Maximizing versus satisficing: Happiness is a matter of choice', *Journal of Personality and Social Psychology* **83**(5), 1178–1197.

Snyderman, M. and Rothman, S. (1987), 'Survey of expert opinion on intelligence and aptitude testing', *American Psychologist* **42**(2), 137.

Sunde, U., Zegners, D. and Strittmatter, A. (2022), 'Speed, quality, and the optimal timing of complex decisions: Field evidence', *arXiv preprint arXiv:2201.10808* .

Taylor, M. P. (2016), 'Are high-ability individuals really more tolerant of risk? a test of the relationship between risk aversion and cognitive ability', *Journal of Behavioral and Experimental Economics* **63**, 136–147.

Thaler, R. H., Tversky, A., Kahneman, D. and Schwartz, A. (1997), 'The effect of myopia and loss aversion on risk taking: An experimental test', *The Quarterly Journal of Economics* **112**(2), 647–661.

Tversky, A. and Kahneman, D. (1981), 'The framing of decisions and the psychology of choice', *Science* **211**(4481), 453–458.

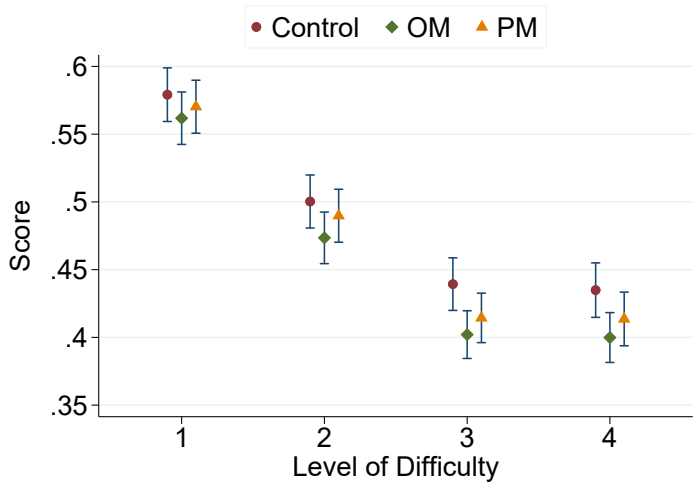
Wakker, P. P. and Yang, J. (2019), 'A powerful tool for analyzing concave/convex utility and weighting functions', *Journal of Economic Theory* **181**, 143–159.

Zapechelnyuk, A. (2015), 'An axiomatization of multiple-choice test scoring', *Economics Letters* **132**, 24–27.

Zhang, X. and Zhong, S. (2020), Putting preference for randomization to work. Working Paper.

Appendix A Additional results

Figure A1: Score by difficulty level



Notes: The brackets indicate 95% confidence intervals with standard errors clustered at subject level.

Figure A2: Score by question order

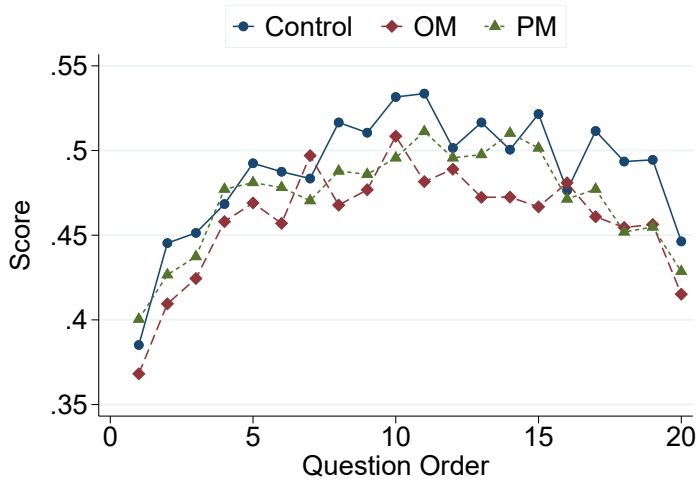
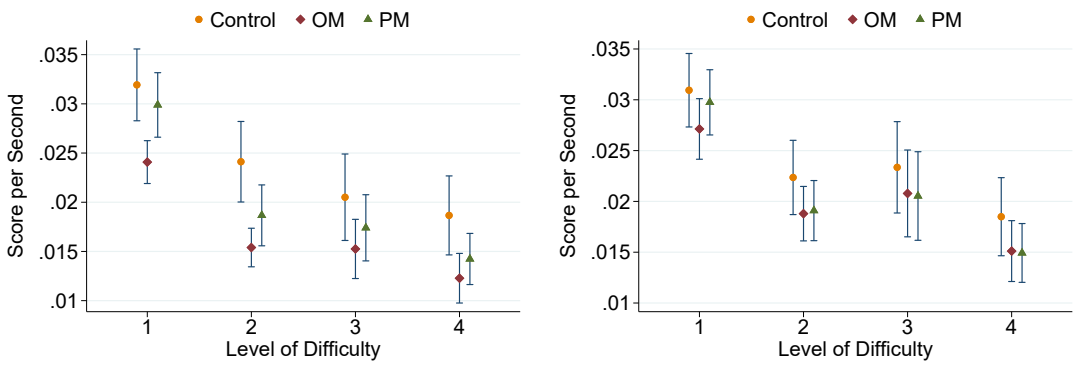


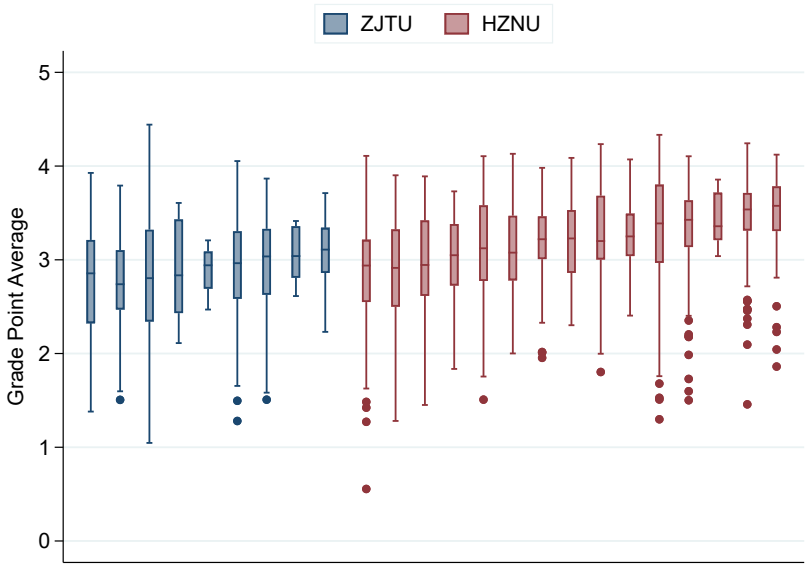
Figure A3: Return on time investment by condition and time constraint

(a) Time Constraint 40 minutes (b) Time Constraint 20 minutes



Notes: Brackets indicate 95 % confidence intervals with the standard errors clustered at subject level.

Figure A4: Summary of raw GPA by university and faculty



Notes: Box plot of the raw 5-point GPA for the participants across different faculties. Within each university the faculties are ordered by the average raw GPA.

Table A1: Question-level analysis of mixing

	(1)	(2)	(3)	(4)
PM	-0.121^{***}	-0.121^{***}	-0.118^{***}	-0.118^{***}
	(0.012)	(0.012)	(0.012)	(0.012)
Time constraint (20 min)	0.049^{***}	0.050^{***}	0.050^{***}	0.048^{***}
	(0.012)	(0.012)	(0.012)	(0.012)
Difficulty level				
Two-Rule		0.075^{***}	0.075^{***}	0.075^{***}
		(0.005)	(0.005)	(0.005)
Three-Rule		0.131^{***}	0.131^{***}	0.131^{***}
		(0.006)	(0.006)	(0.006)
Four-Rule (hardest)		0.129^{***}	0.129^{***}	0.129^{***}
		(0.006)	(0.006)	(0.006)
Question order		0.007^{***}	0.007^{***}	0.007^{***}
		(0.000)	(0.000)	(0.000)
Practice score			-0.040^{***}	-0.038^{***}
			(0.006)	(0.006)
Logged practice time			0.051^{***}	0.046^{***}
			(0.012)	(0.013)
Mistakes in control q.			0.003	0.005
			(0.009)	(0.009)
Male				-0.051^{***}
				(0.013)
Age				0.004^{**}
				(0.002)
Second university				0.013
				(0.013)
Constant	0.288^{***}	0.136^{***}	-0.049	-0.090
	(0.011)	(0.011)	(0.077)	(0.084)
Observations	39079	39079	39079	39079
Clusters	1989	1989	1989	1989
R-squared	0.023	0.046	0.059	0.063

Notes: Responses by subjects in the Control are excluded from the analysis because they cannot choose to mix; a small number of questions (1.7%) which are not answered by the subject due to timeout are also excluded from the sample. The dependent variable is a dummy indicator for whether the subject answered the question with mixing. The difficulty level of the question is measured by the number of "rules", i.e., the number of changes to be identified among the shapes. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Standard errors clustered at subject level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A2: Interaction effects between time constraint and treatments

Dependent Variable:	Proportion of mixing OM and PM only			IQ score Pooled sample		
	(1)	(2)	(3)	(4)	(5)	(6)
OM				−0.131*	−0.132**	−0.126**
				(0.067)	(0.059)	(0.058)
PM	−0.122***	−0.118***	−0.118***	−0.093	−0.092	−0.087
	(0.017)	(0.017)	(0.017)	(0.066)	(0.059)	(0.059)
TC	0.044**	0.044**	0.042**	−0.411***	−0.356***	−0.349***
	(0.019)	(0.018)	(0.018)	(0.064)	(0.057)	(0.057)
TC × OM				0.037	0.033	0.026
				(0.088)	(0.079)	(0.079)
TC × PM	0.004	0.003	0.001	0.041	0.039	0.038
	(0.024)	(0.024)	(0.024)	(0.089)	(0.078)	(0.078)
Ability controls	No	Yes	Yes	No	Yes	Yes
Other characteristics	No	No	Yes	No	No	Yes
Observations	1989	1989	1986	2986	2986	2980
R-squared	0.055	0.086	0.097	0.039	0.234	0.241

Notes: "TC" is a dummy indicator for the 20-minute time constraint. "TC × OM" and "TC × PM" interact the time constraint with the treatment dummies to test whether the treatment effects differ by time constraint. "Ability controls" include practice score, logged practice time, number of mistakes in the comprehension questions. "Other characteristics" include gender, age and university. Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A3: Question-level analysis of IQ score

	(1)	(2)	(3)	(4)
OM	-0.028^{**}	-0.028^{**}	-0.028^{***}	-0.028^{***}
	(0.012)	(0.012)	(0.011)	(0.010)
PM	-0.019	-0.019	-0.019^{*}	-0.019^{*}
	(0.012)	(0.012)	(0.010)	(0.010)
Time constraint (20 min)	-0.089^{***}	-0.089^{***}	-0.072^{***}	-0.071^{***}
	(0.010)	(0.010)	(0.009)	(0.009)
Difficulty level				
Two-Rule		-0.084^{***}	-0.084^{***}	-0.084^{***}
		(0.005)	(0.005)	(0.005)
Three-Rule		-0.155^{***}	-0.154^{***}	-0.154^{***}
		(0.005)	(0.005)	(0.005)
Four-Rule (hardest)		-0.157^{***}	-0.157^{***}	-0.157^{***}
		(0.005)	(0.005)	(0.005)
Question order		0.019^{***}	0.019^{***}	0.019^{***}
		(0.001)	(0.001)	(0.001)
Question order squared		-0.001^{***}	-0.001^{***}	-0.001^{***}
		(0.000)	(0.000)	(0.000)
Practice score			0.081^{***}	0.079^{***}
			(0.004)	(0.004)
Logged practice time			0.032^{***}	0.037^{***}
			(0.009)	(0.009)
Mistakes in control q.			-0.050^{***}	-0.051^{***}
			(0.005)	(0.005)
Male				0.036^{***}
				(0.009)
Age				-0.002^{**}
				(0.001)
Second university				-0.021^{**}
				(0.009)
Constant	0.541^{***}	0.547^{***}	0.133^{**}	0.136^{**}
	(0.010)	(0.011)	(0.053)	(0.055)
Observations	58713	58713	58713	58713
Clusters	2986	2986	2986	2986
R-squared	0.009	0.030	0.093	0.095

Notes: A small number of questions (1.7%) which are not answered by the subject due to timeout are excluded from the sample. The dependent variable is the score from the question. The difficulty level of the question is measured by the number of "rules", i.e., the number of changes to be identified among the shapes. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Standard errors clustered at subject level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A4: Return of time investment by condition

	(1)	(2)	(3)	(4)
OM	-0.005^{***}	-0.006^{***}	-0.006^{***}	
	(0.001)	(0.002)	(0.002)	
PM	-0.003^{***}	-0.002	-0.002	
	(0.001)	(0.002)	(0.002)	
Difficulty level	-0.004^{***}	-0.004^{***}	-0.004^{***}	-0.004^{***}
	(0.000)	(0.001)	(0.001)	(0.001)
Difficulty level × OM		0.000	0.000	0.000
		(0.001)	(0.001)	(0.001)
Difficulty level × PM		-0.001	-0.001	-0.001
		(0.001)	(0.001)	(0.001)
Question order	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Question order squared	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Male			0.006^{***}	
			(0.001)	
Age			0.000	
			(0.000)	
Second university			0.005^{***}	
			(0.001)	
Practice score			-0.002^{***}	
			(0.000)	
Logged practice time			-0.007^{***}	
			(0.001)	
Mistakes in control q.			0.004^{***}	
			(0.001)	
Constant	0.018^{***}	0.018^{***}	0.057^{***}	0.016^{***}
	(0.001)	(0.002)	(0.007)	(0.001)
Subject fixed effects	No	No	No	Yes
Observations	58713	58713	58713	58713
Clusters	2986	2986	2986	2986
R-squared	0.010	0.010	0.019	0.010

Notes: A small number of questions (1.7%) which are not answered by the subject due to timeout are excluded from the sample. The dependent variable is return on time investment (ROTI), measured by the question score (maximum 1) divided by the time spent on the question (measured in seconds). The difficulty level of the question is measured by the number of "rules", i.e., the number of changes to be identified among the shapes. It is treated as a continuous variable and interacted with the two treatment dummies to test whether OM and PM helps the subject improve time allocation for harder questions. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Standard errors clustered at subject level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A5: Summary Statistics and Covariate Balance

	Summary Statistics	Check for selection	
		OM	PM
	(1)	(2)	(3)
Risk aversion	-0.000	-0.004	0.017
	[0.782]	(0.011)	(0.012)
FD in Gate problem	0.946	0.036	-0.052
	[0.225]	(0.037)	(0.040)
FD in Card problem	0.533	0.002	0.012
	[0.499]	(0.017)	(0.018)
Extraversion	0.575	0.036	-0.060
	[0.184]	(0.050)	(0.053)
Agreeableness	0.681	-0.075	0.001
	[0.138]	(0.064)	(0.065)
Conscientiousness	0.568	0.082	-0.096
	[0.144]	(0.065)	(0.065)
Neuroticism	0.594	0.066	-0.006
	[0.145]	(0.065)	(0.067)
Openness to experience	0.748	0.044	-0.042
	[0.156]	(0.057)	(0.060)
Regret	0.661	0.050	0.030
	[0.123]	(0.078)	(0.077)
Maximization	0.635	-0.136*	0.082
	[0.123]	(0.080)	(0.081)
Observations	2986	2986	2986
R-squared		0.003	0.004

Notes: Column (1) reports summary statistics of the preference and personality traits elicited by the questionnaire after the IQ test (mean and standard deviation in the square brackets). In Columns (2) and (3) we regress the dummy indicators of the assignment to the OM and PM treatments on the subject characteristics; the robust standard errors are reported in the parentheses.

Table A6: Effects of Personality Characteristics on Mixing and IQ score

(a) Dependent variable: Mixing

	(1) Extraver.	(2) Agreeable.	(3) Conscient.	(4) Neurotic.	(5) Openness	(6) Maximiz.	(7) Regret
Personality	0.066	0.051	−0.027	0.083	−0.203***	−0.108	−0.102
	(0.052)	(0.064)	(0.062)	(0.064)	(0.059)	(0.075)	(0.080)
Personality × PM	−0.027	−0.064	−0.022	−0.110	0.134*	0.115	0.103
	(0.066)	(0.085)	(0.082)	(0.084)	(0.077)	(0.100)	(0.101)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ability controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1986	1986	1986	1986	1986	1986	1986
R-squared	0.102	0.101	0.101	0.101	0.107	0.101	0.101

(b) Dependent variable: IQ score

	(1) Extraver.	(2) Agreeable.	(3) Conscient.	(4) Neurotic.	(5) Openness	(6) Maximiz.	(7) Regret
Personality	−0.261	−0.067	−0.671***	0.234	0.520***	0.314	0.017
	(0.159)	(0.210)	(0.201)	(0.192)	(0.188)	(0.234)	(0.247)
Personality × OM	0.168	−0.292	0.331	−0.559**	0.236	0.194	0.340
	(0.215)	(0.288)	(0.271)	(0.275)	(0.257)	(0.322)	(0.328)
Personality × PM	0.029	−0.251	0.817***	−0.002	−0.099	−0.099	0.136
	(0.219)	(0.294)	(0.275)	(0.270)	(0.252)	(0.322)	(0.324)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ability controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
OM vs. PM p-value†	0.507	0.885	0.064	0.044	0.172	0.351	0.504
Observations	2980	2980	2980	2980	2980	2980	2980
R-squared	0.250	0.250	0.252	0.250	0.256	0.250	0.249

Notes: Each column estimates the effects of one personality characteristic as indicated above the column number. "Ability controls" include practice score, logged practice time, number of mistakes in the comprehension questions. "Other characteristics" include age, university, risk aversion, and false diversification choices on Gate problem and Card problem. †Tests of whether the coefficient of Personality differs in OM and PM (i.e., the p-value for a comparison between the coefficients of "Personality × OM" and "Personality × PM").

Table A7: Replication of Table 1 excluding subjects with missing administrative data

	(1)	(2)	(3)
PM	-0.127^{***}	-0.124^{***}	-0.124^{***}
	(0.013)	(0.012)	(0.012)
Time constraint (20 min)	0.043^{***}	0.042^{***}	0.041^{***}
	(0.012)	(0.013)	(0.012)
Practice score		-0.045^{***}	-0.043^{***}
		(0.006)	(0.006)
Logged practice time		0.044^{***}	0.038^{***}
		(0.013)	(0.013)
Mistakes in control q.		0.004	0.006
		(0.009)	(0.009)
Male			-0.053^{***}
			(0.014)
Age			0.010
			(0.007)
Second university			0.005
			(0.014)
Constant	0.295^{***}	0.169^{**}	0.033
	(0.012)	(0.078)	(0.150)
Observations	1867	1867	1866
R-squared	0.059	0.096	0.106

Notes: The dependent variable is the subject's proportion of questions answered with mixing. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Subjects in the Control are excluded from the analysis because they cannot choose to mix. Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A8: Replication of Table 2 excluding subjects with missing administrative data

	(1)	(2)	(3)
OM	-0.121^{***}	-0.120^{***}	-0.118^{***}
	(0.045)	(0.041)	(0.041)
PM	-0.085^{*}	-0.078^{*}	-0.075^{*}
	(0.046)	(0.041)	(0.041)
Time constraint (20 min)	-0.384^{***}	-0.330^{***}	-0.326^{***}
	(0.037)	(0.034)	(0.033)
Practice score		0.297^{***}	0.290^{***}
		(0.014)	(0.014)
Logged practice time		0.074^{**}	0.092^{***}
		(0.035)	(0.035)
Mistakes in control q.		-0.183^{***}	-0.188^{***}
		(0.020)	(0.020)
Male			0.134^{***}
			(0.037)
Age			-0.016
			(0.021)
Second university			-0.055
			(0.039)
Constant	0.276^{***}	-0.992^{***}	-0.808^{*}
	(0.039)	(0.209)	(0.419)
Observations	2796	2796	2793
R-squared	0.040	0.225	0.231

Notes: The dependent variable is IQ score standardized over the full sample. "Practice score" is the count of questions the subject answers correctly out of the five practice IQ-test questions. "Logged practice time" is the natural log of the length the subject spends on the practice questions. "Mistakes in control q." is the number of mistakes the subject makes in answering the comprehension questions for the test procedure. "Female" and "Second university" are dummy indicators and "Age" is a continuous measure (in years). Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Appendix B Experimental Instructions

[*] indicates differences across treatments

IQ test

Thank you for participating in this quiz! Every participant only gets one chance to take the quiz. If you exit by closing this page, the existing progress may be lost when you re-open the page.

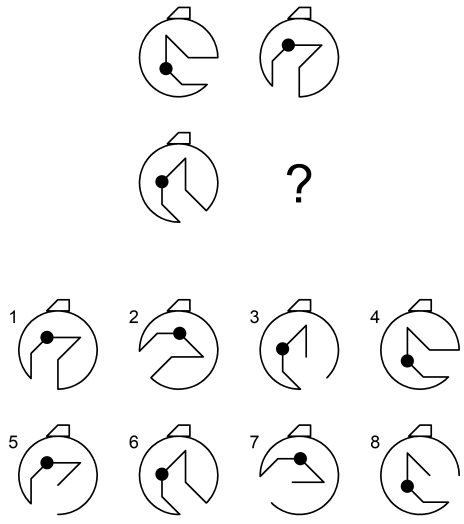
Please read the instruction carefully. Once you click on the "Next" button, you will not be able to go back to the previous screen. If you are ready, please click on "Next" to start.

In the upcoming intelligence quotient (IQ) test, you are asked to work on a quiz with a series of questions.

In each question, a matrix with three shapes is presented (see the examples below). A fourth (bottom-right) shape is missing. Your task is to find the missing shape out of the 8 options presented below the matrix. Each question may involve different types of changes between the shapes in the matrix. You can work either from left to right (thus in the rows) or from top to bottom (thus in the columns). There may be a single change, or multiple changes. It is your task to recognize these changes and apply the same changes to find out the option which fits in the bottom-right place.

For each question, there is only one correct option. Five examples are provided for your reference. Examples 1–3 have only one change, while Example 4 and 5 have multiple changes.

Example 1



Please key in the correct option:

If the option you enter is incorrect, the correct option will be displayed on the top of screen.

[The layout for Example 2–4 is identical and hence omitted.]

[**Control only:**
In the quiz, please indicate your answer by clicking on the option. Notice: You can choose only one option. You will score 1 point if the option you choose is the correct option, and zero otherwise. For example, suppose that Option 5 is the correct answer,

- If you choose Option 5 as your answer, you will receive 1 point. Otherwise, you will receive zero.]

[**Outcome Mixing (OM) treatment only:**
In the quiz, please indicate your answer by clicking on the option. Notice: You can either choose one option, or multiple options:

- If you choose one option, you will score 1 point if it is the correct option, and zero otherwise.
- If you choose more than one options, and if the correct option is among your chosen options, you will receive a fractional score, 1 divided by the number of chosen options, and zero otherwise.

For example, suppose that Option 5 is the correct answer,

- If you only choose one option and it is Option 5, you will receive 1 point. If it is not Option 5, you will receive zero.
- If you choose Options 1, 5, 7 and 8, since the correct option (Options 5) is among the options you choose, and you choose four options, you will score 1 divided by 4, which is 0.25 point.]

[**Probability Mixing (PM) treatment only:**
In the quiz, please indicate your answer by clicking on the option. Notice: You can either choose one option, or multiple options:

- If you choose one option, you will score 1 point if it is the correct option, and zero otherwise.
- If you choose more than one options, the computer will randomly select one option among your chosen options to be your answer. Each of your chosen options will have an equal chance of being selected. You will score 1 point if the randomly selected answer is correct, and zero otherwise.

For example, suppose that Option 5 is the correct answer,

- If you only choose one option and it is Option 5, you will receive 1 point. If it is not Option 5, you will receive zero.

- If you choose Options 1, 5, 7 and 8, the computer will randomly select one of the four options. If Option 5 is selected, you will receive 1 point; if Option 1, 7 or 8 is selected, you will receive zero.]

You will have 20 questions to solve. Your total scores will be the sum of scores for these 20 questions.

You are allowed to attempt each question only once. That is, after you submit the answer to a question, you will not be able to review or revise your answer.

You may find some questions to be easier than the others, but the questions are put in a random order. That is, you will not necessarily encounter a more difficult question when you move on to the next question.

[**Long time constraint only:** Please note that there is an overall time limit of 40 minutes for completing the quiz. You will see a timer on the screen. When the time is up, all of the remaining unsolved questions will give you zero score.]

[**Short time constraint only:** Please note that there is an overall time limit of 20 minutes for completing the quiz. You will see a timer on the screen. When the time is up, all of the remaining unsolved questions will give you zero score.]

When the study is over, 10 participants will be randomly chosen from all participants. Those participants will receive 100RMB fixed bonus. Moreover, the test score will be converted into extra bonus: 15RMB per point. The total score of this test is 20 points. That means the maximum amount of bonus is 400RMB. We will contact the 10 participants by email and transfer the bonus using WeChat Pay.

Before we begin the quiz, please answer the following questions to make sure you understand the rules. Please follow the instructions on your screen.

Comprehension questions

1. How many questions are there in the quiz, and how long do you have to work on the quiz? Please choose:

- 10 questions, 20 minutes
- 10 questions, 40 minutes
- 20 questions, 20 minutes
- 20 questions, 40 minutes

(Correct answer: "20 questions, 20 minutes" for subjects who have the short time constraint, "20 questions, 40 minutes" for subjects who have the long time constraint.
Message if an incorrect answer is given: "There are 20 questions in the quiz and you have 20 (40) minutes to complete the quiz.")

2. On question 1 of the quiz, suppose the correct answer is Option 7. What score will you receive on this question if you choose Option 3?

(Correct answer: 0
Message if incorrect answer is given: "Because the answer you chose is incorrect, you will receive 0 point.)

3. On question 9 of the quiz, suppose the correct answer is Option 4 and you choose Option 4. What score will you receive?

(Correct answer: 1
Message if incorrect answer is given: "Because the answer you chose is correct, you will receive 1 point.")

[**OM treatment only:**

4. On question 12 of the quiz, suppose the correct answer is Option 6 and you choose Options 6 and 8. What score will you receive? Please keep two decimal places for your answer.

(Correct answer: 0.50
Message if incorrect answer is given: "The correct answer is among the two options that you choose, therefore your score will be 1/2, or 0.50 point.")

5. On question 23 of the quiz, suppose the correct answer is Option 2 and you choose Options 1 and 3. What score will you receive?

(Correct answer: 0
Message if incorrect answer is given: "The correct answer is not among the two options that you choose, therefore your score will be 0 point.")]

[**PM treatment only:**

4. On question 12 of the quiz, suppose the correct answer is Option 6, and you choose Options 6 and 8. What score will you receive if the computer randomly selects Option 6 as your answer?

What score will you receive if the computer randomly selects Option 8 as your answer?

(Correct answer: 1, 0
Message if incorrect answer is given: "If the computer selects option 6 as your answer, you will receive 1 point; if the computer selects option 8 as your answer, you will receive 0 point.")

5. On question 23 of the quiz, suppose the correct answer is Option 2 and you choose Options 1 and 3. What score will you receive if the computer randomly selects Option 1 as your answer?

What score will you receive if the computer randomly selects Option 3 as your answer?

(Correct answer: 0, 0
Message if incorrect answer is given: "If the computer selects option 1 as your answer, you will receive 0 point; if the computer selects option 3 as your answer, you will receive 0 point. Because the correct option is not among the options you choose, you will receive 0 regardless of which option is selected by the computer.")]

Please click "Start" to start the IQ test. Once you click on "Start", there will be a timer displayed on the screen for your reference.

Questionnaires

This is the end of the IQ test. Thank you for your participation. Please follow the instruction to finish the questionnaire. It is expected to take no more than 10 minutes. After you finish the questionnaire, your score in IQ test and feedback from questionnaire will be provided for your reference.

Risk - Hypothetical Choices

Please imagine the following situation: you can choose between

- Getting a draw where you would have is 50% chance to get 300RMB and 50% chance to get nothing
- Getting a fixed amount of payment.

You may choose between fixed payment or lucky draw. We will show you 5 different situations.

Situation 1/5

What would you prefer:

- Draw: 50% chance to get 300RMB, 50% chance to get nothing.
- Fixed payment: 160RMB

(The layout for situations 2–5 is identical. The fixed payment x changes according to the answers provided for the previous situations:

- In situation 2, x increases (decreases) by 80RMB from situation 1 if the draw (the fixed amount) is chosen.
- In situation 3, x increases (decreases) by 40RMB from situation 2 if the draw (the fixed amount) is chosen.
- In situation 4, x increases (decreases) by 20RMB from situation 3 if the draw (the fixed amount) is chosen.
- In situation 5, x increases (decreases) by 10RMB from situation 4 if the draw (the fixed amount) is chosen.)

Big Five Inventory

How well do the following statements describe your personality?

1 for Completely disagree, 2 for Strongly disagree, 3 for Disagree, 4 for Neutral, 5 for Agree, 6 for Strongly agree, 7 for Completely agree.

I see myself as someone who ...

(1) ... is reserved
(2) ... is generally trusting
(3) ... tends to be lazy
(4) ... is relaxed, handles stress well
(5) ... has few artistic interests
(6) ... is outgoing, sociable
(7) ... tends to find fault with others
(8) ... does a thorough job
(9) ... gets nervous easily
(10) ... has an active imagination

Maximizer Scale

Read each item carefully. Please select the number that best describes you.

1 for Completely disagree, 2 for Strongly disagree, 3 for Disagree, 4 for Neutral, 5 for Agree, 6 for Strongly agree, 7 for Completely agree.

(1) No matter what it takes, I always try to choose the best thing.
(2) I don't like having to settle for "good enough"
(3) I am a maximizer.
(4) No matter what I do, I have the highest standards for myself.
(5) I will wait for the best option, no matter how long it takes.
(6) I never settle for second best.
(7) I am uncomfortable making decisions before I know all of my options.
(8) Whenever I'm faced with a choice, I try to imagine what all the other possibilities are, even ones that aren't present at the moment.
(9) I never settle.

Regret Scale

Read each item carefully. Please select the number that best describes you.

1 for Completely disagree, 2 for Strongly disagree, 3 for Disagree, 4 for Neutral, 5 for Agree, 6 for Strongly agree, 7 for Completely agree.

(1) Whenever I make a choice, I'm curious about what would have happened if I had chosen differently.
(2) Whenever I make a choice, I try to get information about how the other alternatives turned out.
(3) If I make a choice and it turns out well, I still feel like something of a failure if I find out that another choice would have turned out better.
(4) When I think about how I'm doing in life, I often assess opportunities I have passed up.
(5) Once I make a decision, I don't look back.

Risk - Self-Report

Please tell us, overall, how willing or unwilling you are to take risk?

Please use a 10-point scale to evaluate your preference. 0 stands for completely unwilling to take risk and 10 stands for completely willing to take risk. You can also use any number between 0 and 10, that is, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 to answer.

False Diversification

Please imagine following two scenarios:

1. You are waiting for your friend in a mall. There are four entrances in the mall. According to statistics, the proportions of visitors entering each entrance are as follows:

21% east entrance, 27% west entrance, 32% south entrance, 20% north entrance.

You have no idea which entrance your friend will get in from, so you need to choose one to wait for your friend.

How will you assign the probability for each entrance?

East Entrance _____%
West Entrance _____%
South Entrance _____%
North Entrance _____%

(Message if the four numbers do not add up to 100: "Please check your response. The four numbers you enter should add up to 100 and the probability should not exceed 100%.")

2. Suppose you participate in the following game. There are 100 cards:

   36 green cards, 25 blue cards, 22 yellow cards, 17 brown cards.

   You will randomly draw 5 cards and guess the color for each card. For each card, you will get 10RMB if your guess is right, 0 if your guess is wrong.

   How will you guess the colors of these five cards?

   First Card: Green / Blue / Yellow / Brown
   Second Card: Green / Blue / Yellow / Brown
   Third Card: Green / Blue / Yellow / Brown
   Fourth Card: Green / Blue / Yellow / Brown
   Fifth Card: Green / Blue / Yellow / Brown

IQ Test Experience

Next you will be able to view your score in the IQ test. Do you want to know your relative performance compared to the other participants? Yes/No

If your choice is "Yes" we will email you your relative performance after the study is finished. (For example, your score is higher than X% participants.)

Have you ever taken this test or a similar test before? Yes/No

According to your experience in this test, please answer following question.

In this test, do you think your performance could benefit if you get to some extra time for working on the quiz? Yes/No If so, how many minutes?

[**OM and PM treatments only:**

Did you ever choose multiple options? Yes/No

If so, why did you do it? If not, why didn't you choose it? Please provide a brief explanation:]