Asymmetries in Attribution of Success and Failure

Ruchi Avtar[*]

September 2025

Abstract

This paper studies how individuals interpret outcomes that combine skill and chance, and how these interpretations shape beliefs and decisions. In a controlled experiment, participants solve cognitive tasks and compete in a tournament where outcomes are partly determined by random adjustments. This design allows Bayesian benchmarks for belief updating about both relative ability and luck. I find systematic asymmetries: participants update beliefs about luck in line with Bayes, but distort ability beliefs in self-serving ways, particularly after losses. These biased attributions drive competitive choices, highlighting the role of ego-protection in shaping persistence and mobility.

[*] I am grateful to my advisors Andrew Schotter and Sevgi Yuksel for their guidance and invaluable advice on this project. I am also very thankful for comments and suggestions from Alex Ballyk and Raquel Fernandez. Rishika Ajay Tulysan provided excellent assistance with the experiment. This study was reviewed by the Institutional Review Board at New York University. I gratefully acknowledge the research support provided by Center for Experimental Social Science, New York University.

1 Introduction

How do individuals interpret signals of success and failure, and how do these interpretations shape their subsequent choices? In many economic settings such as applying for jobs or promotions, competing for contracts, educational testing, or entering a tournament, outcomes depend on both internal factors (ability and effort) and external ones (luck, timing etc). After such outcomes, individuals must decide whether to persist, switch strategies, or disengage. These decisions matter: they shape career paths, competitive dynamics, and market efficiency.

A central determinant of these choices is how people attribute their success or failure outcomes. Do they see success as evidence of ability or as the product of luck? Do they see failure as revealing low ability, or simply a bad break? A growing body of work documents that people often misattribute the drivers of their outcomes, leading to distorted beliefs about their own ability and the role of chance (see for example Miller and Ross (1975), Bénabou and Tirole (2002), Eil and Rao (2011)). Such misattributions matter since they can shape and alter self-confidence, generate persistence or exit in competitive environments, and ultimately shape labor market and financial decisions.

This paper studies these attribution processes in a controlled experimental setting. I am interested in studying the asymmetries that arise in choices and belief updating after outcomes that can plausibly be attributed to both skill and chance. To isolate the possible forces at work, I design a lab experiment. A laboratory setting allows me to control the structure of outcomes, measure priors and posteriors precisely, and cleanly separate ability-based and luck-based components of performance. This precision is critical for testing whether people make biased choices, and whether those can be linked to outcome-dependent belief updating errors.

The experiment proceeds in two parts. First, participants solve Raven's matrices (John and Raven (2003)), generating a measure of cognitive ability. Second, outcomes in head-to-head matchups are determined by both own performance and a random adjustment term that is equally likely to help or hinder. Subjects observe whether they win or lose, and then update their beliefs about (i) their relative ability in the absence of any adjustment and (ii) the role of luck through the adjustment term. A neutral ("robot") control condition allows me to separate general inferential errors from ego-relevant motivated reasoning, while treatments vary the range of randomness to test sensitivity to signal informativeness. Finally, choices over whether to re-compete against the same or a new opponent under different adjustment circumstances provide behavioral evidence on how attribution may be shaping decisions.

The results reveal two main findings. First, there is a striking asymmetry in choice preferences. Even though the outcome includes an element of noise, winners overwhelmingly prefer to stick with the same counterpart they just beat, while losers prefer to switch away, even when counterfactual scenarios such as comparisons without noise, or with a newly drawn noise term, suggest such choices may not be payoff-maximizing. While the preferences over the possible role of luck and noise in subsequent comparisons are broadly similar across groups, some evidence suggests that losers prefer more noise than winners. This asymmetry highlights how success and failure, even when partially driven by chance, can systematically shape persistence and mobility in competitive settings, with important consequences.

Second, looking at whether these choices appear to be driven by attribution distortions in beliefs, that indeed is the case. Not only are there clear attribution distortions in beliefs, they are also asymmetrical across dimensions. Participants exhibit motivated reasoning when updating beliefs about their own ability, particularly discounting negative signals and updating in the opposite direction to what Bayes predictions would suggest, while remaining conservative after positive ones. In contrast, updates about luck remain largely consistent with Bayesian inference. Further a comparison involving the same type of belief updating framed in terms of neutral 'robots' shows smaller belief distortions. Together this outcome-dependent asymmetry indicates that distortions are not due to misunderstanding statistics, but rather to ego-protection when self-image is at stake.

Along with documenting new evidence on choices following chance-driven outcomes, this paper contributes to the literature on belief updating, overconfidence, and decision-making under uncertainty in three key ways. First, it links behavioral choices to attribution distortions: I show that the asymmetries in persistence and switching after winning or losing map directly to biased updating about ability, highlighting how distorted beliefs spill over into consequential decisions. Second, the experiment provides evidence on multi-dimensional updating from a single signal, allowing me to track how the same outcome simultaneously informs beliefs about relative ability and about luck. Whereas prior work has typically examined only one dimension or abstracted from non-Bayesian updating (reviewed by Benjamin (2019)), I leverage the joint nature of the signal to document systematic asymmetries across both. Third, the design highlights the sensitivity of these biases to contextual factors: widening the variance of the random adjustment term changes the informativeness of outcomes, and subjects are responsive to this structure. When comparing the responses to the exact same updating task for neutral ("robot") agents, belief distortions are smaller, pointing to motivated reasoning rather than statistical misunderstanding as the key mechanism.

Taken together, these findings shed light on the cognitive mechanisms through which chance-driven outcomes shape beliefs, choices, and competitive dynamics. By linking biased belief updating to observable behavior, the paper shows how ego-protection can sustain persistence among winners, promote exit or switching among losers, and ultimately shape dynamics in competitive markets. These results have clear real-world parallels and implications for labor market behavior, educational testing, and strategic decision-making. In the labor market, workers who succeed may cling to the colleagues and settings that validated their ability, while those who fail may seek to distance themselves from unfavorable comparisons. At the same time, both groups appear to distort self-attributions more than they distort their statistical reasoning, suggesting that workplace beliefs about ability and luck are shaped less by cognitive limits than by self-protection and self-enhancement. Such biases can shape reapplication decisions, persistence, and mobility in ways that standard rational models of learning may miss.

The rest of the paper is structured as follows: Section 2 discusses and situates this paper in the related literature, Section 3 describes the experimental design, Section 4 analyzes the data and presents the key results, and Section 5 concludes with discussing future lines of research.

2 Literature Review

Attribution theory in psychology posits that people explain outcomes in self-serving ways - crediting successes to internal factors such as ability or effort and blaming failures on external factors like bad luck (Miller and Ross (1975); Weiner (1985), Mezulis et al. (2004)). This “self-attribution bias” has been observed across many domains, particularly when outcomes are identity-relevant (Campbell and Sedikides (1999)).

In economic settings, attributional reasoning is important for tournament behavior, where performance is judged relative to others and outcomes may be partly stochastic. Prior experimental work (e.g., Shastry et al. (2020)) shows that noisy feedback about relative performance elicits differential attributions to ability versus luck, which can shape willingness to enter future competitions. Unlike some previous studies where feedback perfectly reveals rank (Ertac (2011)), I introduce explicit randomness into performance comparisons, making attribution a key inference problem. Further, almost all previous literature on this topic deals with beliefs about relative ranks in a big group (Coutts (2019), Shastry et al. (2020), Möbius et al. (2022) etc), while I focus on performance relative to another individual. My design also relates to the ambiguity literature (Machina and Siniscalchi (2014)), as participants know the distribution of the random adjustment but not its realization, creating scope for ambiguity-averse updating.

Another related literature examines how feedback influences persistence and performance in subsequent tasks. Some experimental literature demonstrates that individuals are more likely to continue competing, investing effort, or re-entering tournaments after success than after failure (Gill and Prowse (2014), Buser and Yuan (2019)). Failures often trigger disengagement or avoidance, while successes reinforce persistence and willingness to re-compete. However, other work like Buser (2016) find instead an increase in challenge seeking after a loss by men. These behavioral patterns are consistent with attributional asymmetries: if winners interpret outcomes as revealing ability, they gain confidence and continue; if losers attribute outcomes to external noise or luck, they may persist, but if they attribute it to low ability, they may prefer to switch contexts or avoid further comparisons. By connecting belief updating to subsequent opponent choices, this paper provides direct evidence that attribution affects not only internal beliefs but also observable decisions.

Within economics, a parallel literature has explored deviations from Bayesian updating. Bayesian inference provides a normative framework for how beliefs should be revised in light of new evidence. Classic experimental work in psychology and economics has shown, however, that individuals often deviate systematically from this benchmark (Tversky and Kahneman (1973), Grether (1980) for example). Such deviations include updating too little relative to Bayes’ rule (conservatism), responding more strongly to favorable than to unfavorable signals (asymmetric updating), and base-rate neglect among others. While early studies documented under- and over-reaction in neutral settings (for example Massey and Wu (2005)), more recent work highlights that biases can be amplified in ego-relevant contexts. Eil and Rao (2011) and Möbius et al. (2022) show that when feedback concerns personal ability, individuals tend to over-weight good news and under-weight bad news. These distortions align with models of belief-based utility (Bénabou and Tirole (2002)), where beliefs serve both informational and psychological functions. Related evidence (Coutts (2019); Möbius et al. (2022)) indicates that asymmetry and conservatism are reduced in affectively neutral control tasks, suggesting that ego-protection drives part of the bias. This paper adds to this literature, by including a neutral robot control to test the differences in belief updating

A final contribution of this paper is to examine joint belief updating across multiple dimensions. Much of the literature has studied single signals about ability, neglecting the fact that real-world outcomes often conflate skill and chance. Recent work has begun to explore multidimensional updating (Buser et al. (2018); Zimmermann (2020); Coutts et al. (2024)), but evidence remains scarce. By simultaneously eliciting beliefs about both ability and luck, and by varying the informativeness of the outcome signal, this experiment provides new insights into how individuals allocate explanatory weight across dimensions. This structure allows me to separate general inferential difficulty from motivated distortions. The finding that higher variance treatments shift weight toward luck in line with Bayesian predictions suggests that people understand the informational structure of the environment—but only apply this correctly when ego is not threatened.

3 Experimental Design

The goal of this experiment was twofold: allow us to study asymmetries in how people attribute success and failure outcomes to external factors like luck vs internal factors like own merit or ability; and see how that affects subsequent decision making. The experiment studies belief updating on both attributional dimensions of interest under a noisy feedback set-up, and then presents choices to subjects to determine subsequent preferences. Given the possible element of ego-relevance and motivated reasoning involved when updating beliefs about your own success or failure, the design involves a neutrally worded control about robots where the entire set-up is structurally the same. Thus the control provides us a benchmark for thinking about how people make inferences on multiple dimensions from a single ambiguous signal.

The experiment has two parts: Part 1 involves a cognitive performance task, while Part 2 comprised of a sequence of incentivized belief-elicitation and decision-making questions.[1] In this section I will first mention the experimental procedure and briefly describe the implementation and incentives. Then I will describe the experiment in more detail, and end the section by discussing the treatments and setting out the Bayesian benchmarks.

Experimental Procedures

The experiment was conducted online via Qualtrics during July 2025 and consisted of two main parts. 221 subjects completed the experiment on Prolific, based on the typical Prolific eligibility screeners of being at least 18 years of age who have completed at least 100 prior submissions on Prolific with an approval rating of 95% or greater, and chose the United States as their residence. In addition to a fixed $3 participation payment for a 15-minute study, participants could earn up to an additional $2 in bonus payment, determined by a random selection of either Part 1 or Part 2 for payout.

If Part 1 was selected for payment, participants earned $0.10 for each correct response. If Part 2 was selected, one of the questions was randomly chosen as the question-that-counts, and subjects were paid the amount that they earned in that specific question. Participants were incentivized to answer accurately, and were told in the instructions that answering accurately would maximize their possible bonus payment. Subjects could click on a hyperlink to view a demonstration of how the most accurate

[1] Full experimental materials are provided in Appendix Section A.4.

answer would maximize their chance of the largest possible bonus. Depending on the type of question, the bonus payment was determined using the outcome under the subject's preferred choice, the quadratic scoring rule, or the Becker-DeGroot-Marschak (BDM) method. See Appendix Section A.2 for more details on the exact incentives.

3.1 Part 1: Cognitive Task

Participants completed as many Raven's Progressive Matrices (John and Raven (2003)) as possible within a four-minute time limit, with a maximum of 20 questions. I will refer to these question as puzzle questions in the experiment and here. Questions were randomly drawn from Sets C, D and E of the original Raven's test, arranged in increasing difficulty without explicitly informing participants. Raven's matrices measure abstract reasoning and fluid intelligence, and are widely used in economics and psychology.[2] One untimed practice question was shown prior to the task to familiarize participants with the format. Each question required selecting the image that best completed a visual pattern (See Practice Question 1 in Appendix Section A.4).

Performance in this Part 1 task, i.e. the number of correct puzzles answered will serve as the basis for subsequent comparisons in Part 2. Subjects were not given feedback about their raw performance before moving on. While they knew that a second part would follow, they were not informed about its nature or that it would involve eliciting beliefs about their absolute and relative performance.

3.2 Part 2: Beliefs and Choices

3.2.1 Prior Beliefs

Part 2 began by grounding participants in their own expectations about performance. They first reported their beliefs about how they had performed in Part 1, both in absolute terms and relative to others. For relative performance, they provided (i) the mean probability of scoring more than a randomly selected counterpart, and (ii) a full probability distribution for the score difference, reported in pre-specified buckets (the percent chance of scoring at least 8 puzzles less than another person, between 4 and 7 puzzles less, ... , exactly the same, ..., between 4 and 7 puzzles more and so on).[3] These priors establish a personalized benchmark against which subsequent updating can be measured.

3.2.2 Tournament Setup

Next participants were introduced to the tournament-like stage, where their raw score from Part 1 would be compared to that of a randomly selected counterpart. Before comparison, however, the subjects were informed that a random adjustment 'X' would be added to their score to generate an adjusted score. The range of possible values for X depended on the treatment condition that subjects were randomly assigned to: in Treatment 1 (narrow adjustment), the computer would randomly draw a value of X from

[2] See Benito-Ostolaza et al. (2016) and Gill and Prowse (2016) for applications of Raven's matrices in behavioral economics. The Raven's matrices task was specially chosen for this experiment as a real effort task with probable variation in performance across subjects, not correlated with gender.

[3] See Part 2 Question 3 in Appendix Section A.4.

$\{-3, -2, -1, 1, 2, 3\}$, while for those in Treatment 2 (wide adjustment), X would randomly be drawn from $\{-7, -6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6, 7\}$.

In both treatments, the range was displayed, and participants were told there was a 50% chance of a positive adjustment and a 50% chance of a negative one. Tournament outcomes were determined by comparing the subject's adjusted score to the randomly drawn counterpart's unadjusted score.[4] If the own adjusted score was higher than the counterpart's score, subjects were informed that they ***won***; if it was lower they ***lost***; and ties were decided by a fair coin toss. Crucially, participants were told they would only learn the outcome - not their own score, or the value of the computer drawn X, or their counterpart's score - ensuring that any updating relied solely on the win/loss signal. Understanding was verified through worked examples and comprehension checks.

3.2.3 Updated Beliefs

Immediately after learning the tournament result, participants engaged in two belief-updating tasks designed to separately capture attribution to luck and attribution to relative ability. All probability questions were presented as a slider from 0 to 100 that subjects could click on for their response. First, they reported the posterior probability that the adjustment term X had been in the direction of their outcome (positive if they won, negative if they lost). This captures a direct measure of perceived luck. Second, they gave their posterior probability of winning against the same counterpart if there had been no adjustment in an attempt to isolate perceived relative ability after the outcome. Together, these measures provide a two-dimensional view of updating from the same noisy signal, against which Bayesian benchmarks can be calculated from participants' earlier priors (more on this shortly).

3.2.4 Choices

Updating beliefs is one thing; acting on them is another. To see how attributions translated into tangible competitive preferences, participants were next asked to choose between facing the same counterpart again, facing a new randomly selected counterpart, or stating indifference, under three scenarios: (i) Using the same adjustment value as before (Same X), (ii) Using no adjustment at all (No X), and (iii) Drawing a new adjustment from the same range as in their treatment (New X). If a loss is attributed to the counterpart's high ability, a new opponent might be preferred; if it is attributed to bad luck from the adjustment, the same opponent under different adjustment rules may seem more appealing. These choices reveal whether attribution patterns affect real-world-relevant competitive behavior, not just abstract probability reports.

Having explored opponent preferences, participants were then asked to choose the range that they preferred the random adjustment term X to be drawn from, to be applied in a new match against a different opponent: No adjustment $\{0\}$, Narrow range $\{-3, -2, -1, 1, 2, 3\}$, or Wide range $\{-7, -6, \ldots, -1, 1, \ldots, 6, 7\}$. This decision captures a different behavioral dimension: whether participants strategically embrace or avoid randomness in determining outcomes. Those attributing a prior loss to low ability may prefer no adjustment; those attributing it to bad luck may prefer a wider range to amplify the role of chance.

[4] Alternatively, the design could have been worded as adding an adjustment to both scores. To keep the interpretation simple, the adjustment was only applied to the subject's own score allowing for both positive and negative adjustments, which can also be interpreted as a net adjustment.

3.2.5 Robot Control Condition

To benchmark updating without ego involvement, participants also completed a robot control task. They were shown the same prior distribution they themselves had reported earlier but framed as belonging to the difference in strength between Robot A and Robot B. After seeing Robot A's win/loss outcome based on its adjusted strength relative to Robot B, subjects reported (i) the probability the adjustment was in the direction of Robot A's outcome, and (ii) the probability that Robot A would win against Robot B with no adjustment. Importantly, subjects were informed that the adjustment term was drawn from the same range as in their own tournament, and the win/loss signal provided was exactly their own outcome from earlier. By comparing these responses to those from the self-relevant condition, I can identify distortions in updating that arise specifically from self-serving or self-protective motives, and also document underlying difficulties in belief updating on two dimensions from a one signal.

Treatments

The two treatment arms - narrow vs. wide adjustment ranges - directly manipulate the potential influence of luck on the win/loss outcomes. If participants are Bayesian, a wider range should lead to higher posterior attribution to luck and smaller updates to perceived ability. Comparing responses across treatments will allow me to test whether participants recognize this difference in potential luck and incorporate it into both their belief updating and their strategic competitive choices.

Bayesian Benchmarking

A key feature of the experimental design is that it allows for the construction of clear Bayesian benchmarks for belief updating. This is possible because (i) priors over own relative ability were elicited directly in the form of a belief distribution before the noisy outcome signal, and (ii) the signal structure is fully known and transparent to participants. In particular, the only information subjects observe after the tournament is whether they won or lost against their counterpart, with the knowledge that an independent adjustment X has been added to their score drawn from a known distribution. Given this setup, each subject's posterior beliefs about both relative ability without adjustment and luck in terms of getting a positive adjustment can be derived using Bayes' rule, providing a normative standard against which actual reported beliefs can be compared.

To look at the Bayesian posterior, let us start by defining a few variables. Let S_i denote subject i's raw score from Part 1, the number of puzzle questions they answered correctly, and S_j denote the raw score of the randomly chosen counterpart j. Let $D = S_i - S_j$ denote the raw score difference between the subject and their counterpart. Before observing the outcome, each participant provides a prior distribution $\pi(d)$ over D.

Since the prior distribution was elicited using an interval method, there are different ways I can approximate the distribution for Bayesian calculations. The simplest way would be to assign probability weights to the mid points of the intervals, while more sophisticated methods like approximating a normal distribution or Beta distribution over the entire range is also possible. For the analysis in this paper, I will assume a scaled Beta distribution from -9 to +9 (in line with the bucket ranges provided), and find α and β parameters that best fit each subject's reported probability distribution. This choice was driven by the overall better fit of beta distributions given the reported interval beliefs, relative to normal distribution. These fitted beta distributions for each subject will allow me to compute $\Pr(D > 0)$ and conditional probabilities for any relevant thresholds.

The randomly selected adjustment term X that is added to subject i's score is drawn from a known discrete support with uniform probability, and participants are informed of the 50/50 chance of positive/negative adjustment. This allows exact computation of likelihoods. In our setup, the distribution $f_X(x)$ is a uniform prior on $\mathcal{X} = \{-N, -(N-1), \ldots, -1, 1, \ldots, (N-1), N\}$ where $N = 3$ in Treatment 1, and $N = 7$ in Treatment 2.

The outcome $Y \in \{\text{Win, Loss}\}$ denotes the win or loss signal that is determined by $D + X$. So the Bayesian posteriors can be calculated mathematically using the discrete sum over X values combined with the fitted Beta distribution for D.

3.2.6 Updating about Relative Ability

The first dimension of interest is how participants update beliefs about their own relative ability compared to their counterpart, conditional on the tournament outcome. Specifically, after learning whether they won or lost, they report the posterior probability that they would win without any adjustment against the same counterpart.

The Bayesian posterior distribution of scoring higher than their counterpart without adjustment, i.e. $Pr(D > 0)$, conditional on the observed outcome Y being a win, is given by:[5]

$$\begin{aligned}\Pr(D > 0 \mid \text{Win}) &= \frac{\Pr(D > 0 \ \wedge \ D + X > 0)}{\Pr(D + X > 0)} \\ &= \frac{\displaystyle\sum_{x \in \mathcal{X}} f_X(x) \ \Pr(D > -x) \ \Pr(D > 0 \mid D > -x)}{\displaystyle\sum_{x \in \mathcal{X}} f_X(x) \ \Pr(D > -x)}\end{aligned}$$

For a participant who observes a loss, the posterior probability that they had higher ability than their counterpart without adjustment is:

$$\begin{aligned}\Pr(D > 0 \mid \text{Loss}) &= \frac{\Pr(D > 0 \ \wedge \ D + X < 0)}{\Pr(D + X < 0)} \\ &= \frac{\displaystyle\sum_{x \in \mathcal{X}} f_X(x) \ \Pr(D > -x) \ \Pr(D > 0 \mid D < -x)}{\displaystyle\sum_{x \in \mathcal{X}} f_X(x) \ \Pr(D < -x)}\end{aligned}$$

3.2.7 Updating About Luck

The second dimension of interest is the extent to which subjects attribute outcomes to the adjustment X, i.e., whether their win or loss is perceived to have been driven by receiving a favorable or unfavorable draw.

[5] See Appendix Section A.3 for a detailed explanation of the Bayes calculation and a worked out example for clarity.

The posterior probability that the participant received a positive adjustment given a win is:

$$\Pr(X > 0 \mid \text{Win}) = \frac{\Pr(X > 0 \ \wedge \ D + X > 0)}{\Pr(D + X > 0)} = \frac{\displaystyle\sum_{x>0} f_X(x) \ \Pr(D > -x)}{\displaystyle\sum_{x\in\mathcal{X}} f_X(x) \ \Pr(D > -x)}.$$

Analogously, the Bayes posterior probability that the subject received a negative adjustment given a loss is:

$$\Pr(X < 0 \mid \text{Loss}) = \frac{\Pr(X < 0 \ \wedge \ D + X < 0)}{\Pr(D + X < 0)} = \frac{\displaystyle\sum_{x<0} f_X(x) \ \Pr(D < -x)}{\displaystyle\sum_{x\in\mathcal{X}} f_X(x) \ \Pr(D < -x)}.$$

Together, these two sets of Bayesian benchmarks will allow me to test not only whether participants update beliefs in line with Bayes' rule, but also whether deviations are systematically biased in self-serving directions. See Appendix Section A.3 for more details.

4 Results

In this section, I present the main experimental results. I begin by looking at the performance and prior beliefs that subjects hold, then turn to comparing subjects' behavioral choices and dopcumenting the systemtatic difference there. To better understand what might be driving these differences, I next turn to look at whether there are differences in belief updating, focusing on the reported posterior beliefs about subjects' relative ability and the role of luck against the Bayesian benchmarks derived from their stated priors. This allows me to identify systematic deviations in updating and to test whether these deviations differ after wins versus losses. I then examine heterogeneity across treatments, focusing on how the informativeness of the adjustment distribution affects updating. Finally, I contrast ego-relevant outcomes with the control condition involving neutral agents ("robots"), which isolates motivated reasoning from general inferential errors.

4.1 Overview

A total of 221 participants took part in this experiment. Since participants were randomly assigned an adjustment value and other opponent, I will exploit the fact that conditional on own performance, the eventual win or loss outcome is effectively random (just like Shastry et al. (2020)). In my experiment, 93 participants won their tournament, while the remaining 128 subjects lost.[6] Qualtrics randomly assigned subjects into the two treatments, resulting in 126 subjects in Treatment 1, and 95 in Treatment 2.

[6] For the initial 20 participants, the other scores were values taken from a pilot conducted with NYU students in March 2025. For the remaining participants, the other score values were taken from the responses of these initial 20 subjects. The analysis in the paper includes all participants, but dropping the initial 20 does not change the conclusions.

Remember that for the tournament-like setup, subjects first submitted their initial beliefs for the probability distribution of score differences (the number of puzzle questions solved correctly) between themselves and a randomly drawn counterpart (the ***prior***). They then had a randomly drawn adjustment value "X" added to their score (the ***adjustment***), and this adjusted score was compared with that of a randomly drawn ***counterpart***. Depending on the treatment, the adjustment term was either drawn from {-3, -2, -1, 1, 2, 3} (Treatment 1) or from {-7, -6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6, 7} (Treatment 2). In either treatment, there was a 50% chance of receiving a positive adjustment and a 50% chance of receiving a negative adjustment, with the expected value of the adjustment being zero. A comparison between the subject's adjusted score and their counterpart's score determined the win/loss outcome (the ***signal***) - if their adjusted score was higher, they received a signal that they *'won'*; if it was lower they *'lost'*, and if it was exactly equal, a fair coin toss would determine their outcome. Subjects were then asked for their updated beliefs about the adjustment term, and ability relative to that specific counterpart (the ***posteriors***). Next participants were asked to make choices between having their score compared to the same counterpart, a new randomly chosen person and stating indifference under two different counterfactual scenarios. Lastly participants had to respond to the same belief updating questions but when framed for neutral "robots".

4.2 Prior Beliefs and Confidence

First let's look at how well participants performed in the Raven's puzzle matrices task in Part 1 of the experiment. Remember that subjects were given 4 minutes to solve as many puzzles as possible up to a maximum of 20. Figure 1 shows a histogram of the actual number of puzzles participants solved correctly. On average participants solved 8.56 puzzles correctly in the allotted time.

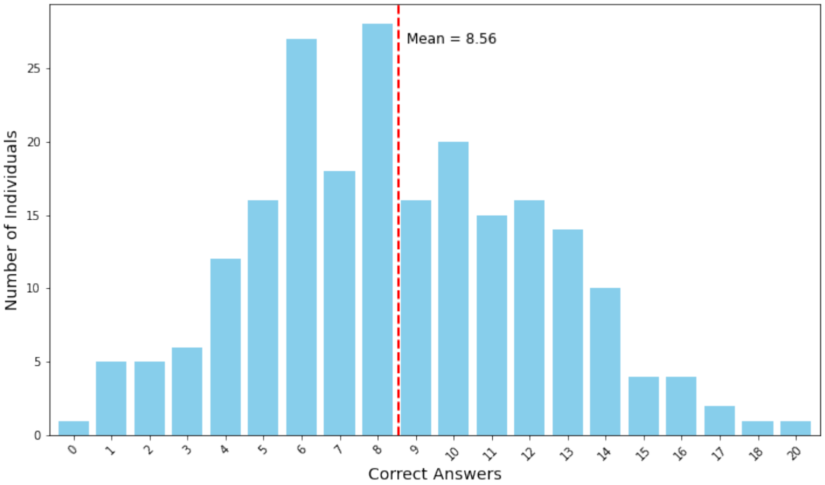


Figure 1: Histogram of Number of Correct Answers

Note: The histogram depicts the number of subjects based on the number of Raven's puzzle questions they solved correctly in Part 1, for all subjects in the experiment.

Since I am interested in looking at differences between those who went on to win/lose their tournament, I will begin differentiating between these two groups. Notice that at this point in the experiment, participants did not know about the upcoming tournament, and so this categorization is ex-post for the purposes of this analysis.

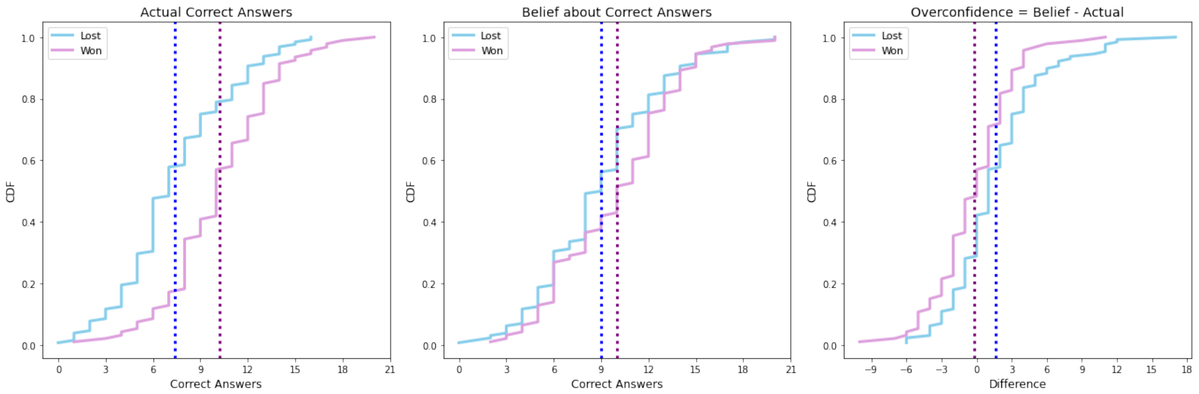


Figure 2: Cumulative Distributions by Outcome

Note: The CDFs depict the following, split by those who won, and those who lost: The left-most depicts the actual number of puzzle questions solved correctly, the middle subplot shows how many questions subjects think they solved correctly, and the subplot on the right plots overestimation, defined as the difference between what subjects think and how many they actually solved correctly.

Starting with the leftmost graph in Figure 2, these cumulative distributions (CDFs) depict the number of puzzle questions that participants solved correctly. On average, those who went on to win solved 10.22 questions, while those who lost solved 7.36. (Refer to Figure A1 for the histogram.) The graph in the middle depicts the distributions of reported (mean) beliefs about own performance in Part 1.[7] Notice that despite significant differences (p value from a Kolmogorov-Smirnoff test is p=0.00) in actual performance, there is a smaller difference in the beliefs that subjects hold about their absolute performance across the two groups, as seen in Column 1 and 2 of Table 1. While those who won solve 2.8 more questions correctly on average relative to those who lost, they only believe they solved 1 question more.

Next I am interested in overconfidence, and seeing how that may affect belief updating. There are multiple ways of thinking about overconfidence.[8] As an initial measure, we can look at overestimation of one's actual ability or performance, defined here as the difference between the beliefs subjects hold about their performance, and their actual performance. The rightmost graph in Figure 2 plots the CDFs for this measure of overconfidence in terms of overestimation. On average, 52% of subjects report beliefs higher than their actual performance, with only 12% responding with the correct value. On average all subjects believe they have score 0.86 puzzles more than they actually have, suggesting a slight overconfidence in absolute ability in general.

Notice that despite performing worse on average, those who go on to lose are more overconfident about their performance ($p < 0.05$), both in absolute terms and relative to those who won. (Column 3 and 4 in Table 1 below). As you would expect, after controlling for their own score, there is no longer any

[7] Subjects were asked how many questions they think they solved correctly. The incentivization used was to elicit mean beliefs instead of the more standard modal belief elicitation. See Appendix Section A.2 for more details.

[8] See Moore and Healy (2008) for a detailed discussion about overconfidence, overestimation, over-placement and overprecision.

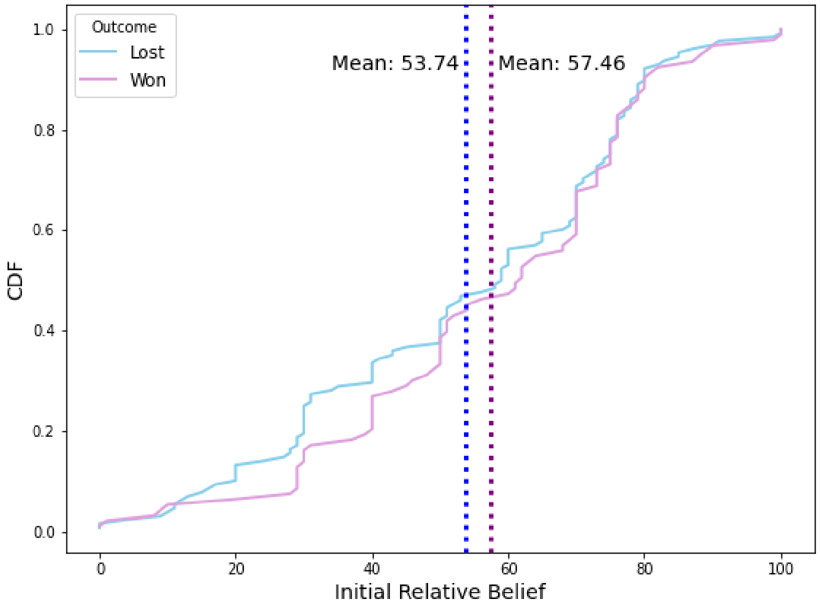


Figure 3: Distribution of Initial Relative Ability

Note: The CDF depicts the responses split by those who won and those who lost of the question: What do you think is the percent chance that you solved more puzzle questions correctly than a random other person (on a scale of 0-100)?

difference across winners and losers in their level of overestimation, with those scoring higher less likely to over estimate their performance.

Since the focus is on situations where others are involved, beliefs about relative performance are even more important. Participants were asked what was probability that they have scored more than a random other participant who has solved the exact same task, giving us a measure of over-placement as discussed in Moore and Healy (2008). Figure 3 plots the cumulative distribution of these belief responses. Although those who go on to win hold slightly higher beliefs about their relative ability, the differences in the two distributions are not statistically significant. Even when controlling for actual performance, there are no differences in initial beliefs about relative ability, although scoring one more question correctly corresponds to only a 1.19% average increase in beliefs (Column 6).

Result 1: On average people overestimate their own absolute performance, but not so much their relative performance. Between those who go on to win versus those who lose, losers are more overconfident about absolute ability, while winners are slightly more confident about relative ability.

4.3 Choices

What ultimately matters in practice is how individuals act. In the labor market, for example, we rarely observe people's internal beliefs about why they succeeded or failed; what is visible are the decisions they make in response - whether to reapply for a promotion, change employers, or withdraw altogether. To capture this behavioral dimension, my experiment examines subjects' actual choices, providing a direct

Table 1: Performance and Confidence

	Actual Performance	Beliefs	Overestimation	Overestimation	Overplacement	Overplacement
	1	2	3	4	5	6
Won	2.856***	0.989*	-1.866***	-0.624	3.720	0.308
	(0.471)	(0.529)	(0.490)	(0.467)	(3.177)	(3.116)
Score				-0.435***		1.195***
				(0.0711)		(0.408)
Constant	7.359***	9.000***	1.640***	4.843***	53.742***	44.950***
	(0.312)	(0.349)	(0.348)	(0.676)	(2.156)	(4.140)
N	221	221	221	221	221	221

Note: Standard errors in parentheses. Won refers to those that went on to receive the 'Won' outcome in the tournament setup, and Score refers to the actual number of puzzle questions solved correctly in Part 1 of the experiment. Column 1 looks at the differences in actual performance. Column 2 reports beliefs about the number of correctly solved questions. Columns 3 and 4 report regression coefficients for overestimation, defined as the difference between reported beliefs and actual performance. Column 5 and 6 report coefficients from a regression on the initial reported probability of having scored higher than a random other person (on a scale of 0-100).

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

lens on how attribution patterns translate into concrete actions and I start by looking at these.

4.3.1 Choices over counterparts

I begin by examining subjects' preferences over whom their performance will be compared against after receiving their outcome. Subjects could choose to have their comparison against the current counterpart, a random new person, or state that they were indifferent under three different settings: given the current realization of the random adjustment term X (Same X), under a counterfactual where no adjustment is made (No X), and under a new draw of X from the same range as used for their outcome (New X). Figure 4 plots the share of subjects making each choice using a stacked bar chart. This figure shows a striking divergence between winners and losers.

In the baseline situation with the current (same) draw of X (the leftmost bar for both winners and losers), staying with the 'current counterpart' is deterministic, resulting in the same outcome previously communicated. As expected, losers predominantly abandon their counterpart in favor of a new person and winners strongly favor remaining with their counterpart. The comparison of interest is how these choices change under the counterfactual conditions, the next two bars in each subgraph. Notice how the counterfactual conditions reduce these extremes: losers became somewhat more likely to stay with their counterpart, while winners become more likely to report indifference.

To test these patterns formally, I estimate a multinomial logit model with "current counterpart" as the base category and include interactions between winning and the decision condition. The results, reported in Table 2, confirm the descriptive patterns. While the table coefficients depict log-odds of a choice, I interpret the marginal effects here for ease of understanding. For losers, the baseline probability of switching to a new person is 71%, with just 12% staying with their counterpart. Winners, on the

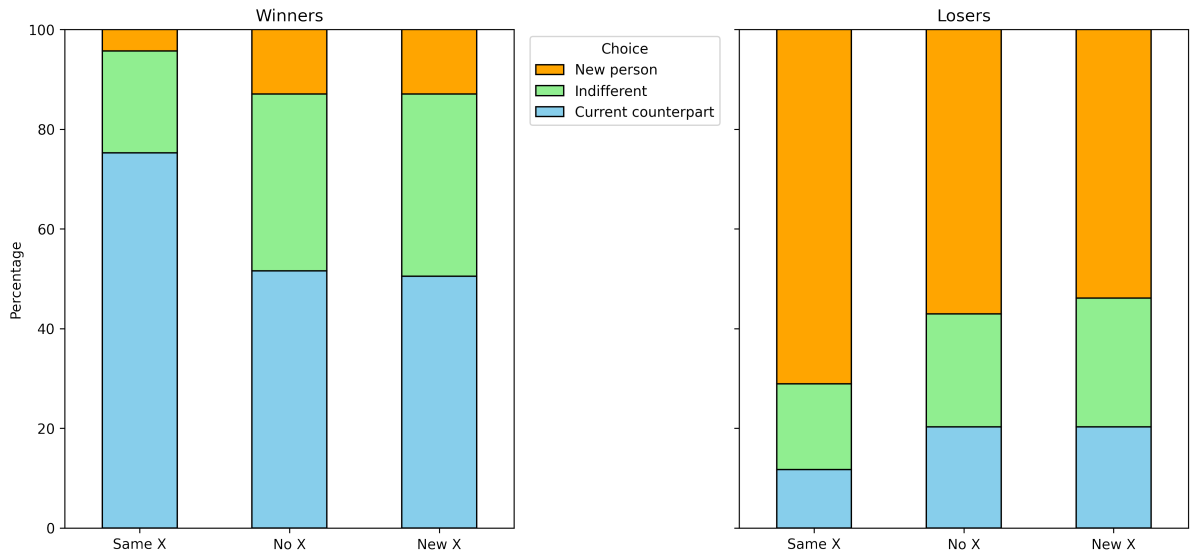


Figure 4: Choices Across the Different Settings

Note: The stacked bar graphs depict the proportion of subjects who won and lost respectively, that chose the different options. Same X refers to the comparison of scores using the same adjustment that was used to determine the initial win/loss outcome, No X is comparison without any adjustment term being included, and New X is a comparison of score when a new adjustment term is drawn.

other hand, display the opposite pattern: 75% stay with their counterpart and only 4% switch to a new person. Under alternative decision framings, winners become substantially less rigid: only around half remain with their counterpart, while indifference rises to over one-third and switching increases modestly to 13%. Losers also moderate somewhat, with the probability of switching falling to about 54–57% and staying with the counterpart rising to around 20%. A joint test confirms that these effects differ systematically across winners and losers (p = 0.007). What is interesting to note is that there appears to be hardly any difference in choice preferences when no adjustment is to be made and the situation where a new adjustment term is to be drawn.

These choice patterns highlight how attribution of noisy outcomes shapes subsequent behavior. Winners' tendency to stick with their counterpart suggests that they interpret success as reflecting true ability, reinforcing the desirability of the status quo. Losers, by contrast, frequently abandon their counterpart, consistent with attributing failure to external circumstances rather than to their own lack of skill. Importantly, when the choice is re-framed under counterfactual conditions, both groups become less extreme, indicating that attribution is not fixed but malleable. This points to a mechanism by which motivated beliefs translate into real decisions: individuals not only update beliefs in a self-serving way but also adjust their actions to preserve favorable self-perceptions or to distance themselves from unfavorable ones.

4.3.2 Preferences over distributions

I next turn to subjects' choices about the distribution of the random adjustment term X when facing a new comparison against a random person. Here, subjects could choose no adjustment, having X be

Table 2: Multinomial Logit Regression of Choice of Opponent

	Indifferent	New Person
Won	-1.837***	-4.709***
	(0.428)	(0.591)
Decision: No X	-0.273	-0.770**
	(0.428)	(0.360)
Decision: New X	-0.143	-0.826**
	(0.423)	(0.360)
Won × No X	1.208**	2.248***
	(0.551)	(0.707)
Won × New X	1.129**	2.325***
	(0.547)	(0.708)
Own Score	0.051	0.015
	(0.032)	(0.030)
Constant	0.004	1.695***
	(0.428)	(0.383)
Observations	663	
Base Outcome	Current Counterpart	
Base Decision	Same X	

Note: Robust standard errors in parentheses. Own score is measured by number of correct answers in Part 1. The decisions No X asks about a situation with no adjustment term, and New X is a new draw of the adjustment term from the same range, relative to the Same X baseline. The base category for the dependent variable `choice` is "Current Counterpart". Coefficients are from a multinomial logit regression reflecting the log-odds of choosing a new opponent, or being indifferent relative to picking the current counterpart.

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

drawn from the narrow range {-3,-2,-1,1,2,3}, or having X be drawn from the wider range {-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7}. Figure 5 shows the distribution of choices separately for winners and losers, and by treatment condition.

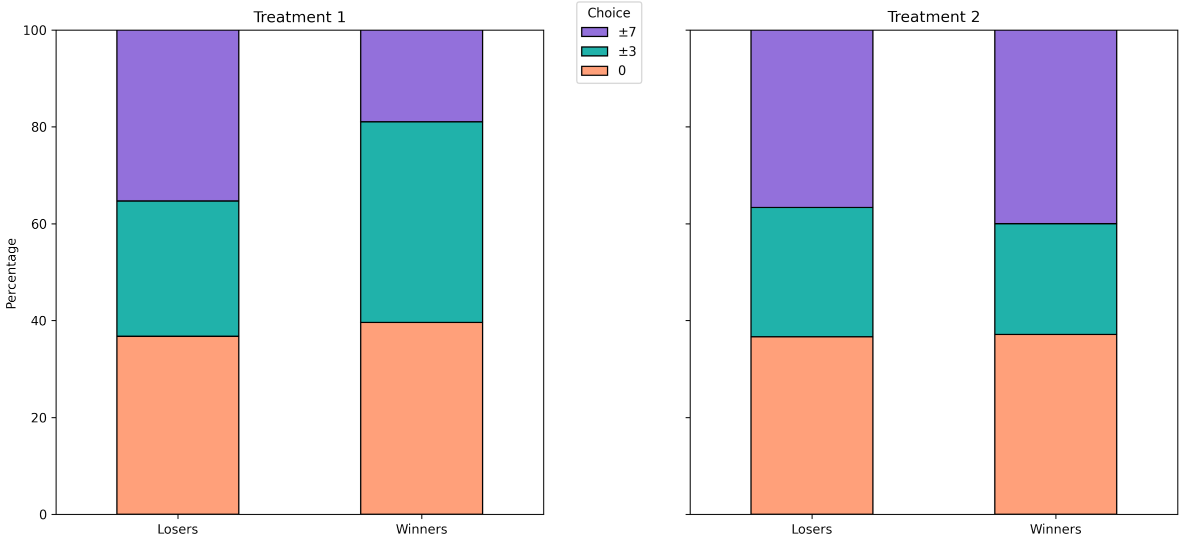


Figure 5: Choices about distribution of the adjustment term

Note: The stacked bar graphs depict the proportion of subjects who won and lost respectively within each treatment, that chose the different options. ± 7 refers to preferring a new adjustment term be drawn from {-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7} when having the score compared to a new randomly drawn person, ± 3 is preferring it be drawn from {-3,-2,-1,1,2,3}, and 0 is preferring having no adjustment made.

Unlike the strong asymmetries observed in counterpart choices, here we find much weaker effects. Winners were somewhat more likely to favor narrowing the distribution, while losers showed slightly greater demand for widening it, but overall the distributions of choices were similar. The multinomial logit model reported in Table 3 confirms that neither winning nor the treatment condition significantly predicts distribution choice (joint p = 0.276).

The distribution-choice results suggest that while subjects clearly react to *who* they are paired with, they are less sensitive to *how* luck enters the comparison. Across treatments, winners display a mild preference for narrowing the adjustment range, consistent with wanting outcomes to reflect ability more clearly, whereas losers show a modest tilt toward widening the range, consistent with seeking greater scope to attribute outcomes to luck. However, these tendencies are weak relative to the sharp asymmetries observed in counterpart choices. Moreover, the lack of systematic treatment differences indicates that subjects' preferences for uncertainty are not strongly shaped by the statistical properties of the adjustment distribution itself. Put differently, subjects appear more concerned with distancing or reinforcing a social counterpart than with altering the underlying variance of luck, suggesting that attribution operates more through social framing and relative comparisons than through risk preferences over randomness.

Result 2: Choices reveal a clear self-serving asymmetry in the social domain - winners overwhelmingly stay with their counterpart while losers abandon theirs - yet preferences over adjustment distributions remain broadly similar across groups, with only weak evidence that winners prefer less noise and losers

Table 3: Multinomial Logit Regression of Adjustment Range Preference

	Narrow Adjustment	Wide Adjustment
Won	1.186**	0.349
	(0.477)	(0.552)
Treatment 2	0.147	0.273
	(0.465)	(0.461)
Won × Treatment 2	-1.018	0.135
	(0.753)	(0.723)
Own Score	-0.236***	-0.283***
	(0.052)	(0.057)
Constant	1.482***	1.990***
	(0.517)	(0.501)
Observations	221	
Base Outcome	No Adjustment {0}	

Note: Robust standard errors in parentheses. Base outcome is "No Adjustment" {0}. Treatment 2 is the wide range in the original tournament. The dependent variable is the choice of adjustment range from which to have an X drawn for comparison with a new random opponent. Coefficients reflect log-odds of choosing the narrow ({-3,-2,-1,1,2,3}) of wide ({-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7}) adjustment versus no adjustment.

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

more.

These patterns mirror how workers respond to noisy outcomes in the labor market. After success or failure, individuals often reassess whether their performance reflects higher or lower relative ability compared to colleagues, or whether outcomes were simply the result of chance. The experiment shows that such attributions primarily affect comparisons with others: winners seek to preserve comparisons that reinforce their apparent superiority, while losers prefer to move away from counterparts that highlight a deficit. By contrast, individuals show little inclination to adjust the role of randomness itself, just as workers rarely focus on the noisiness of evaluation processes. Instead, attribution works through social comparisons, shaping whether people continue to measure themselves against the same peers or seek a different reference point, rather than through preferences over the underlying distribution of luck.

To better understand what might be the underlying factors driving these differences in choices after a success or failure outcome, I now turn to digging deeper into the belief updating that participants are doing after they get their outcome. In the next subsection, I will look at whether there are systematic differences in how participants update their beliefs about the two dimensions of interest.

4.4 Belief Updating Behavior

My experimental design involves belief updating on two dimensions after the single ambiguous signal: the probability of receiving a favorable/unfavorable adjustment (X) and the probability of winning against the same counterpart without any adjustment (i.e. the probability of having scored more puzzles correctly than the other person). Using the prior reported by each individual about their relative ability, the given prior about the adjustment term (50/50), and the signal (win/loss), we can calculate a Bayesian benchmark for the posteriors of both the dimensions of interest as shown in Section 3 and Appendix Section A.3.

Using the Bayesian calculations, I start by looking at updating of beliefs relative to the prior for both the dimensions of interest (luck and relative ability). Although earlier studies (Barron (2021), Coutts (2019), Möbius et al. (2022)) exclude subjects who reported posterior beliefs that were updated in the opposite direction compared to the Bayesian prediction (i.e. an upward shift in own relative ability beliefs after a loss signal), this behavior seems relevant in my context and so will not be dropped.

4.4.1 Updating about Relative Ability

When thinking about relative ability, the prior belief is the initial probability with which you think you scored more than a random person, without any adjustment to your score, so simply the probability of doing better than someone else. Although subjects reported a single value for this belief from Part 2 Question 2, there is more information in the entire belief distribution, and so the prior being used for this analysis is the probability of scoring more than a counterpart without adjustment as taken from the fitted beta distribution on the reported probability distribution.

The Bayesian posterior tells us how much a rational agent would update, given the individual's prior, and the win/loss signal they received. Defining a ***belief update*** as the difference between the posterior and prior, I start off by looking at how far subject's belief updates are from the Bayesian benchmark.

Figure 6 is a scatter plot where each dot depicts an individual in the experiment. The x-axis represents the difference between the Bayesian posterior and the prior, thus depicting how much a rational agent should update their beliefs in lieu of the signal. The y-axis represents the difference between the subject's reported posterior (referred to here as the empirical posterior) and the same prior, thus depicting how much individuals in my experiment actually updated their beliefs. Individuals who received a win signal are shown by purple diamonds, and those who lost are depicted by blue circles. The black markers represent the group averages.

The immediate takeaway from this graph is what appears to be a mirror image along the x=0 vertical line, with Bayes suggesting a reduction in beliefs after a loss, and updating beliefs positively after a win. However if you think about this more, notice that all individuals in the bottom left quadrant and top right quadrant are those that updated beliefs in the direction that Bayes suggests. For those that won, on average Bayes suggest increasing beliefs by 38.3 percentage points, while subjects only increase by 14 on average. This pattern is driven by 60% of respondents who are conservative, i.e. they do not update beliefs to the extent that Bayes suggests (as depicted by the dots being below the 45 degree line in the top right quadrant). 24% of subjects actually decrease their beliefs about their ability to win without adjustment.

In sharp contrast is the belief updating after receiving a loss signal. 73% of those who lost with adjustment are in the top left quadrant, implying they increased their beliefs about the probability of doing better than the counterpart without any adjustment, which is in the opposite direction relative to the

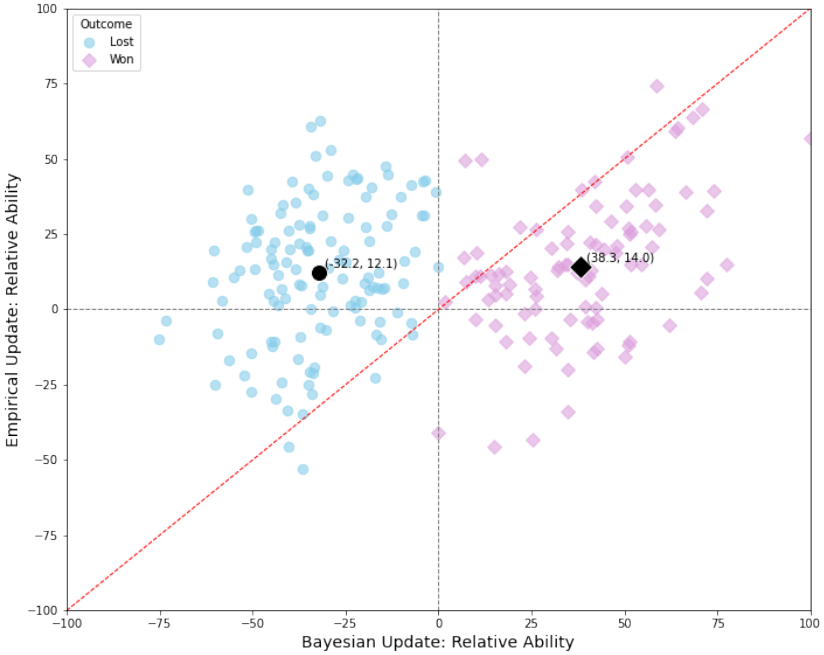


Figure 6: Scatter Plot of Belief Updating vs Bayesian: Relative Ability

Note: Each dot represents an individual in the experiment. Both axes depict belief updates, i.e. the difference between the posterior and the individual's prior about their relative ability (scoring more than their opponent without any score adjustment). The x-axis shows the Bayes prediction of how much should an individual update their belief after the noisy signal, while the y-axis shows how much the subject actually updated their beliefs.

Bayes prediction. Only 27% of all subjects decrease their beliefs after the loss. Although Bayes suggests an approximately 32 percentage point reduction in beliefs about relative ability on average, subjects are instead increasing beliefs by 12% over the prior. It is important to note that there were no statistically significant differences in the prior relative ability beliefs between those who won and lost, even when using the beta distribution assumption. Together this suggests a strong tendency of attributing the loss outcome to an unfavorable adjustment term.

Given the richness of data I have, I can look if there are any asymmetries across winners and losers in how they specifically update their beliefs about the direction of the adjustment term they received, which is what I turn to next.

4.4.2 Updating about Luck

Remember that across both treatments, subjects were clearly informed that they had a 50% chance of receiving a positive adjustment, and a 50% chance of receiving a negative adjustment. Given this unambiguous and simple prior, I am interested in looking at how individuals update on the probability of having received a positive adjustment term.[9]

[9] In the experiment subjects were asked to provide the posterior probability for having received an adjustment term in the direction of their outcome (so having received a positive adjustment if they won, and having received a negative

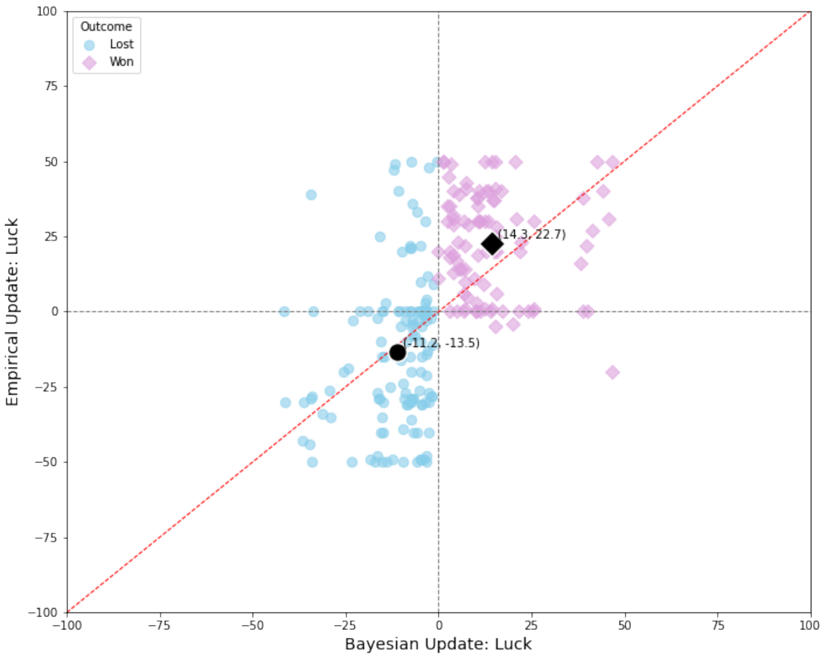


Figure 7: Scatter Plot of Belief Updating vs Bayesian: Luck

Note: Each dot represents an individual in the experiment. Both axes depict belief updates, i.e. the difference between the posterior and the individual's prior about the role of luck (the probability of having received a positive adjustment term). The x-axis shows the Bayes prediction of how much should an individual update their belief after the noisy signal, while the y-axis shows how much the subject actually updated their beliefs.

Figure 7 now depicts the updating by individuals about the probability of having received a positive adjustment relative to the 50/50 prior on the y-axis, and the Bayesian update on the x-axis. As earlier, the update is defined as the difference between any posterior and the 50/50 prior. It is important to note here that although the prior about the adjustment term was the same for everyone, the posterior depends on each individual's prior about their relative ability since the signal incorporates both dimensions.[10] Thus it will not be the case that the Bayes posterior will be exactly 50/50 for everyone, and the variation along the x-axis is expected.

Just like with the updating about relative ability, we expect the Bayesian updating to be in opposite directions based on the signal. However, in contrast to what we saw for the ability updating, subjects appear to be updating much more in line with Bayes when asked about the adjustment term. Both winners and losers are not conservative on average, if anything they seem to be slightly over updating after winning on average.

adjustment if they lost). However, for ease of interpretation of results and comparing them across the two dimensions, I will be presenting the probabilities in terms of having received a positive adjustment for all subjects. For those who lost, this just means taking the complementary probability (i.e. 100 - their reported value), and a similar adjustment for the Bayesian posterior from Section 3.

[10] Refer to the Bayesian benchmark calculations in Appendix Section A.3 for more clarification through an example.

Result 3: Winners are conservative about updating on their relative ability, while losers overwhelmingly update in the wrong direction, relative to Bayes predictions. In contrast, both sides are nearly Bayesian when updating about luck in terms of the adjustment term.

It is interesting to see the differences that emerge between the updating being done on two different dimensions from the same signal. You would think that as the belief updating about relative ability suggests people thinking they have a higher chance of winning without adjustment after a loss using adjustment, they should also be the ones to think there was a very low probability that they received a positive adjustment term. However I find that the rates of updating beliefs about the adjustment term are very similar to those who didn't increase their beliefs about their relative ability even after winning.

One explanation for the differences could be coming from the different priors: for the adjustment term, subjects were given an unambiguous and simple 50/50 prior. However, for relative ability, subjects provided their own prior which may have made updating more complex. This may suggest individuals being farther away from Bayesian, but it wouldn't explain the systematic tendency for those who lost to positively update their beliefs about being better at the puzzle task than their counterpart if no adjustment were to take place. This result seems to suggest something more is going on beyond just different priors.

What instead seems like a stronger argument in favor for the differences seen is the context of the updating. When individuals are considering the adjustment term, they are thinking about an abstract number randomly selected by a computer. Instead, when they are thinking about their relative ability, a strong element of motivated reasoning becomes involved. Even though both posteriors are asked in a context that has ego-relevance: the subject having won or lost with the adjustment, it appears that there is a stronger motivated reasoning logic behind relative ability, but not as much behind the adjustment term that may have played a role in the outcome.

4.5 Treatments

The purpose of the two treatments was to see how people update about the two dimensions - luck and relative ability, under two different possible ranges of the adjustment term X. Even though the probability of having received a positive or negative adjustment is perfectly symmetrical (50/50 prior) in both treatments, the higher variance in possible values of X in Treatment 2 affects how much people should update on both dimensions. I are interested in seeing whether people understand that the feedback signal is more about this adjustment than their own relative ability, and compare the signal informativeness that people extract for both dimensions.

To get at this, I will turn to scatter plots similar to Figure 6 and Figure 7 but now split the responses by treatment. Looking at Figure 8 first, notice how at first glance there does not seem to be much of a difference between the two treatments.

When looking at belief updates about relative ability, notice that the Bayes prediction suggests much less updating under Treatment 2, both for losers and especially more for winners. This is intuitive because the informativeness of the win/loss signal about underlying relative ability depends on how much variation comes from luck. When the adjustment range is narrow, a win provides relatively stronger evidence that one's raw ability exceeded the opponent's, so the Bayesian benchmark calls for a larger upward

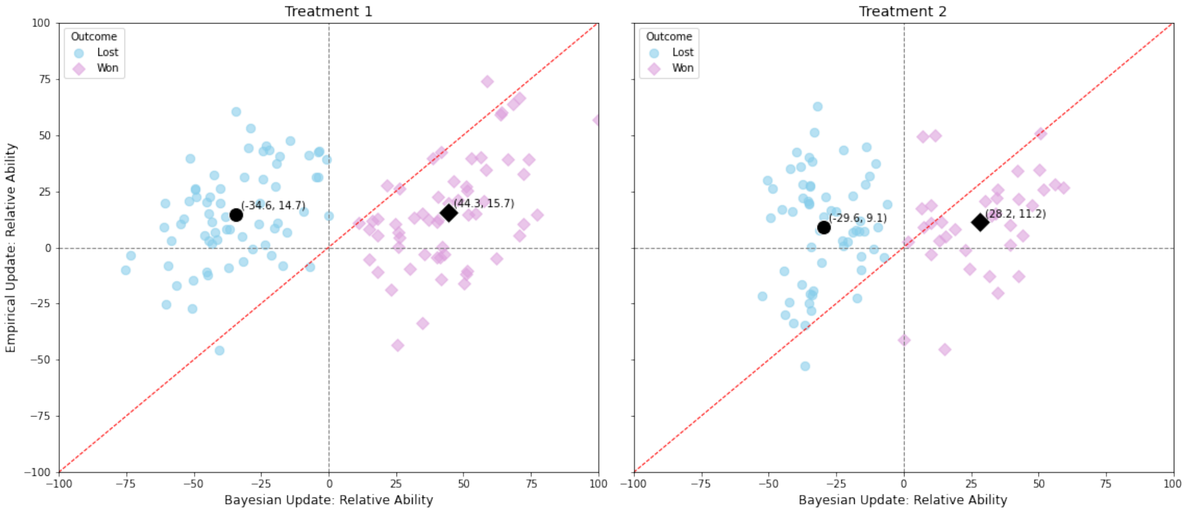


Figure 8: Scatter Plot of Belief Updating vs Bayesian by Treatment: Relative Ability

Note: Each dot represents an individual in the experiment, where the responses are split by which treatment the subject was in (Treatment 1 had the adjustment term drawn from a narrow range, Treatment 2 had it drawn from a wider range). Both axes depict belief updates, i.e. the difference between the posterior and the individual's prior about their relative ability (scoring more than their opponent without any score adjustment). The x-axis shows the Bayes prediction of how much should an individual update their belief after the noisy signal, while the y-axis shows how much the subject actually updated their beliefs.

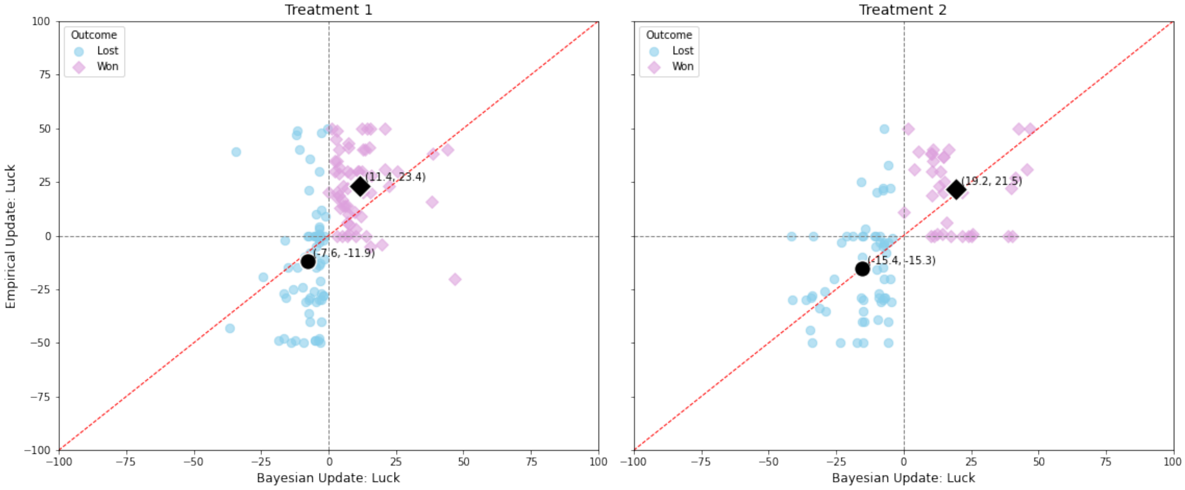


Figure 9: Scatter Plot of Belief Updating vs Bayesian by Treatment: Luck

Note: Each dot represents an individual in the experiment, where the responses are split by which treatment the subject was in (Treatment 1 had the adjustment term drawn from a narrow range, Treatment 2 had it drawn from a wider range). Both axes depict belief updates, i.e. the difference between the posterior and the individual's prior about the role of luck (having received a positive adjustment term). The x-axis shows the Bayes prediction of how much should an individual update their belief after the noisy signal, while the y-axis shows how much the subject actually updated their beliefs.

belief revision about relative ability. Conversely, in Treatment 2, where the adjustment term has higher variance, the same win is less diagnostic of ability; it is more likely that the favorable outcome was driven by the adjustment rather than underlying skill. As a result, the rational (Bayesian) posterior implies a smaller upward shift in relative ability beliefs after a win in Treatment 2 compared to Treatment 1. On average, after a win Bayes suggests updating beliefs by 44.3 percentage points when the range of adjustment is low, but only suggests a 28.2 percentage point increase with the wider range.

In the case of a loss, the informativeness of the signal again depends on the range of the adjustment term. In Treatment 1 the adjustment range is small, so a loss is stronger evidence that one's raw ability may be lower than the opponent's. The Bayesian benchmark therefore calls for a larger downward revision in relative ability beliefs after a loss. In Treatment 2 however, the adjustment introduces much more noise. Losing by an adjusted score could easily have been driven by receiving a highly unfavorable adjustment rather than being genuinely weaker than the opponent. Thus, the Bayesian posterior implies a smaller reduction in relative ability beliefs after a loss in Treatment 2 compared to Treatment 1. The Figure also shows that while a majority of subjects update in the wrong direction relative to Bayes after a loss, this effect is slightly less in Treatment 2.

However, turning to Figure 9 shows much more of a difference across treatments. This is because even though the 50/50 prior remains the same, and the expected value remains 0 across both treatments, the range differs. Remember subjects were only asked about the direction of the adjustment term, not what value they think they received.

When updating beliefs about the adjustment term itself, the logic works in reverse to that for relative ability. Because the prior is a simple 50/50, the informativeness of the win/loss signal about whether the adjustment was positive depends on how wide the adjustment distribution is. In Treatment 1, outcomes are less diagnostic about the adjustment direction, since differences in raw scores can more easily swamp the relatively small adjustment. By contrast, in Treatment 2, the adjustment range is larger, so the win/loss signal provides stronger evidence about whether the adjustment was positive or negative. As a result, Bayesian updating about the adjustment term should be more pronounced in Treatment 2 than in Treatment 1. This is evidenced in Figure 9, where subjects appear to be updating along Bayesian predictions on average. Winners in Treatment 1 actually seem to be overestimating the probability of having received a positive adjustment, updating beliefs by 23.4 percentage points on average, relative to the average Bayes update of 11.4 percentage points.

Result 4: When the adjustment range is wider, outcomes become less informative about ability but more informative about luck - subjects' updating reflects this shift in signal informativeness.

4.6 Motivated Reasoning: Self-Serving Attribution in Ability vs. Luck

Research has long shown behavioral biases in belief updating relative to Bayesian predictions. A central challenge in interpreting biased updating is whether it reflects a genuine misunderstanding of how informative signals are, or whether people strategically bend beliefs when outcomes concern themselves. To speak more directly to this literature and to provide a clean way to distinguish between these two possibilities, I included a control condition for belief updating. As described in the experimental design, towards the end of the study, subjects were asked to respond to beliefs regarding two robots (Robot A and B). The setup in terms of the prior, range of possible adjustment values, and the signal were exactly the same as earlier in the ego-relevant condition when they were responding about their own performance in the Raven's puzzles task. That is, the distribution of initial weight lifting capacity differences between the robots was exactly what the subject had provided as their probability distribution about relative performance in the earlier Raven's puzzles; they were informed that the computer-drawn adjustment term X was from the same possible distribution as earlier, and they received the same win/loss outcome signal for Robot A as they had received earlier for themselves. Subjects were then asked to provide two posterior beliefs with the exact same wording as earlier: the probability the adjustment term was positive/negative depending on the outcome, and the probability Robot A was stronger than B if no adjustment was applied.[11]

This is useful because it provides us two responses for each subject under the exact same prior, noise and signal set-up, where one is about them (will refer to this as the ego condition), and the other using neutral wording about robots (will refer to this as the control condition). This comparison allows me to isolate whether the systematic deviations from Bayesian updating observed earlier reflect a genuine misunderstanding of signal informativeness, or whether they are driven by self-serving motives in interpreting feedback about one's own ability. When updating about the robots, subjects face the same statistical problem but without ego or identity concerns. If their updating aligns more closely with Bayesian predictions in the robot condition, this would suggest that the distortions in the self condition stem from motivated reasoning rather than a lack of statistical understanding. Conversely, if subjects show similar patterns in both cases, it would imply that the deviations reflect cognitive or inferential limits rather than self-serving biases.

I will now compare the belief updating by subjects on both dimensions of interest across these ego and control conditions by looking at the difference in posterior belief reported under the ego condition (about self) and the posterior belief reported for the same thing under the control condition (about robots). A positive value of this difference suggests holding higher posterior belief when it is about yourself.

Figure 10 depicts histograms and the fitted kernel distributions of this difference, where the left panel is for the belief about scoring higher than the other person/robot without any adjustment (relative ability), and the right panel is for the probability of receiving a positive adjustment term. Only 7% and 13% of subjects report the same values across both conditions in each of the belief updating dimensions respectively, with there being quite a spread in the responses.

Looking at the contrast between those who won and lost, Figure 11 depicts the cumulative distributions of this difference by outcome. The left panel shows that there is no asymmetry across winners and losers in updating about the probability of doing better than the other without adjustment. However, the right panel shows a significant (p = 0.00 for the K-S test) difference when updating about the role of luck in terms of the adjustment for the outcome, where winners differently tend to update their belief more when it is about their own performance, but less when it is about robots.

Further of interest is seeing whether within each win/loss outcome there are systematic differences across the two belief updating dimensions. Figure 12 depicts the cumulative distribution for those who won on the left and those who lost on the right. Winners appear to show more of a difference in updating

[11] See Question 13 and 14 in Appendix Section A.4.

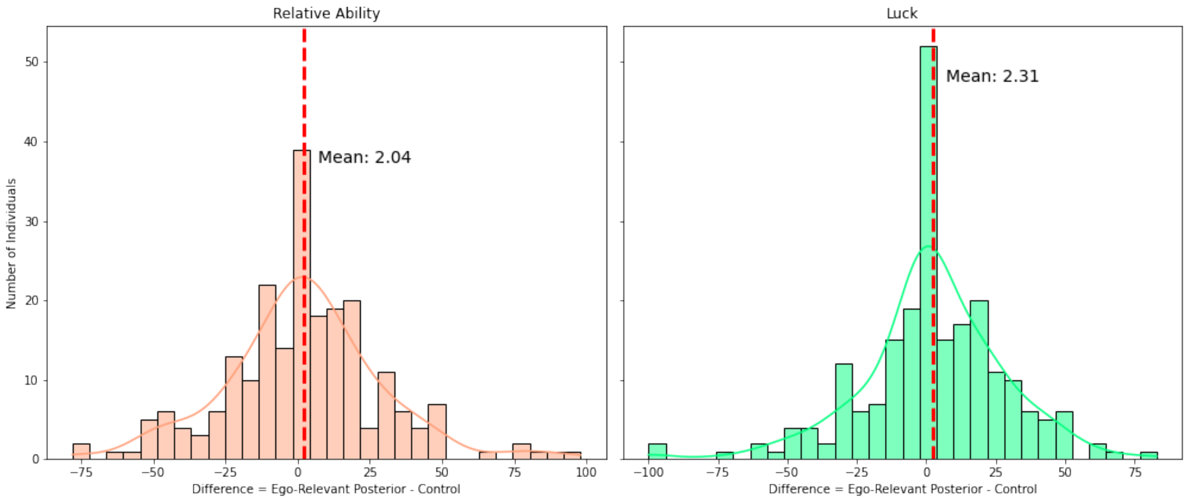


Figure 10: Difference in Belief Updating Across Ego and Control Conditions

Note: Both histograms depict the difference between the posterior reported under the ego-relevant condition (when asked about their own outcome) and the posterior reported in the control condition (when asked about the "robot" outcomes). The fitted curve is the based on the kernel desisty estimation.

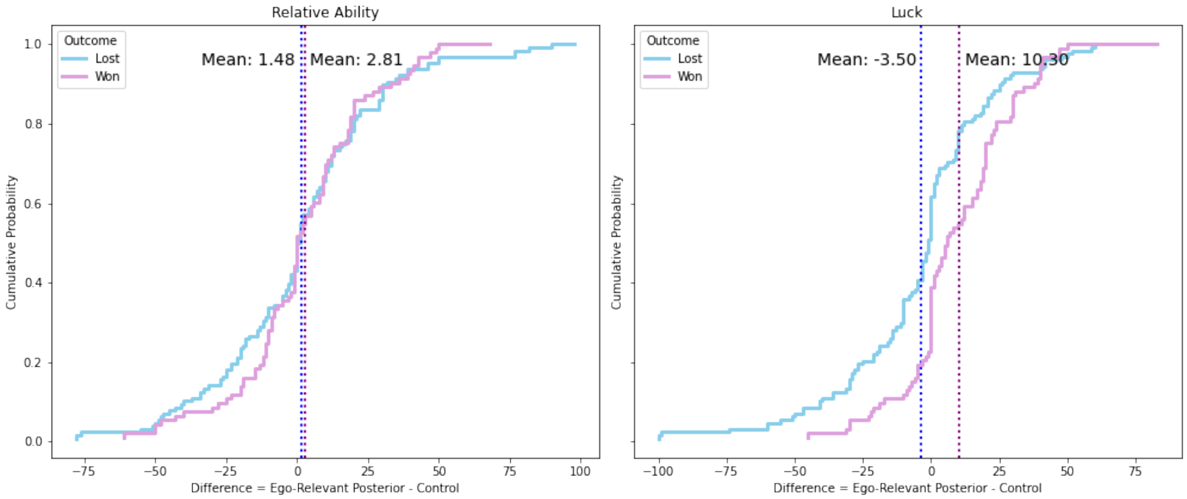


Figure 11: CDFs of Difference between Ego and Control Condition by Outcome

Note: Both CDFs depict the difference between the posterior reported under the ego-relevant condition (when asked about their own outcome) and the posterior reported in the control condition (when asked about the "robot" outcomes), split by the win/loss outcome, for each of the updating dimensions.

across the ego and control condition when updating about the role of the adjustment term, relative to updates regarding relative ability ($p < 0.01$ for the K-S test), while those who lost exhibit similar levels of self serving bias across both dimensions.

Taken together, the comparison of ego and control conditions provides clear evidence of motivated reasoning in how individuals interpret feedback. When the task is framed in terms of their own performance,

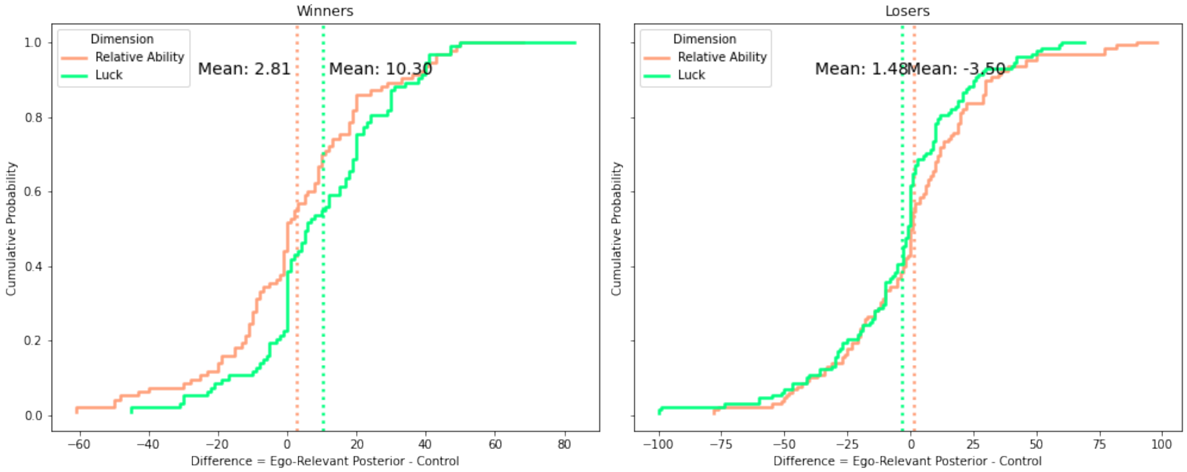


Figure 12: Cumulative Distributions Split by Outcome

Note: Both CDFs depict the difference between the posterior reported under the ego-relevant condition (when asked about their own outcome) and the posterior reported in the control condition (when asked about the "robot" outcomes), split by the two updating dimensions, for those who won and lost.

subjects' posteriors systematically diverge from those they report under an otherwise identical robot condition. These distortions are not uniform: subjects appear relatively consistent across ego and control conditions when updating beliefs about relative ability, but they display significantly larger gaps when updating beliefs about the role of the adjustment term. In particular, winners amplify the role of luck when outcomes are framed as their own, while downplaying it when the same outcome is attributed to robots.

While winners show stronger divergences between ego and control conditions when updating about the adjustment term, losers' responses are more muted. Losers display self-serving bias in both attribution dimensions, but the gap between ego and control is less pronounced than for winners. This suggests that after negative outcomes, individuals may already be motivated to diffuse responsibility across both ability and luck, leaving less room for systematic differences between ego-relevant and neutral contexts. In other words, losers seem to engage in "broad-based defensive updating," treating both relative ability and luck in ways that soften the sting of failure, whereas winners selectively reinterpret luck to bolster the meaning of success. This asymmetry suggests that deviations from Bayesian updating are not simply due to cognitive limits in processing signals, but are shaped by self-serving motives - individuals adjust their interpretation of randomness to protect or enhance ego-relevant beliefs about success.

Result 5: Comparing ego and control conditions reveals systematic self-serving bias: belief distortions are modest for relative ability but significantly larger for updates about the role of the adjustment term, with winners in particular inflating ego-relevant interpretations of luck.

5 Conclusion

This paper examined how individuals update beliefs and make choices when outcomes reflect both ability and random noise. By designing an experiment that disentangles beliefs about relative ability from beliefs about luck, and by including ego-involved and neutral control conditions, I provide new evidence on the role of motivated reasoning in shaping belief updating and subsequent decisions.

The findings show a consistent pattern: distortions emerge primarily when feedback touches on identity-relevant dimensions. Participants are broadly Bayesian when reasoning about luck, and they adjust appropriately when the informativeness of signals changes. But when feedback implicates their own ability, losers in particular deviate sharply from Bayesian benchmarks, and winners inflate self-serving attributions of luck. These distortions directly affect choices. While preferences over the distribution of randomness remain similar, winners and losers diverge starkly in their willingness to persist with the same counterpart. Success induces persistence with the same counterparts, while failure prompts disengagement.

These results contribute to our understanding of belief formation and decision-making in noisy environments such as education, labor markets, and organizations. They suggest that attributional distortions - assigning credit and blame to ability versus luck - are central drivers of post-outcome behavior. In practical terms, this helps explain why high performers often double down on existing paths while low performers are more likely to disengage or seek fresh comparisons, independent of the true informational value of outcomes.

There are, however, limitations to this study. First, while the experimental setting allows for clean identification of attribution channels, it abstracts from many features of real-world environments, such as repeated interactions, strategic communication, or reputational concerns. The choice set was simplified, in reality individuals can not only stay or switch but also invest in skill, exit competition, or change strategies altogether. Second, the stakes in the experiment are modest relative to real labor market or educational settings, where higher incentives and longer-term consequences may amplify or attenuate motivated reasoning. Third, while the design distinguishes between ability and luck attributions, it cannot fully capture other psychological mechanisms, such as affect regulation or social comparison motives, that may interact with attributional biases.

Future research could extend these insights in several directions. One avenue is to study how attribution-driven updating evolves in repeated environments, where the same individuals face feedback over time and choices accumulate into career trajectories. Another is to examine whether interventions, such as reframing outcomes, providing statistical training, or emphasizing the role of external factors and team-based performance can mitigate self-serving biases in attribution. Finally, it would be valuable to test these mechanisms in higher-stakes or naturally occurring settings, such as educational testing, hiring, or promotion processes, where the consequences of distorted attributions are particularly salient.

Overall, the evidence presented here highlights that individuals are not universally poor statisticians; rather, they are motivated interpreters of noisy outcomes. Belief distortions are targeted, emerging most strongly when feedback threatens or enhances the ego. Understanding this asymmetry is crucial for designing policies and practices that account for how people process and act upon feedback.

References

Barron, K. (2021). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics 24*(1), 31–58.

Bénabou, R. and J. Tirole (2002). Self-confidence and personal motivation. *The quarterly journal of economics 117*(3), 871–915.

Benito-Ostolaza, J. M., P. Hernández, and J. A. Sanchis-Llopis (2016). Do individuals with higher cognitive ability play more strategically? *Journal of Behavioral and Experimental Economics 64*, 5–11.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1 2*, 69–186.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science 62*(12), 3439–3449.

Buser, T., L. Gerhards, and J. Van Der Weele (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty 56*(2), 165–192.

Buser, T. and H. Yuan (2019). Do women give up competing more easily? evidence from the lab and the dutch math olympiad. *American Economic Journal: Applied Economics 11*(3), 225–252.

Campbell, W. K. and C. Sedikides (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of general Psychology 3*(1), 23–43.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics 22*(2), 369–395.

Coutts, A., L. Gerhards, and Z. Murad (2024). What to blame? self-serving attribution bias with multi-dimensional uncertainty. *The Economic Journal 134*(661), 1835–1874.

Eil, D. and J. M. Rao (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics 3*(2), 114–138.

Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization 80*(3), 532–545.

Gill, D. and V. Prowse (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics 5*(2), 351–376.

Gill, D. and V. Prowse (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy 124*(6), 1619–1676.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics 95*(3), 537–557.

John and J. Raven (2003). Raven progressive matrices. In *Handbook of nonverbal assessment*, pp. 223–237. Springer.

Machina, M. J. and M. Siniscalchi (2014). Ambiguity and ambiguity aversion. In *Handbook of the Economics of Risk and Uncertainty*, Volume 1, pp. 729–807. Elsevier.

Massey, C. and G. Wu (2005). Detecting regime shifts: The causes of under-and overreaction. *Management Science 51*(6), 932–947.

Mezulis, A. H., L. Y. Abramson, J. S. Hyde, and B. L. Hankin (2004). Is there a universal positivity bias in attributions? a meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological bulletin 130*(5), 711.

Miller, D. T. and M. Ross (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological bulletin 82*(2), 213.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2022). Managing self-confidence: Theory and experimental evidence. *Management Science 68*(11), 7793–7817.

Moore, D. A. and P. J. Healy (2008). The trouble with overconfidence. *Psychological review 115*(2), 502.

Shastry, G. K., O. Shurchkov, and L. L. Xia (2020). Luck or skill: How women and men react to noisy feedback. *Journal of Behavioral and Experimental Economics 88*, 101592.

Tversky, A. and D. Kahneman (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology 5*(2), 207–232.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review 92*(4), 548.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review 110*(2), 337–363.

A Appendix

A.1 Additional Figures

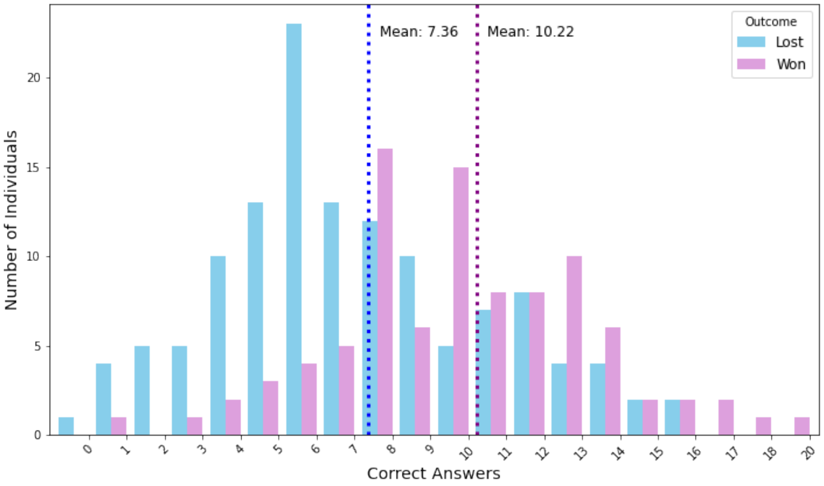


Figure A1: Histogram of Number of Correct Answers By Outcome

A.2 Experiment Incentives

In Part 2 of the experiment subjects were asked different belief and choice questions. The incentives associated with the different kind of questions are as follows.

For questions where subjects were asked to guess the value of something, if their guess matched the actual value, they received a bonus of \$2. The value of the bonus reduced the farther their response was to the actual value, where the bonus was calculated as $\$(2 - 0.1 * ((\text{your guess - actual value})^2))$. For example, suppose the guess is 3 but the answer is 5. Then the bonus for this question would be $\$(2 - 0.1 * ((3 - 5)^2)) = \$1.6$.

For some questions subjects were asked to guess the percent chance of some outcome being true on a slider that ranges from 0% to 100%. They were told: Imagine that you are given two options: 1) If the outcome specified in the question is true, you earn the \$2 bonus, and if the outcome is not true, you receive no bonus. 2) There is a box with 100 balls, where Y balls are black, and the remaining balls are white. A ball will be randomly drawn from the box, and if a black ball is drawn, you receive the \$2 bonus. You receive no bonus if a white ball is drawn. You are asked to report the smallest value of Y that makes you indifferent between these two choices. The computer will then draw a random integer Z between 0 and 100. If the drawn number, Z, is less than your stated Y, you will receive the \$2 bonus if the outcome specified in the question is true. If instead Z is greater than your reported Y, the box will be filled with Z black balls and (100 - Z) white balls, and one ball drawn randomly. You will receive the \$2 bonus is a black ball is drawn. In order to secure the largest chance you earning the \$2 bonus, you should report Y as your most-accurate guess of the outcome being true.

Lastly, when subjects were are asked to make a choice between a few options, the outcome would be determined by the scenario under their preferred choice. If the outcome specified in the question was true, they would earn the $2 bonus.

A.3 Bayesian Benchmarking

Section 3 discussed the calculation of Bayesian benchmarks for the posterior probabilities regarding our two dimensions of interest - relative ability without adjustment, and receiving an adjustment term in the direction of the outcome. Here I will go through the details of the calculation in more depth, and provide a worked out example for clarity.

I had earlier defined the following: Let S_i denote subject i's raw score from Part 1, the number of puzzle questions they answered correctly, and S_j denote the raw score of the randomly chosen counterpart j. Let $D = S_i – S_j$ denote the raw score difference between the subject and their counterpart. Before observing the outcome, each participant provides a prior distribution $\pi(d)$ over D, on which I fit a Beta distribution. X is the randomly drawn adjustment term from the distribution $f_X(x)$ which is a uniform prior on $\mathcal{X} = \{-N, -(N-1), \ldots, -1, 1, \ldots, (N-1), N\}$. The outcome $Y \in \{\text{Win, Loss}\}$ denotes the win or loss signal that is determined by $D + X$.

Remember that given our treatment structure,
For Treatment 1:

$$\mathcal{X} = \{-3, -2, -1, 1, 2, 3\}, \quad f_X(x) = \frac{1}{2N} = \frac{1}{6}.$$

Similarly, for Treatment 2,

$$\mathcal{X} = \{-7, -6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6, 7\}, \quad f_X(x) = \frac{1}{2N} = \frac{1}{14}.$$

Starting with the posterior probabilities about relative ability in terms of scoring higher than their counterpart without an adjustment, after receiving a win with adjustment:

$$\Pr(D > 0 \mid \text{Win}) = \frac{\Pr(D > 0 \ \wedge \ D + X > 0)}{\Pr(D + X > 0)}$$

Focusing on the numerator first, since X is discrete, the joint probability is computed as a sum:

$$\Pr(D > 0 \ \wedge \ D + X > 0) = \sum_{x \in \mathcal{X}} f_X(x) \ \Pr(D > -x) \ \Pr(D > 0 \mid D > -x),$$

where

$$\Pr(D > -x) = 1 - F_{\text{beta}}(-x),$$

and

$$\Pr(D > 0 \mid D > -x) = \begin{cases} 1, & x < 0, \\ \dfrac{1 - F_{\text{beta}}(0)}{1 - F_{\text{beta}}(-x)}, & x > 0. \end{cases}$$

The denominator is

$$\Pr(D + X > 0) = \sum_{x \in \mathcal{X}} f_X(x) \ \Pr(D > -x),$$

so the Bayes posterior becomes

$$\begin{aligned}
\Pr(D > 0 \mid \text{Win}) &= \frac{\sum_{x \in \mathcal{X}} f_X(x)\ \Pr(D > -x)\ \Pr(D > 0 \mid D > -x)}{\sum_{x \in \mathcal{X}} f_X(x)\ \Pr(D > -x)} \\
&= \frac{\frac{1}{2N}[\sum_{x<0}(1 - F_{\text{beta}}(-x)) * 1 + \sum_{x>0}(1 - F_{\text{beta}}(-x))(\frac{1 - F_{\text{beta}}(0)}{1 - F_{\text{beta}}(-x)})]}{\frac{1}{2N}\sum_{x \in \mathcal{X}}(1 - F_{\text{beta}}(-x))} \\
&= \frac{N(1 - F_{\text{beta}}(0)) + \sum_{x<0}(1 - F_{\text{beta}}(-x))}{\sum_{x \in \mathcal{X}}(1 - F_{\text{beta}}(-x))}
\end{aligned}$$

Analogously, after a loss:

$$\Pr(D > 0 \mid \text{Loss}) = \frac{\Pr(D > 0\ \wedge\ D + X < 0)}{\Pr(D + X < 0)}$$

For the numerator

$$\Pr(D > 0\ \wedge\ D + X < 0) = \sum_{x \in \mathcal{X}} f_X(x)\ \Pr(D < -x)\ \Pr(D > 0 \mid D < -x),$$

where

$$\Pr(D < -x) = F_{\text{beta}}(-x),$$

and

$$\Pr(D > 0 \mid D < -x) = \begin{cases} 0, & x > 0, \\ 1 - \dfrac{F_{\text{beta}}(0)}{F_{\text{beta}}(-x)}, & x < 0. \end{cases}$$

The denominator is

$$\Pr(D + X < 0) = \sum_{x \in \mathcal{X}} f_X(x)\ \Pr(D < -x),$$

and so the posterior is calculated as:

$$\begin{aligned}
\Pr(D > 0 \mid \text{Loss}) &= \frac{\sum_{x \in \mathcal{X}} f_X(x)\ \Pr(D < -x)\ \Pr(D > 0 \mid D < -x)}{\sum_{x \in \mathcal{X}} f_X(x)\ \Pr(D < -x)} \\
&= \frac{\frac{1}{2N}[\sum_{x<0} F_{\text{beta}}(-x)(1 - \frac{F_{\text{beta}}(0)}{F_{\text{beta}}(-x)}) + \sum_{x>0} F_{\text{beta}}(-x) * 0}{\frac{1}{2N}\sum_{x \in \mathcal{X}} F_{\text{beta}}(-x)} \\
&= \frac{\sum_{x<0}\big(F_{\text{beta}}(-x) - F_{\text{beta}}(0)\big)}{\sum_{x \in \mathcal{X}} F_{\text{beta}}(-x)}.
\end{aligned}$$

Next, looking at the Bayes posterior probability of receiving an adjustment term in a specific direction. This is relatively straightforward, after a win the posterior of having received a positive adjustment is:

$$\begin{aligned}\Pr(X > 0 \mid \text{Win}) &= \frac{\Pr(X > 0 \;\wedge\; D + X > 0)}{\Pr(D + X > 0)} \\ &= \frac{\sum\limits_{x>0} f_X(x)\ \Pr(D > -x)}{\sum\limits_{x\in\mathcal{X}} f_X(x)\ \Pr(D > -x)}.\end{aligned}$$

From the expression for $Pr(D > -x)$ above, this becomes:

$$\Pr(X > 0 \mid \text{Win}) = \frac{\frac{1}{N}\sum\limits_{x>0}(1 - F_{\text{beta}}(-x))}{\frac{1}{2N}\sum\limits_{x\in\mathcal{X}}(1 - F_{\text{beta}}(-x))}$$

Analogously, the posterior probability of having received a negative adjustment after a loss is:

$$\begin{aligned}\Pr(X < 0 \mid \text{Loss}) &= \frac{\Pr(X < 0 \;\wedge\; D + X < 0)}{\Pr(D + X < 0)} \\ &= \frac{\sum\limits_{x<0} f_X(x)\ \Pr(D < -x)}{\sum\limits_{x\in\mathcal{X}} f_X(x)\ \Pr(D < -x)} \\ &= \frac{\frac{1}{N}\sum\limits_{x<0} F_{\text{beta}}(-x)}{\frac{1}{2N}\sum\limits_{x\in\mathcal{X}} F_{\text{beta}}(-x)}\end{aligned}$$

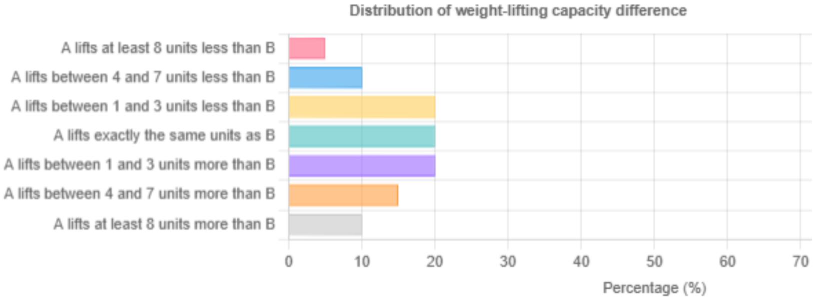


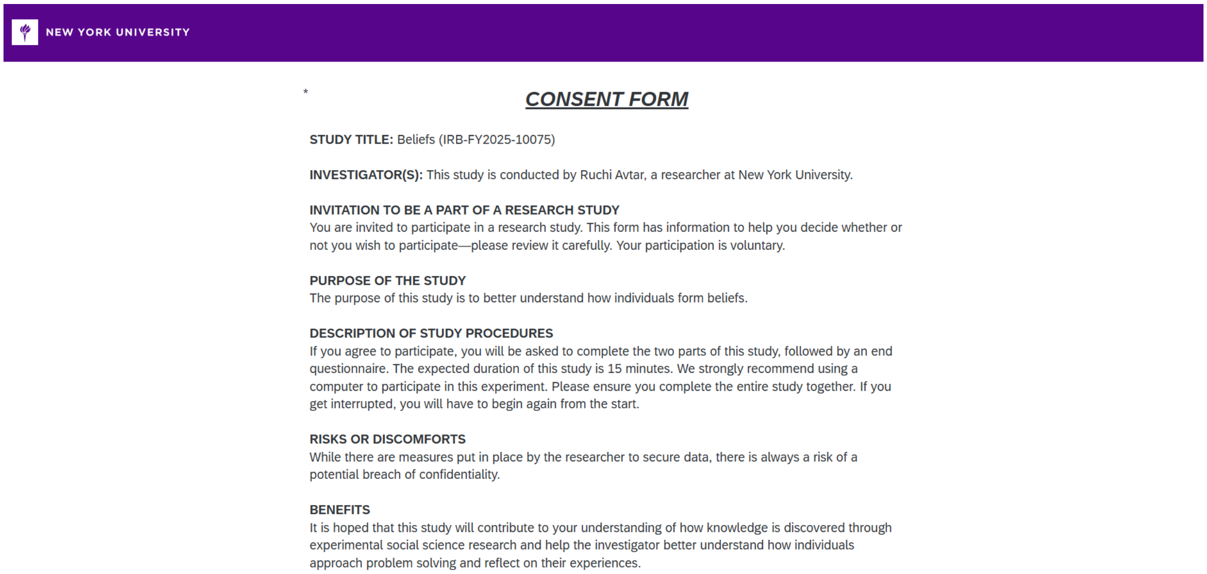
Figure A2: Belief Distribution Example

For more clarity, let's look at a concrete example. Suppose the probability distribution looks like Figure A2. From the empirical distribution, the total probability of Robot A lifting more than B is 45%. Assuming a beta distribution and estimating the best fit, this results in the initial probability being 56.9%. Under Treatment 2, suppose the adjustment term is drawn from {-7,-6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7}. After receiving a signal that Robot A lifted more with adjustment (equivalent to a 'won' signal in the ego-relevant wording), using the Bayes calculation mentioned above, the posterior probability of Robot A lifting more than Robot B without adjustment is 89.2%. In contrast, from the 50/50 prior of receiving a positive adjustment, after the signal, the posterior probability of having received a positive adjustment is 61.1%.

If instead the initial distribution was more skewed to the left, with a 25% probability of Robot A lifting between 4 to 7 units less, 50% probability of it lifting 1 to 3 units less, and 25% chance of it lifting exactly the same, the empirical probability from the reported distribution of Robot A lifting more than B is 0%. Assuming a beta distribution, this becomes 12.5%. If then the signal received is that Robot A lifts more than B, the Bayes posterior about the relative strength of A being higher without adjustment is 34.2%, while the posterior of having received a positive adjustment is now 94.1%.

Together these two instance highlight how the initial belief distribution plays a big role in determining the posterior probabilities for the Bayes predictions.

A.4 Experiment Interface Screenshots



*

CONSENT FORM

STUDY TITLE: Beliefs (IRB-FY2025-10075)

INVESTIGATOR(S): This study is conducted by Ruchi Avtar, a researcher at New York University.

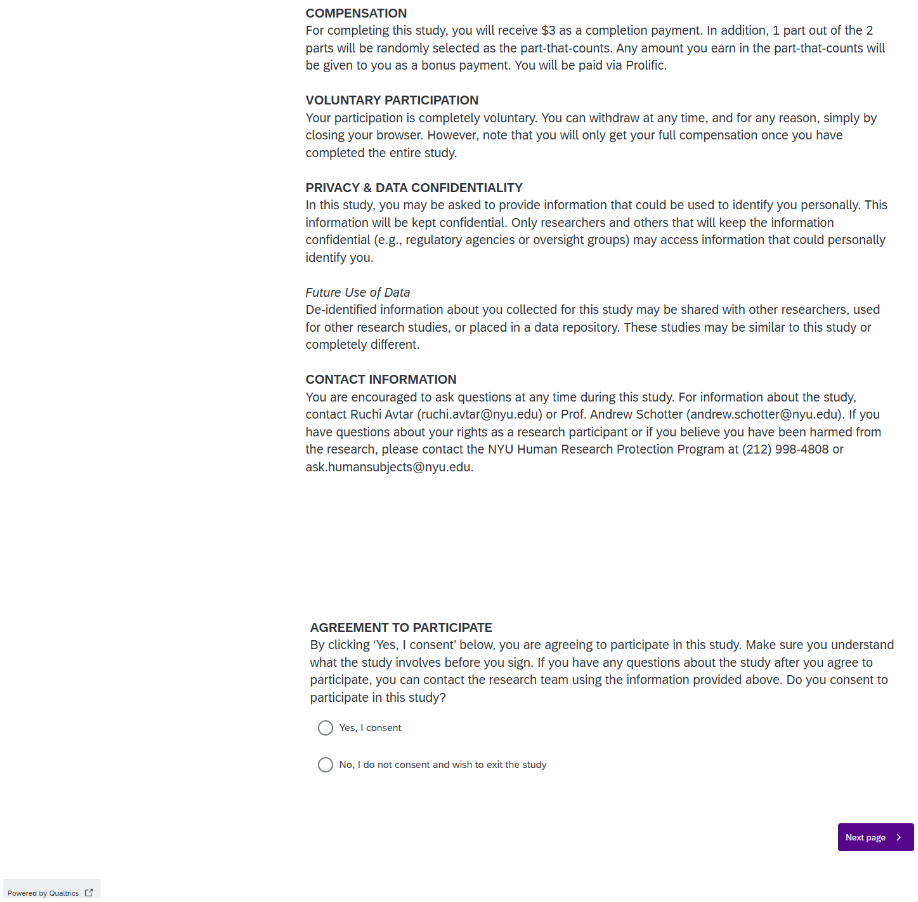
INVITATION TO BE A PART OF A RESEARCH STUDY
You are invited to participate in a research study. This form has information to help you decide whether or not you wish to participate—please review it carefully. Your participation is voluntary.

PURPOSE OF THE STUDY
The purpose of this study is to better understand how individuals form beliefs.

DESCRIPTION OF STUDY PROCEDURES
If you agree to participate, you will be asked to complete the two parts of this study, followed by an end questionnaire. The expected duration of this study is 15 minutes. We strongly recommend using a computer to participate in this experiment. Please ensure you complete the entire study together. If you get interrupted, you will have to begin again from the start.

RISKS OR DISCOMFORTS
While there are measures put in place by the researcher to secure data, there is always a risk of a potential breach of confidentiality.

BENEFITS
It is hoped that this study will contribute to your understanding of how knowledge is discovered through experimental social science research and help the investigator better understand how individuals approach problem solving and reflect on their experiences.



COMPENSATION
For completing this study, you will receive $3 as a completion payment. In addition, 1 part out of the 2 parts will be randomly selected as the part-that-counts. Any amount you earn in the part-that-counts will be given to you as a bonus payment. You will be paid via Prolific.

VOLUNTARY PARTICIPATION
Your participation is completely voluntary. You can withdraw at any time, and for any reason, simply by closing your browser. However, note that you will only get your full compensation once you have completed the entire study.

PRIVACY & DATA CONFIDENTIALITY
In this study, you may be asked to provide information that could be used to identify you personally. This information will be kept confidential. Only researchers and others that will keep the information confidential (e.g., regulatory agencies or oversight groups) may access information that could personally identify you.

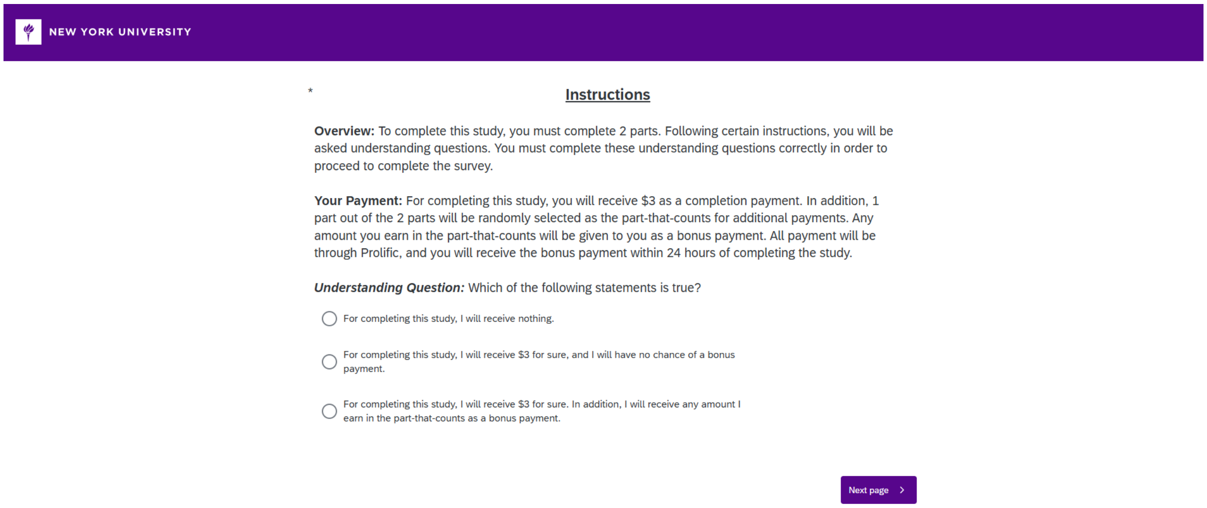
Future Use of Data
De-identified information about you collected for this study may be shared with other researchers, used for other research studies, or placed in a data repository. These studies may be similar to this study or completely different.

CONTACT INFORMATION
You are encouraged to ask questions at any time during this study. For information about the study, contact Ruchi Avtar (ruchi.avtar@nyu.edu) or Prof. Andrew Schotter (andrew.schotter@nyu.edu). If you have questions about your rights as a research participant or if you believe you have been harmed from the research, please contact the NYU Human Research Protection Program at (212) 998-4808 or ask.humansubjects@nyu.edu.

AGREEMENT TO PARTICIPATE
By clicking 'Yes, I consent' below, you are agreeing to participate in this study. Make sure you understand what the study involves before you sign. If you have any questions about the study after you agree to participate, you can contact the research team using the information provided above. Do you consent to participate in this study?

○ Yes, I consent

○ No, I do not consent and wish to exit the study



* **Instructions**

Overview: To complete this study, you must complete 2 parts. Following certain instructions, you will be asked understanding questions. You must complete these understanding questions correctly in order to proceed to complete the survey.

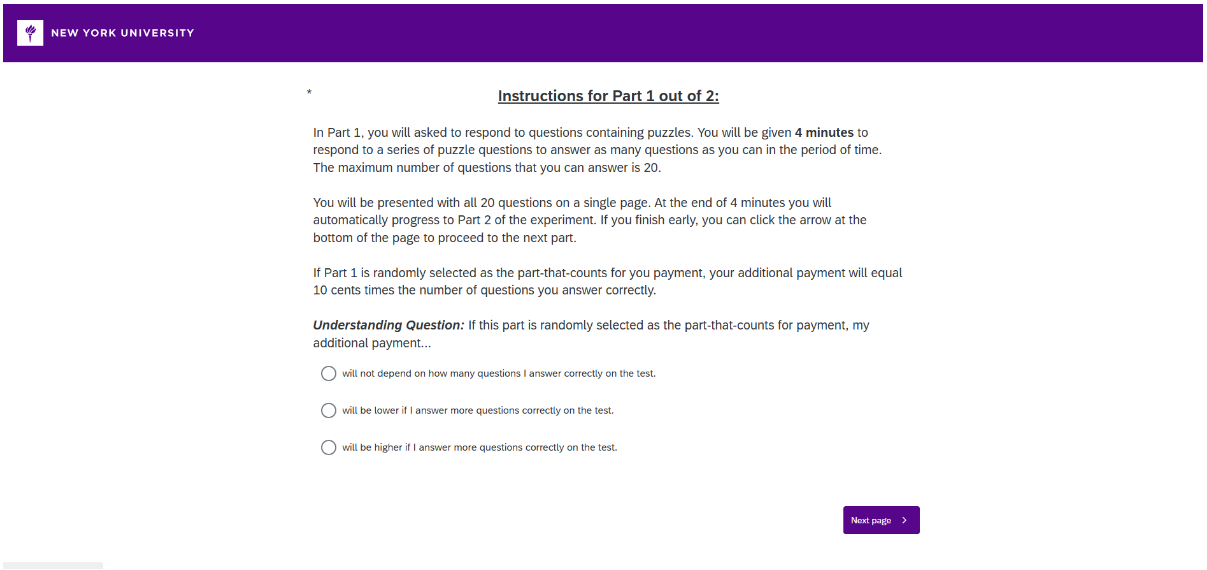
Your Payment: For completing this study, you will receive $3 as a completion payment. In addition, 1 part out of the 2 parts will be randomly selected as the part-that-counts for additional payments. Any amount you earn in the part-that-counts will be given to you as a bonus payment. All payment will be through Prolific, and you will receive the bonus payment within 24 hours of completing the study.

Understanding Question: Which of the following statements is true?

○ For completing this study, I will receive nothing.

○ For completing this study, I will receive $3 for sure, and I will have no chance of a bonus payment.

○ For completing this study, I will receive $3 for sure. In addition, I will receive any amount I earn in the part-that-counts as a bonus payment.



NEW YORK UNIVERSITY

* **Instructions for Part 1 out of 2:**

In Part 1, you will asked to respond to questions containing puzzles. You will be given **4 minutes** to respond to a series of puzzle questions to answer as many questions as you can in the period of time. The maximum number of questions that you can answer is 20.

You will be presented with all 20 questions on a single page. At the end of 4 minutes you will automatically progress to Part 2 of the experiment. If you finish early, you can click the arrow at the bottom of the page to proceed to the next part.

If Part 1 is randomly selected as the part-that-counts for you payment, your additional payment will equal 10 cents times the number of questions you answer correctly.

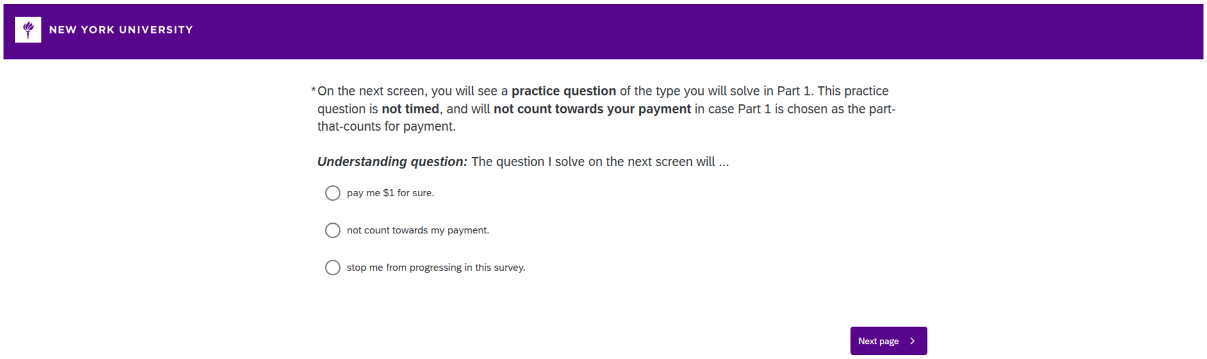
Understanding Question: If this part is randomly selected as the part-that-counts for payment, my additional payment...

○ will not depend on how many questions I answer correctly on the test.

○ will be lower if I answer more questions correctly on the test.

○ will be higher if I answer more questions correctly on the test.

Next page >



NEW YORK UNIVERSITY

*On the next screen, you will see a **practice question** of the type you will solve in Part 1. This practice question is **not timed**, and will **not count towards your payment** in case Part 1 is chosen as the part-that-counts for payment.

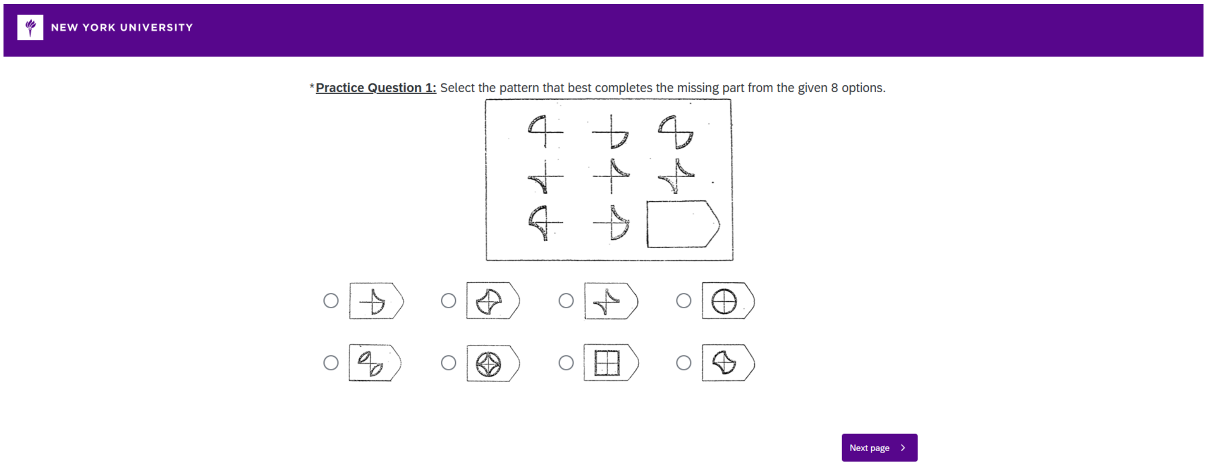
Understanding question: The question I solve on the next screen will ...

○ pay me $1 for sure.

○ not count towards my payment.

○ stop me from progressing in this survey.

Next page >



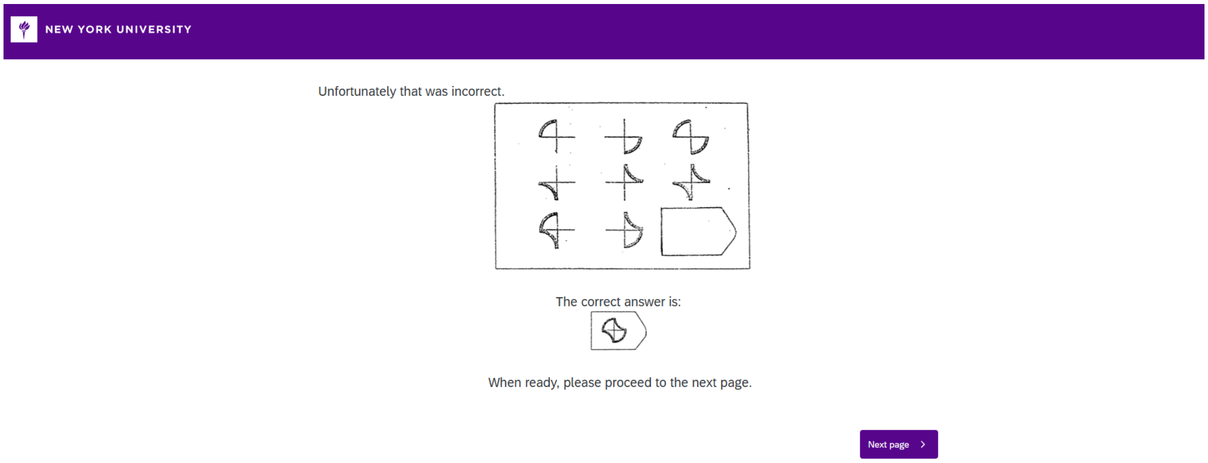
NEW YORK UNIVERSITY

***Practice Question 1:** Select the pattern that best completes the missing part from the given 8 options.

○ ○ ○ ○

○ ○ ○ ○

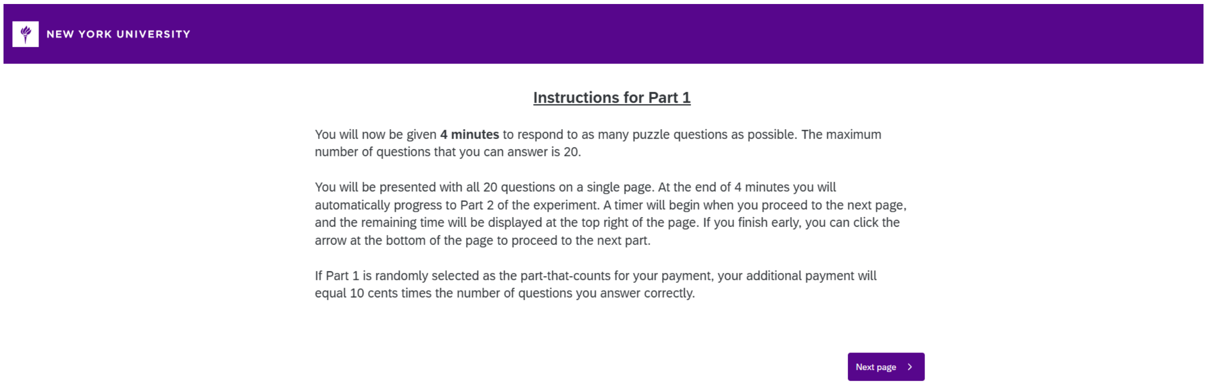
Next page >



Unfortunately that was incorrect.

The correct answer is:

When ready, please proceed to the next page.

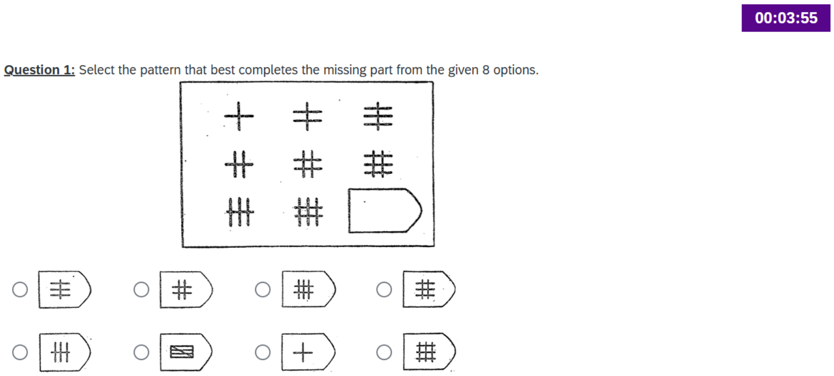


Instructions for Part 1

You will now be given **4 minutes** to respond to as many puzzle questions as possible. The maximum number of questions that you can answer is 20.

You will be presented with all 20 questions on a single page. At the end of 4 minutes you will automatically progress to Part 2 of the experiment. A timer will begin when you proceed to the next page, and the remaining time will be displayed at the top right of the page. If you finish early, you can click the arrow at the bottom of the page to proceed to the next part.

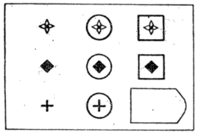
If Part 1 is randomly selected as the part-that-counts for your payment, your additional payment will equal 10 cents times the number of questions you answer correctly.



Question 1: Select the pattern that best completes the missing part from the given 8 options.

Question 2: Select the pattern that best completes the missing part from the given 8 options.

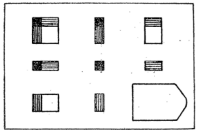
00:03:26



○ ○ ○ ○

○ ○ ○ ○

Question 3: Select the pattern that best completes the missing part from the given 8 options.

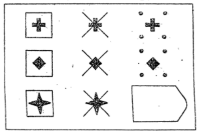


○ ○ ○ ○

○ ○ ○ ○

00:03:11

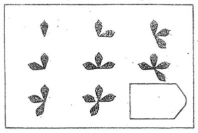
Question 4: Select the pattern that best completes the missing part from the given 8 options.

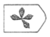


○ ○ ○ ○

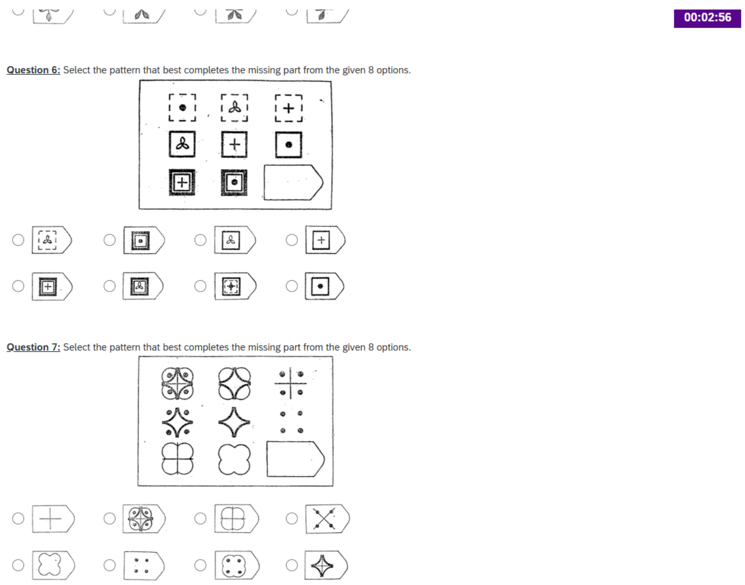
○ ○ ○ ○

Question 5: Select the pattern that best completes the missing part from the given 8 options.



○ ○ ○ ○ 

○ ○ ○ ○



00:02:56

Question 6: Select the pattern that best completes the missing part from the given 8 options.

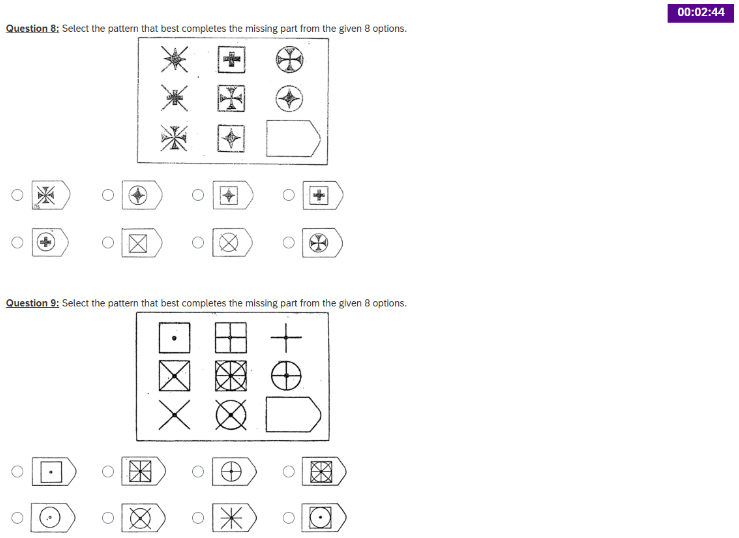
○ ○ ○ ○

○ ○ ○ ○

Question 7: Select the pattern that best completes the missing part from the given 8 options.

○ ○ ○ ○

○ ○ ○ ○



00:02:44

Question 8: Select the pattern that best completes the missing part from the given 8 options.

○ ○ ○ ○

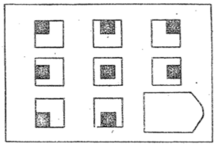
○ ○ ○ ○

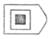
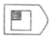
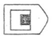
Question 9: Select the pattern that best completes the missing part from the given 8 options.

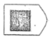
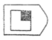
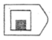
○ ○ ○ ○

○ ○ ○ ○

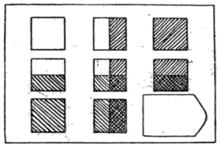
Question 10: Select the pattern that best completes the missing part from the given 8 options.

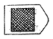
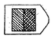
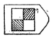
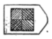


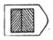
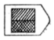
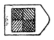
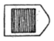
○ ○  ○  ○ 

○ ○  ○  ○ 

Question 11: Select the pattern that best completes the missing part from the given 8 options.

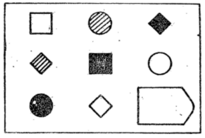


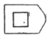
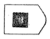
○  ○  ○  ○ 

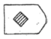
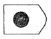
○  ○  ○  ○ 

Question 12: Select the pattern that best completes the missing part from the given 8 options.

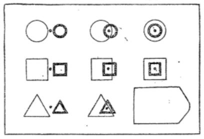
00:02:18

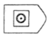


○  ○ ○  ○

○  ○ ○  ○

Question 13: Select the pattern that best completes the missing part from the given 8 options.

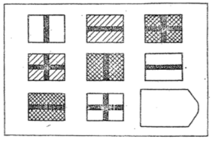


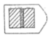
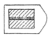
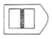
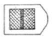
○ ○ ○ ○ 

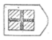
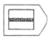
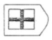
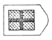
○ ○ ○ ○

00:02:07

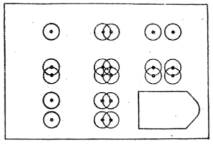
Question 14: Select the pattern that best completes the missing part from the given 8 options.

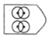
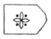


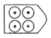
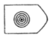
○  ○  ○  ○ 

○  ○  ○  ○ 

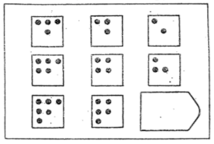
Question 15: Select the pattern that best completes the missing part from the given 8 options.

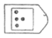
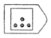
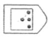
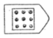


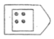
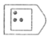
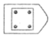
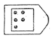
○ ○  ○  ○

○ ○  ○  ○

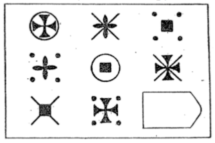
Question 16: Select the pattern that best completes the missing part from the given 8 options.

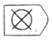
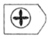
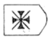
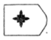


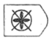
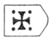
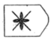
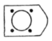
○  ○  ○  ○ 

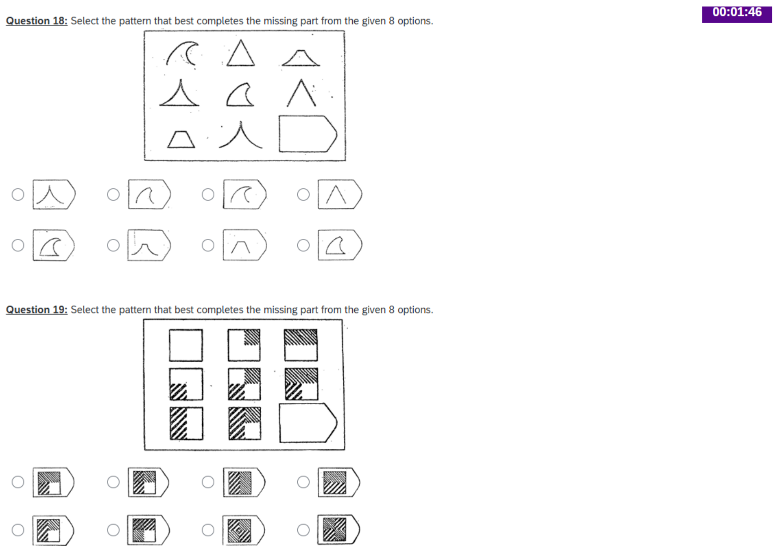
○  ○  ○  ○ 

Question 17: Select the pattern that best completes the missing part from the given 8 options.



○  ○  ○  ○ 

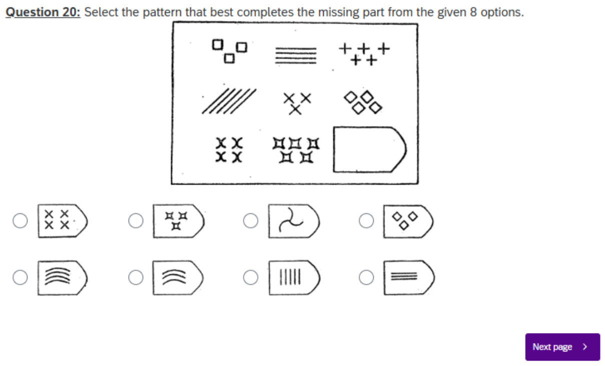
○  ○  ○  ○ 



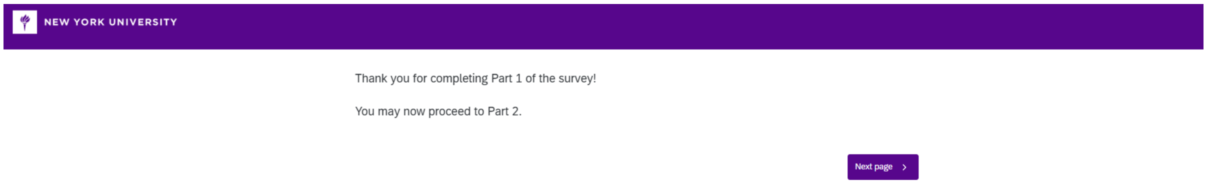
00:01:46

Question 18: Select the pattern that best completes the missing part from the given 8 options.

Question 19: Select the pattern that best completes the missing part from the given 8 options.

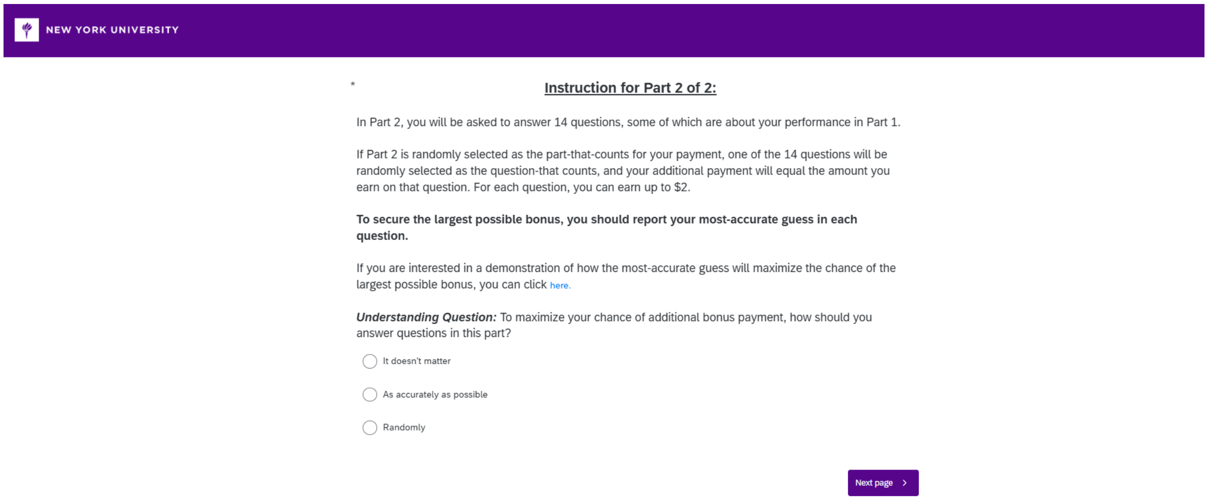


Question 20: Select the pattern that best completes the missing part from the given 8 options.



Thank you for completing Part 1 of the survey!

You may now proceed to Part 2.



* **Instruction for Part 2 of 2:**

In Part 2, you will be asked to answer 14 questions, some of which are about your performance in Part 1.

If Part 2 is randomly selected as the part-that-counts for your payment, one of the 14 questions will be randomly selected as the question-that counts, and your additional payment will equal the amount you earn on that question. For each question, you can earn up to $2.

To secure the largest possible bonus, you should report your most-accurate guess in each question.

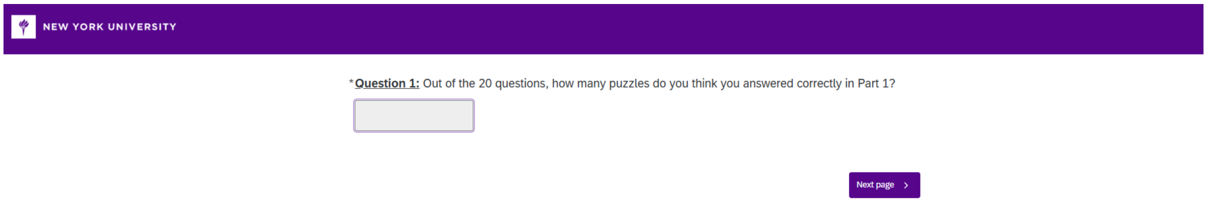
If you are interested in a demonstration of how the most-accurate guess will maximize the chance of the largest possible bonus, you can click here.

Understanding Question: To maximize your chance of additional bonus payment, how should you answer questions in this part?

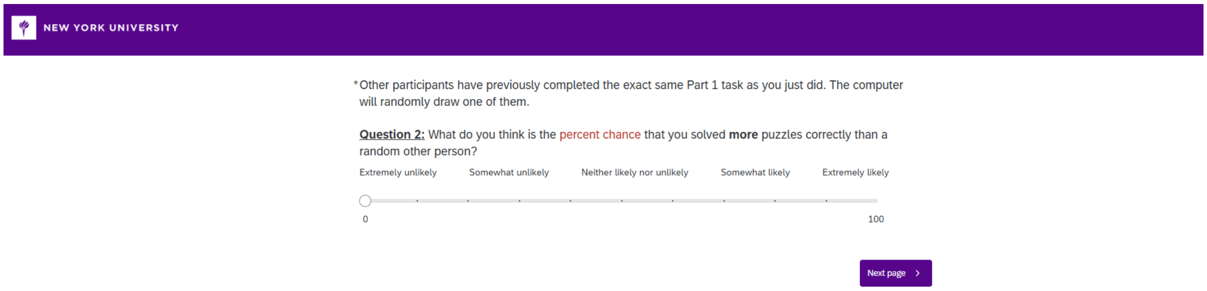
○ It doesn't matter

○ As accurately as possible

○ Randomly

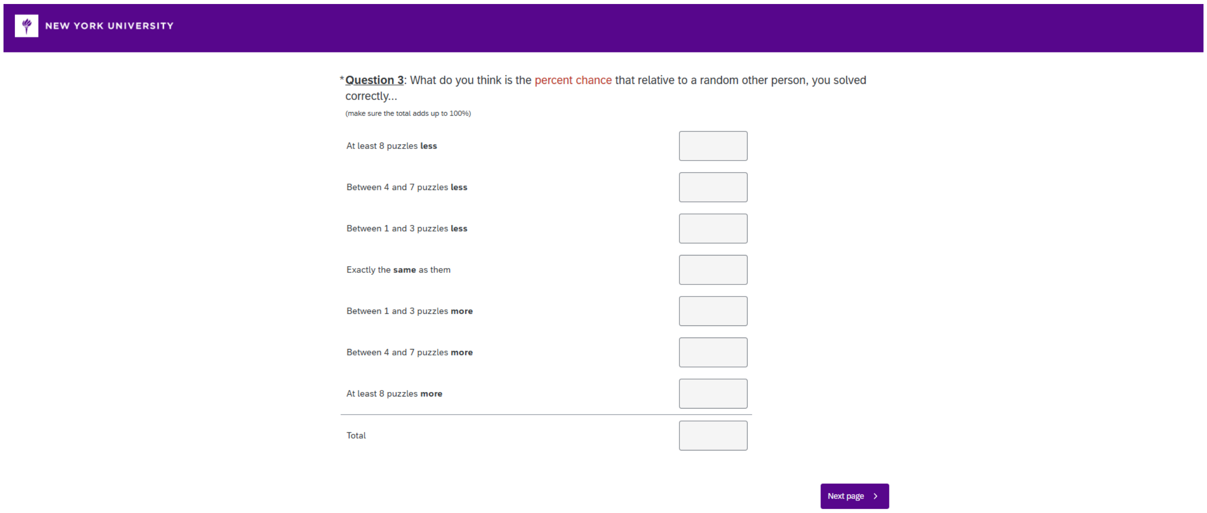


***Question 1:** Out of the 20 questions, how many puzzles do you think you answered correctly in Part 1?



*Other participants have previously completed the exact same Part 1 task as you just did. The computer will randomly draw one of them.

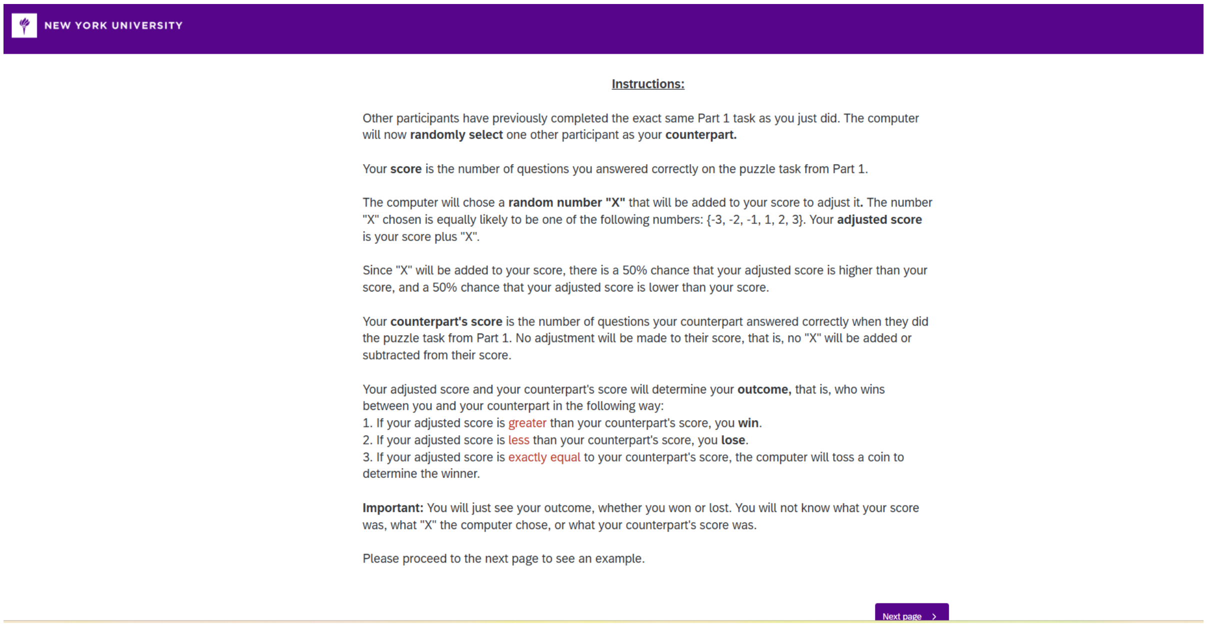
Question 2: What do you think is the percent chance that you solved **more** puzzles correctly than a random other person?



***Question 3**: What do you think is the percent chance that relative to a random other person, you solved correctly...

(make sure the total adds up to 100%)

At least 8 puzzles **less**	
Between 4 and 7 puzzles **less**	
Between 1 and 3 puzzles **less**	
Exactly the **same** as them	
Between 1 and 3 puzzles **more**	
Between 4 and 7 puzzles **more**	
At least 8 puzzles **more**	
Total	



Instructions:

Other participants have previously completed the exact same Part 1 task as you just did. The computer will now **randomly select** one other participant as your **counterpart.**

Your **score** is the number of questions you answered correctly on the puzzle task from Part 1.

The computer will chose a **random number "X"** that will be added to your score to adjust it**.** The number "X" chosen is equally likely to be one of the following numbers: {-3, -2, -1, 1, 2, 3}. Your **adjusted score** is your score plus "X".

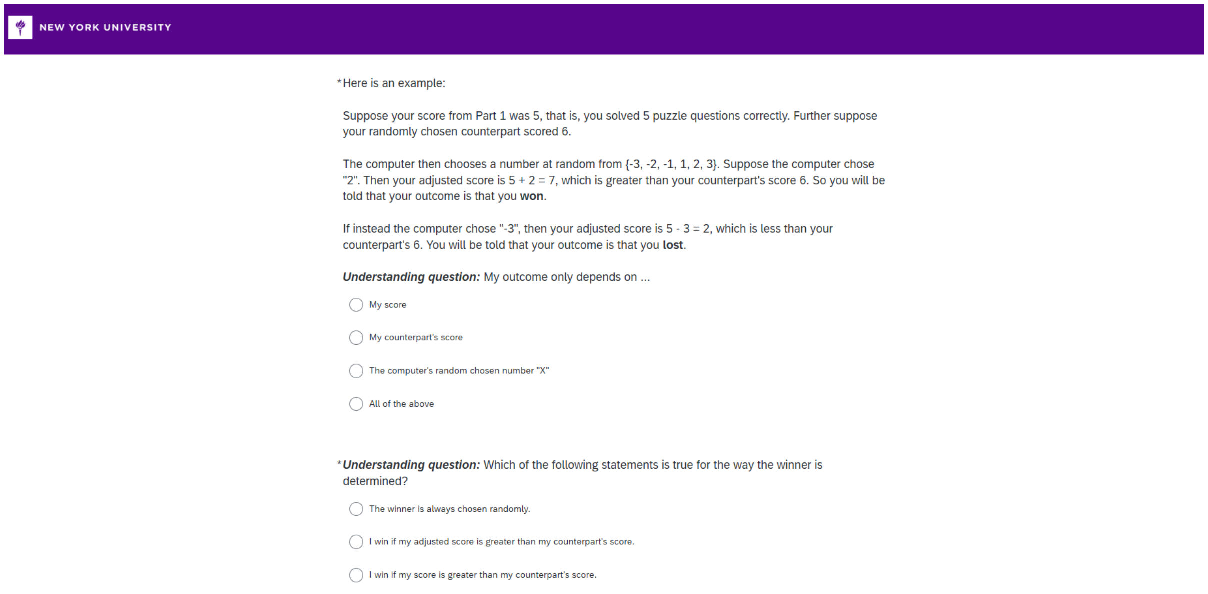
Since "X" will be added to your score, there is a 50% chance that your adjusted score is higher than your score, and a 50% chance that your adjusted score is lower than your score.

Your **counterpart's score** is the number of questions your counterpart answered correctly when they did the puzzle task from Part 1. No adjustment will be made to their score, that is, no "X" will be added or subtracted from their score.

Your adjusted score and your counterpart's score will determine your **outcome,** that is, who wins between you and your counterpart in the following way:
1. If your adjusted score is greater than your counterpart's score, you **win**.
2. If your adjusted score is less than your counterpart's score, you **lose**.
3. If your adjusted score is exactly equal to your counterpart's score, the computer will toss a coin to determine the winner.

Important: You will just see your outcome, whether you won or lost. You will not know what your score was, what "X" the computer chose, or what your counterpart's score was.

Please proceed to the next page to see an example.



*Here is an example:

Suppose your score from Part 1 was 5, that is, you solved 5 puzzle questions correctly. Further suppose your randomly chosen counterpart scored 6.

The computer then chooses a number at random from {-3, -2, -1, 1, 2, 3}. Suppose the computer chose "2". Then your adjusted score is 5 + 2 = 7, which is greater than your counterpart's score 6. So you will be told that your outcome is that you **won**.

If instead the computer chose "-3", then your adjusted score is 5 - 3 = 2, which is less than your counterpart's 6. You will be told that your outcome is that you **lost**.

Understanding question: My outcome only depends on ...

○ My score

○ My counterpart's score

○ The computer's random chosen number "X"

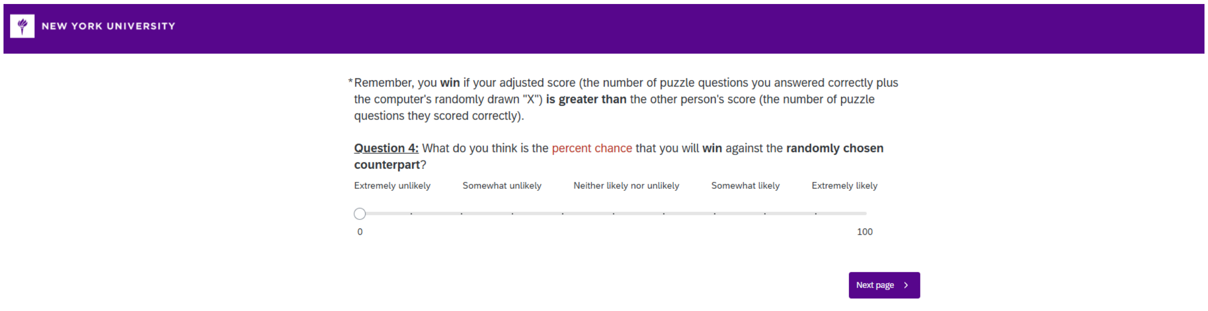
○ All of the above

*___Understanding question:___ Which of the following statements is true for the way the winner is determined?

○ The winner is always chosen randomly.

○ I win if my adjusted score is greater than my counterpart's score.

○ I win if my score is greater than my counterpart's score.

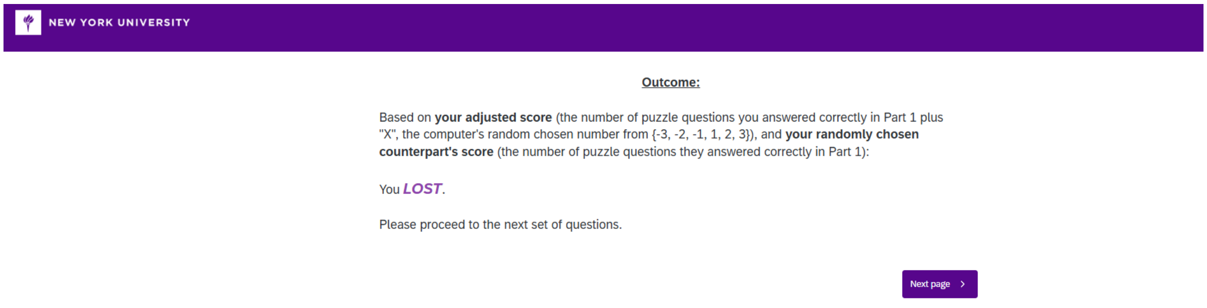


*Remember, you **win** if your adjusted score (the number of puzzle questions you answered correctly plus the computer's randomly drawn "X") **is greater than** the other person's score (the number of puzzle questions they scored correctly).

Question 4: What do you think is the percent chance that you will **win** against the **randomly chosen counterpart**?

Extremely unlikely | Somewhat unlikely | Neither likely nor unlikely | Somewhat likely | Extremely likely

0 100

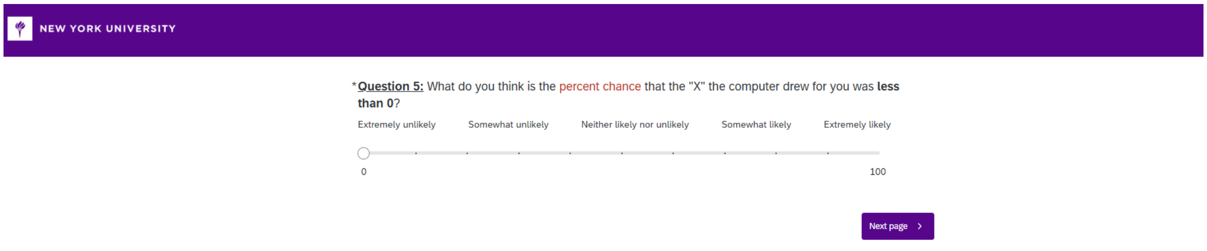


Outcome:

Based on **your adjusted score** (the number of puzzle questions you answered correctly in Part 1 plus "X", the computer's random chosen number from {-3, -2, -1, 1, 2, 3}), and **your randomly chosen counterpart's score** (the number of puzzle questions they answered correctly in Part 1):

You ***LOST***.

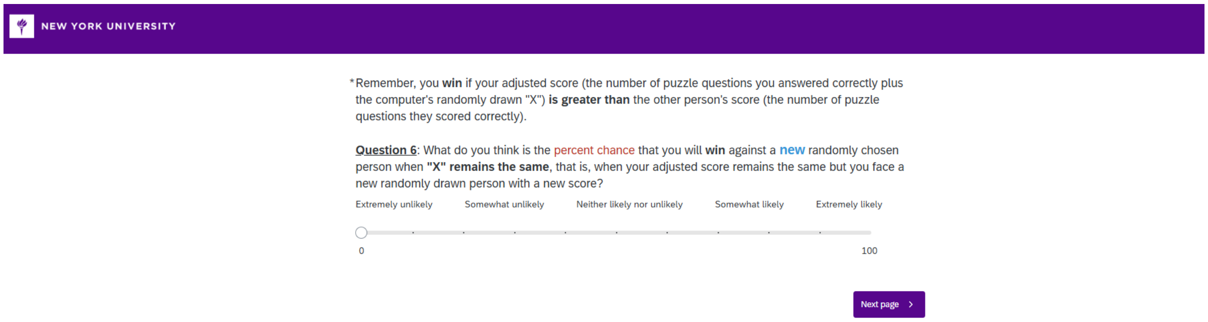
Please proceed to the next set of questions.



***Question 5:** What do you think is the percent chance that the "X" the computer drew for you was **less than 0**?

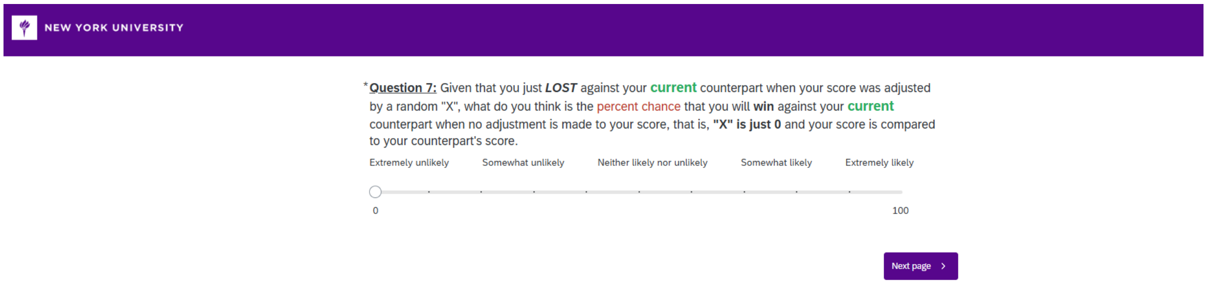
Extremely unlikely | Somewhat unlikely | Neither likely nor unlikely | Somewhat likely | Extremely likely

0 100

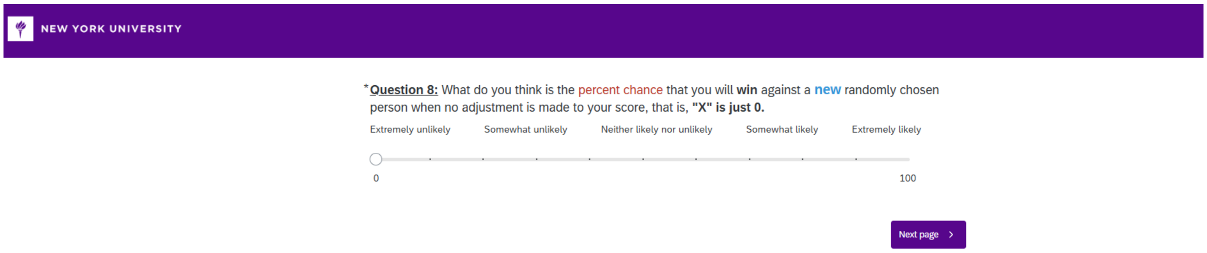


*Remember, you **win** if your adjusted score (the number of puzzle questions you answered correctly plus the computer's randomly drawn "X") **is greater than** the other person's score (the number of puzzle questions they scored correctly).

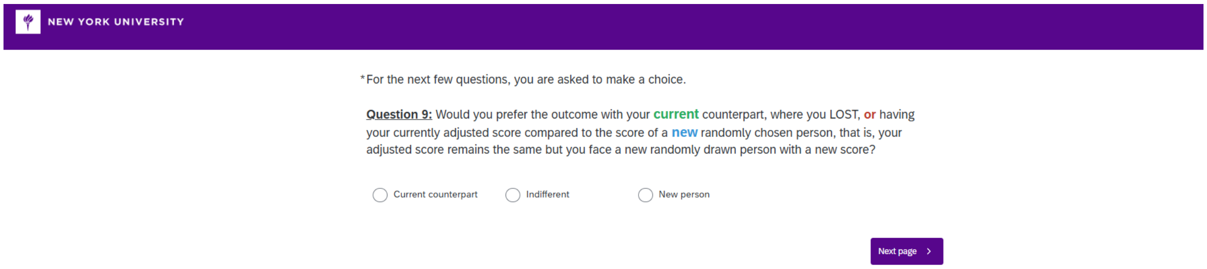
Question 6: What do you think is the percent chance that you will **win** against a **new** randomly chosen person when **"X" remains the same**, that is, when your adjusted score remains the same but you face a new randomly drawn person with a new score?



Question 7:** Given that you just ***LOST against your **current** counterpart when your score was adjusted by a random "X", what do you think is the percent chance that you will **win** against your **current** counterpart when no adjustment is made to your score, that is, **"X" is just 0** and your score is compared to your counterpart's score.



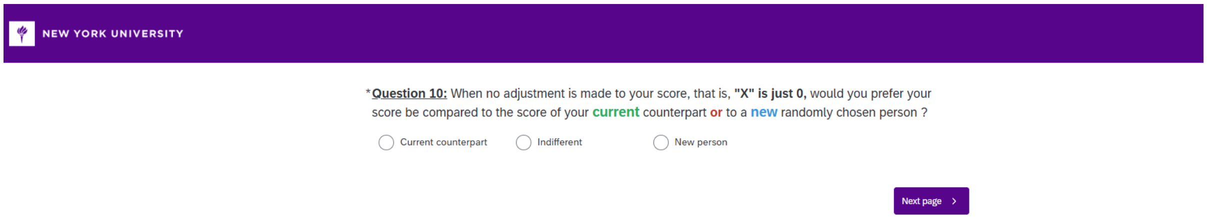
***Question 8:** What do you think is the percent chance that you will **win** against a **new** randomly chosen person when no adjustment is made to your score, that is, **"X" is just 0.**



*For the next few questions, you are asked to make a choice.

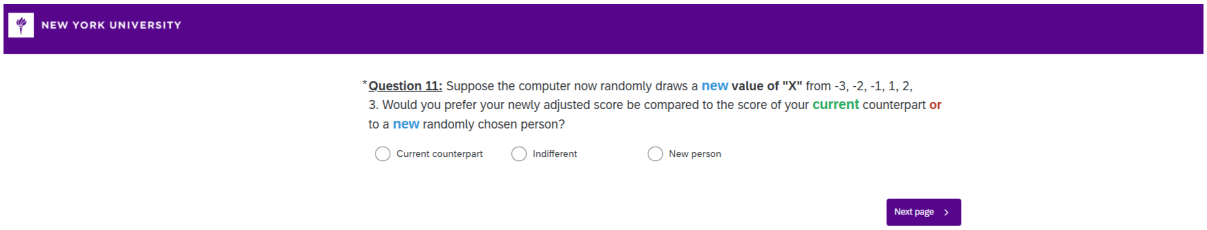
Question 9: Would you prefer the outcome with your **current** counterpart, where you LOST, **or** having your currently adjusted score compared to the score of a **new** randomly chosen person, that is, your adjusted score remains the same but you face a new randomly drawn person with a new score?

○ Current counterpart ○ Indifferent ○ New person



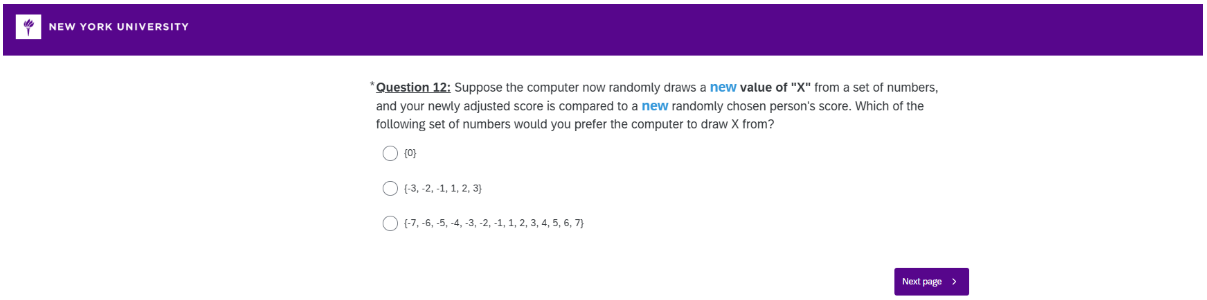
***Question 10:** When no adjustment is made to your score, that is, **"X" is just 0,** would you prefer your score be compared to the score of your **current** counterpart **or** to a **new** randomly chosen person ?

○ Current counterpart ○ Indifferent ○ New person



***Question 11:** Suppose the computer now randomly draws a **new value of "X"** from -3, -2, -1, 1, 2, 3. Would you prefer your newly adjusted score be compared to the score of your **current** counterpart **or** to a **new** randomly chosen person?

○ Current counterpart ○ Indifferent ○ New person

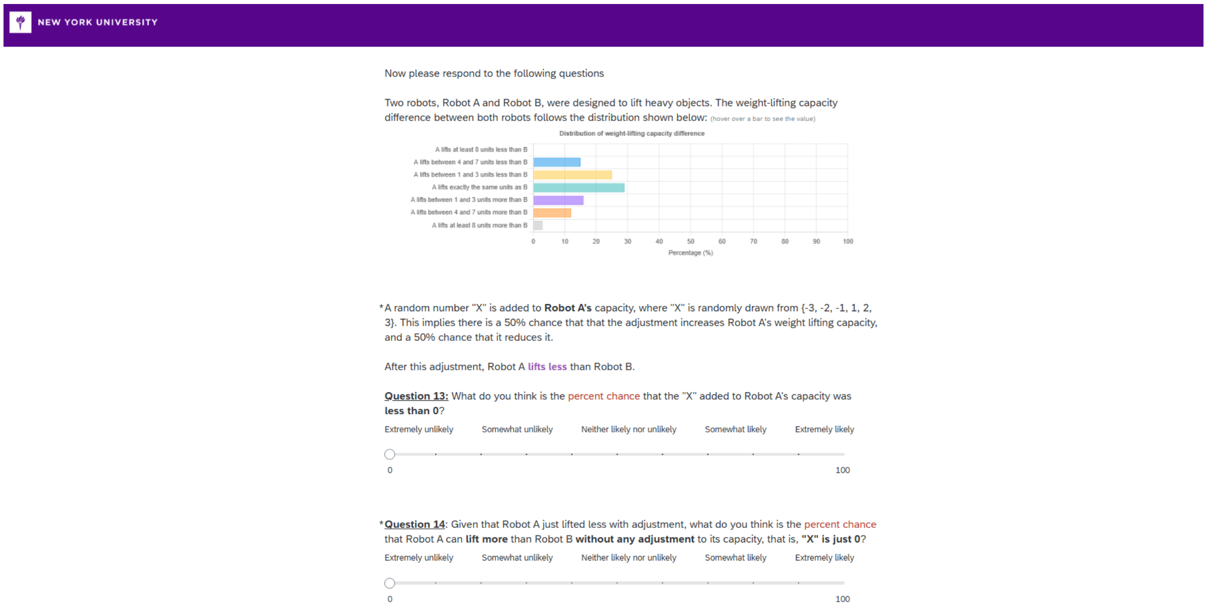


***Question 12:** Suppose the computer now randomly draws a **new value of "X"** from a set of numbers, and your newly adjusted score is compared to a **new** randomly chosen person's score. Which of the following set of numbers would you prefer the computer to draw X from?

○ {0}

○ {-3, -2, -1, 1, 2, 3}

○ {-7, -6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6, 7}



Now please respond to the following questions

Two robots, Robot A and Robot B, were designed to lift heavy objects. The weight-lifting capacity difference between both robots follows the distribution shown below: (hover over a bar to see the value)

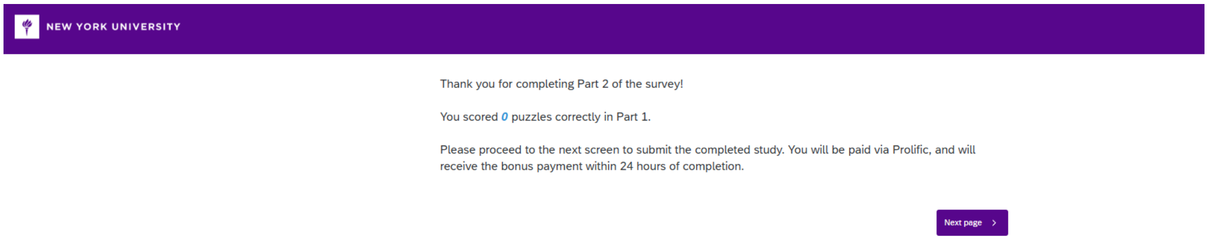
Distribution of weight-lifting capacity difference
A lifts at least 8 units less than B
A lifts between 4 and 7 units less than B
A lifts between 1 and 3 units less than B
A lifts exactly the same units as B
A lifts between 1 and 3 units more than B
A lifts between 4 and 7 units more than B
A lifts at least 8 units more than B
0
10
20
30
40
50
60
70
80
90
100
Percentage (%)

*A random number "X" is added to **Robot A's** capacity, where "X" is randomly drawn from {-3, -2, -1, 1, 2, 3}. This implies there is a 50% chance that that the adjustment increases Robot A's weight lifting capacity, and a 50% chance that it reduces it.

After this adjustment, Robot A lifts less than Robot B.

Question 13: What do you think is the percent chance that the "X" added to Robot A's capacity was **less than 0**?

***Question 14**: Given that Robot A just lifted less with adjustment, what do you think is the percent chance that Robot A can **lift more** than Robot B **without any adjustment** to its capacity, that is, **"X" is just 0**?



Thank you for completing Part 2 of the survey!

You scored ***0*** puzzles correctly in Part 1.

Please proceed to the next screen to submit the completed study. You will be paid via Prolific, and will receive the bonus payment within 24 hours of completion.