Stereotypes and Belief Updating

Prospectus for Oral Exam

Che Sun

November 2021

Biased Belief Updating and the Persistence of Stereotypes

Anujit Chakraborty
Che Sun [*]

November 28, 2021

Abstract

We study whether gender stereotype influences belief updating on stereotypical beliefs about others. Subjects perform belief updating tasks in a gender-stereotypical domain (math) and a stereotype-irrelevant, value neutral domain. We experimentally assign each subject a series 6 of i.i.d. signals that are noisy but informative. We observe their prior and elicit updated posteriors after they receive each new signal. Our experimental design allows for precise estimation of the parameters of the subjects' belief updating rule and a clean structural test against the Bayesian baseline. We contribute to the literature by uncovering a potential channel of stereotype persistence and experimentally identify this channel separately from the other heuristics and biases that result in non-Bayesian updating.

[*] The authors thank Professor Andrés Carvajal and the Department of Economics at UC Davis for funding support for this research.

1 Introduction

There has been substantial change in gender roles and shifts in societal culture around gender during the late 20th century and early 21st century in the United States. However, traditional gender stereotypes are still widespread and persistent in society today (Lueptow et al., 1995, 2001). Despite evidence to the contrary showing no significant difference in the average math ability of men and women (Hyde et al., 2008; Hyde and Mertz, 2009), the gender stereotype that men are better at math and science in women is still pervasive and forms in early childhood (Cvencek et al., 2011; Master, 2021). The male-math and male-science gender stereotypes have been show to be harmful to young girls and women in their motivation to pursue math and science, and make them doubt their own ability and performance in these traditionally male-dominant domains (Master, 2021; Coffman et al., 2019). The harmful effects of gender stereotypes could present challenges for our pursuit of gender equity in traditionally male-dominated fields, as well as our efforts to reduce the gender gap in labor market outcomes in these fields. A natural question is, then, why is the gender stereotype about math ability so resistant to change even though there is plenty of contrary evidence? Our study offers one possible explanation, namely people with gender stereotypical beliefs are more resistant to change in the face of new information which challenges that stereotype.

In conventional economic theory, agents are assumed to be fully rational. When presented with new information, they follow standard Bayesian updating rules when incorporating that information and updating their beliefs accordingly. However, research has shown that people exhibit a variety of heuristics and biases when processing new information which cause them to deviate from the standard assumption of Bayesian updating (Kahneman and Tversky, 1972, 1973; Tversky and Kahneman, 1974; Grether, 1980; El-gamal and Grether, 1995). There is a growing field of literature in economics studying the systematic cognitive mistakes people make while updating beliefs, notable examples of which include conservatism (Albrecht et al., 2013; El-gamal and Grether, 1995; Buser et al., 2018; Kovach, 2021), confirmatory bias (Rabin and Schrag, 1999; Eil and Rao, 2011; Charness and Dave, 2017; Grosjean and Modell, 2004), and asymmetric information processing such as the good news-bad news effect (Eil and Rao, 2011; Ertac, 2011). Unlike simple heuristics such as conservatism, one category of updating mistakes, motivated reasoning, are systematically driven by motivation and take place when updating beliefs in valenced areas. We hypothesize that gender stereotype can lead to biased belief updating by activating motivated reasoning if people are motivated to hold onto their preexisting stereotypes either consciously or subconsciously.

Reasoning and belief formation are often motivated by the goals people want to achieve (Kunda, 1990; Kunda and Sinclair, 1999; Bénabou and Tirole, 2002; Eil and Rao, 2011; Bénabou, 2015; Epley and Gilovich, 2016; Kahan et al., 2017). Kunda (1990) classified motivated reasoning into two categories: one where the motivation is to "arrive at an accurate conclusion", and one where the motivation is to "arrive at a particular, directional conclusion". When truth-seeking is not the only goal and motivation to achieve a particular conclusion comes into play, the set of cognitive processes involving the collection, evaluation, and storing of information can become biased, resulting in non-Bayesian belief updating and motivated beliefs that deviate from reality. Most of the research on motivated reasoning has been in the context of self-serving beliefs about self. In this project we investigate the role of motivated reasoning in the persistence of stereotypes about others by studying how people update stereotypical beliefs in the face of new information. We connect motivated reasoning to the higher resistance offered by stereotypical beliefs (as compared to non-stereotypical beliefs) when confronted with contrary evidence. We contribute to the literature by uncovering a potential channel of stereotype persistence and experimentally identify this channel separately from the other heuristics and biases that lead to non-Bayesian updating.

Our experimental design consists of two stages, the *Pre-Study* and the *Main Study*. In the pre-study, we recruited 136 subjects to complete a math test. The test questions cover algebra, geometry, elementary probability, and trigonometry, and they have a time limit of 15 minutes which is set to induce maximum variance in the scores. The test scores are incorporated into the main study and used to create the events for the belief updating task. We also collect information on the gender of the pre-study subjects, and assign each one of them a randomly generated ID number.

In the main study, we utilize a 2×2 between-subject treatment design, and study the updating behavior of subjects when they update beliefs in stereotypical domains (gender treatment) and non-stereotypical domains (ID treatment). To achieve this, we match to each main study subject two groups of 20 people that are randomly generated from the pre-study subject pool. In the gender treatment, Group A has equal numbers of men and women, and subjects are asked to report their beliefs on the probability of the following events about group A:

- $E_{A1,gender}$ = The top scorer in Group A is a man
- $E_{A2,gender}$ = The average score of the men in Group A is higher than the average

score of the women in Group A

This allows us to establish a baseline measurement of the existing gender stereotypes that subjects hold. In Group B, the relative composition varies and we elicit prior and updated posterior beliefs about the following event for Group B:

- $E_{B1,gender}$ = The top scorer in Group B is a man

The main updating tasks on event $E_{B1,gender}$ about group B includes 6 rounds of updating. After the elicitation of priors, we provide subjects with a series of i.i.d. noisy signals. We have 2 signal accuracy treatments to study whether updating mistakes are affected by the information content of signals. In the high informativeness treatment, signals have 75% accuracy. In the low informativeness treatment, signals have 56% accuracy. The signal structure is public information. We elicit the updated posterior belief of the same event after subjects receive each new signal. In the ID treatment, the design is analogous. When informing subjects of their matched groups, instead of gender, we reveal the composition of people with odd and even ID numbers in their groups. We elicit their beliefs on analogous events about the math performance of people with odd ID vs. people with even ID. The ID treatment serves as a control for baseline updating behavior in the absence of stereotypes, and comparing these two treatments allows us identify any systematic biases in belief updating that are activated by gender stereotypes.

We hypothesize that when confronted with new information, people might systematically overweight information that confirms their stereotypes and underweight information that contradicts their stereotypes. Our study measure and links motivated beliefs, biased updating procedures, and the persistence of stereotypes. We experimentally assigning each subject a series of informative signals that either support or contradict their stereotypical beliefs, which gives us key variation in identifying any asymmetry in information processing. By observing their priors, posteriors, and changes in beliefs repeatedly after they receive both kinds of signals, we can cleanly identify their updating procedures and whether stereotypical beliefs persist despite informative evidence that says otherwise.

There is a growing body of literature that studies biased belief updating that result from motivated reasoning. Experimental evidence has shown that when updating beliefs about own ability, people systematically favor good news and under-weigh bad news, the so-called good news-bad news effect Eil and Rao (2011); Mobius et al. (2014); Zimmermann (2020). This asymmetric pattern of belief updating can lead to persistent overconfidence. Bénabou and Tirole (2002) develop a model in which agents must weight the motivation to maintain self-serving beliefs with the risks of overconfidence, and shows that this can lead to systematic overconfidence. Ertac (2011); Coutts (2019) document the opposite phenomenon where people over-weigh negative information about themselves, which results in them becoming over-pessimistic. Motivated reasoning has also been shown to lead to political belief polarization (Kahan, 2013; Kahan et al., 2017; Taber and Lodge, 2006), and collective delusions in organizations and markets (Bénabou, 2013).

Stereotypes have been shown to affect information processing in the psychology literature. Bodenhausen and Wyer (1985) document subjects who seek out information to confirm their stereotypes, which led to differential recall of stereotype-consistent and stereotype-inconsistent information. Johnston (1996) provides similar experimental evidence that subjects employ an information seeking strategy that exhibit a stereotype preservation bias. Recent evidence in experimental economics has also shown that gender stereotypes can affect how people form and update beliefs about themselves. Coffman (2014) studies the effect of self-stereotyping on idea contribution and finds that conditional on measured ability, subjects are less willing to contribute ideas in areas that are not stereotypically in one's gender domain. Bordalo et al. (2019) show that subjects overestimate their own task performance in domains that are stereotypically associated with their own gender. They also overestimate other subjects' performance in areas associated with the other subjects' gender. Dustan et al. (2020) study the first- and second-order beliefs about performance on a timed math task by gender, and find that event though men and women's first order beliefs do no differ, their second order beliefs differ substantially in a way that is consistent with the existing gender stereotype, whereby the majority believe that most men believe men outscore women in the task. Coffman et al. (2019) study how gender stereotypes influence belief updating about own performance, and find that both men and women respond more to positive feedback in a task domain that is stereotypically associated with their own gender, than compared to when the positive feedback happens in a gender-incongruent domain. We provide the first experimental evidence on the effect of gender stereotype on belief updating about others. Our experimental designs allows us to structurally estimate the updating rule used by subjects, and provides a structural test against Bayesian updating. Our study measures and links gender stereotype to the activation of motivated reasoning, and identifies a potential channel through which gender stereotypes could persist.

The rest of the paper is organized as follows. Section 2 provides a structural framework of belief updating in our environment, which serves as a basis for our main empirical strategy. We also discuss violations of Bayesian updating resulting from widely documented heuristics and biases, and what they entail for the parameters of our framework. At the end of section 2, we provide the details of our empirical strategy. Section 3 provides a detailed discussion of our experimental design. Section 4 discusses the data from our pre-study and main study pilot samples. Section 5 provides some preliminary results from the main study pilot, and we outline some changes to the experimental design and future directions for research.

2 Theoretical Framework for Belief Updating

In this section, we present a simple theoretical framework for belief updating in our experiment. We start with the updating rule of a fully Bayesian agent, then modify the updating rule to allow for deviations from Bayesian updating. Bayesian updating has several testable properties that allow us to experimentally test the updating behavior of subjects against the Bayesian benchmark. The framework also serves as a basis for our empirical strategy, which can structurally estimate parameters of the updating rule. We discuss our empirical strategy in detail at the end of this section as well as possible cases of violations of Bayesian updating and what they entail for the parameters in the framework.

In our environment, subjects complete belief updating tasks on binary events, and they report subjective beliefs on the probability that a given event will happen. For example, in the gender treatment, the event of interest is $E_{B1,gender}$ = the top scorer in their matched 20-person group is a man.[1] In this example, there are two possible states of the world (outcomes of the event). Denote the state of the world by $\theta \in \Theta = \{0, 1\}$. $\theta = 1$ means the event of interest in fact did happen, i.e. the top scorer in the matched group is a man, and $\theta = 0$ means the event of interest did not happen, i.e. the top scorer in the matched group is a woman.[2] Denote the time period by $t \in T = \{0, 1, 2, 3, 4, 5, 6\}$. At time $t = 0$, subjects form their initial priors about the state of the world. From time $t = 1$ to $t = 6$, subjects receive a series of 6 i.i.d signals about the state and update their beliefs. We elicit their updated posterior belief after they observe each signal.

[1] In the main study, we randomly match 2 groups of 20 people with the subjects. These groups are randomly drawn from the subject pool who completed the math test. We elicit beliefs on several events about these 2 groups, both in the gender stereotypical domain and the non-stereotypical domain. For simplicity, we use this event to illustrate the framework in this section.

[2] This is the case of the gender treatment. In the ID treatment, and $\theta \in \Theta = \{O, E\}$ in the ID treatment. $\theta = 1$ means the top scorer in the group has odd ID, and $\theta = 0$ means the top scorer in the group has even ID.

We denote the Bayesian belief about state θ at time t as μ_t^θ and the corresponding subjective belief of agents as $\hat{\mu_t^\theta}$. When agents do not follow Bayesian updating, subjective beliefs can deviate from Bayesian beliefs. The signal received at time t is $s_t \in \{0, 1\}$. Signals are informative and have a 75% accuracy, which means that $P(s = 1|\theta = 1) = P(s = 0|\theta = 0) = 75\%$. Each signal is independently drawn from the same distribution.

After receiving a signal $s_t = 1$ at time t, the Bayesian posterior beliefs are

$$\begin{aligned} \mu_t^1 &= P(\theta = 1|s_t = 1) \\ &= \frac{P(s_t = 1|\theta = 1)\mu_{t-1}^1}{P(s_t = 1)} \end{aligned} \tag{1}$$

and

$$\begin{aligned} \mu_t^0 &= P(\theta = 0|s_t = 1) \\ &= \frac{P(s_t = 1|\theta = 0)\mu_{t-1}^0}{P(s_t = 1)} \end{aligned} \tag{2}$$

Dividing (1) by (2), we have

$$\frac{\mu_t^1}{\mu_t^0} = \frac{P(s_t = 1|\theta = 1)}{P(s_t = 1|\theta = 0)} * \frac{\mu_{t-1}^1}{\mu_{t-1}^0} \tag{3}$$

i.e. posterior likelihood ratio = signal factor × prior likelihood ratio. Taking natural logarithm of both sides of the equation, we have

$$logit(\mu_t^1) = \lambda_1 + logit(\mu_{t-1}^1) \tag{4}$$

where λ_1 is the log signal factor for $s_t = 1$, i.e. $\lambda_1 = log(\frac{P(s_t=1|\theta=1)}{P(s_t=1|\theta=0)})$. Analogously, when the signal is $s_t = 0$, we have

$$logit(\mu_t^1) = \lambda_0 + logit(\mu_{t-1}^1) \tag{5}$$

Therefore, we can rewrite Bayes' Rule in our environment as

$$logit(\mu_t) = \mathbb{I}((s_t = 1)\lambda_1 + \mathbb{I}((s_t = 0)\lambda_0 + logit(\mu_{t-1}) \tag{6}$$

where $\mathbb{I}((s_t = 1)$ is an indicator function that takes value 1 if the signal at time period t is "the top scorer in the group is a man", $s_t = 1$. We use a $logit(\mu_t)$ to denote the log likelihood ratio of the beliefs.

Mobius et al. (2014) point out that Bayesian updating satisfies 3 testable properties: invariance, sufficiency, and stability, which define the core structure of Bayesian updating. They are formally defined as follows:

Definition 1 (Invariance)**.** *A belief updating rule is invariant if it can be written as*

$$logit(\hat{\mu_t}) - logit(\hat{\mu_{t-1}}) = g_t(s_t, s_{t-1}, ...) \tag{7}$$

for a sequence of functions g_t that do not depend on μ_{t-1}.

Invariance means that the change in logit beliefs should only depend on the signals received and should not be affected by the priors $\hat{\mu_{t-1}}$. One notable case of cognitive bias that violates the invariance property is confirmation bias (Rabin and Schrag, 1999; Charness and Dave, 2017; Eil and Rao, 2011), where an agent over-weighs their priors and do not sufficiently incorporate information from the signals.

Given that the belief updating rule satisfies invariance, we can define sufficiency:

Definition 2 (Sufficiency)**.** *A belief updating rule satisfies sufficiency if the belief $\hat{\mu_{t-1}}$ is a sufficient statistic for all the signals received prior to time t.*

Intuitively, sufficiency is the idea that an agent's prior belief should have fully incorporated all previous signals up to that point in time, and therefore only new signals need to be taken into account when forming updated posteriors. Sufficiency implies that $g_t(s_t, s_{t-1}, ...) = g_t(s_t)$. Sufficiency could be violated if an agent follows a chunk-type updating rule, where the agent only updates beliefs once a certain number of signals have been received. A simple example is someone who only believes it is raining after seeing weather reports saying it is indeed raining from two different websites. Notice that the violation of invariance necessarily implies the violation of sufficiency.

Definition 3 (Stability)**.** *A belief updating rule is stable if $g_t = g$ for all t.*

Stability requires that the way an agent incorporates new information, i.e. the updating rule, should not change across time.

Following Mobius et al. (2014), we allow subjective beliefs to follow a generalized belief updating rule which nests Bayesian updating. The simplest version of the updating rule can be written as

$$logit(\hat{\mu_{it}}) = \delta logit(\hat{\mu_{t-1}}) + \beta_1 \mathbb{I}((s_{it} = 1)\lambda_1 + \beta_0 \mathbb{I}((s_{it} = 0)\lambda_0 + \epsilon_{it} \tag{8}$$

Coefficients β_1 and β_0 capture the responsiveness to the two types of signals, and ϵ_{it} is an idiosyncratic error in the updating process. Equation (8) forms the basis for our empirical strategy, whereby running this regression allows us to structurally estimate the parameters of the subjects' updating rule and compare it against the Bayesian benchmark. If invariance is satisfied, then $\delta = 1$. We can test stability by estimating (8) for all 6 rounds of belief updating and test whether the β coefficients are the same across different rounds. If subjects' updating satisfies stability, we can then pool data from all rounds to increase power. To test for sufficiency, we follow Mobius et al. (2014) and include lagged signal factors in the following regression:

$$\begin{aligned} logit(\hat{\mu_{it}}) &= \delta logit(\hat{\mu_{t-1}}) + \beta_1 \mathbb{I}((s_{it} = 1)\lambda_1 + \beta_0 \mathbb{I}((s_{it} = 0)\lambda_0 \\ &+ \sum_{\tau=1}^{t-1} \beta_{t-\tau}[\mathbb{I}((s_{i,t-\tau} = 1)\lambda_1 + \mathbb{I}((s_{i,t-\tau} = 0)\lambda_0] + \epsilon_{it} \end{aligned} \tag{9}$$

Sufficiency is satisfied if coefficients on the lagged signal factors $\beta_{t-\tau}$ are zero.

A widely documented bias in belief updating is conservatism, where people do not place enough weight on new information, which results in posterior beliefs that are too close to their priors (Mobius et al., 2014; Albrecht et al., 2013; El-gamal and Grether, 1995; Buser et al., 2018; Eil and Rao, 2011). If subjects exhibit conservatism, we should see $\beta_1 < 1$ and $\beta_0 < 1$ in (8). Since conservatism is a heuristic also commonly observed in non-valenced belief domains, we should also expect to see the same pattern for coefficients in the ID treatment.

Another common updating mistake is confirmatory bias, where people with a prior belief that favors one state of the world will over-update given signals that support their current hypothesis about the world, and under-update if the signal contradicts their current hypothesis (Rabin and Schrag, 1999; Charness and Dave, 2017). If subjects exhibit confirmatory bias, the β coefficients would depend on their priors. Specifically, if $\mu^{\theta}_{t-1} > 0.5$, then $\beta_\theta > \beta_{\theta'}$, where θ is a particular state of the world and θ' the alternative state in our world of binary events.

To test for differential patterns of belief updating in stereotypical vs. non-stereotypical domains, we estimate regression (8) for the gender and ID treatments and test for the null of equality of coefficients across treatments. There are two possibilities for stereotype activated updating biases. The first possibility is external-stereotype activated bias, where the widespread gender stereotype about superior math ability of men influence the updating of subjects, even if they themselves might not hold this stereotype. In this case, we hypothesize that subjects would exhibit asymmetry in updating in the gender stereotypical domain, where they over-weigh signals that confirm the gender stereotype and under-weigh signals that contradict the gender stereotype, regardless of priors. The second possibility is internal-stereotype activated bias, where the existing stereotypes subjects hold (priors) influence how they incorporate new information. This case can result in similar patterns in updating to confirmatory bias. We measure internal stereotype by eliciting subjects' belief on the top scorer being man in a randomly drawn 20-person group with equal numbers of men and women. We place a question at the very end of the survey asking them whether they think there exists a stereotype that men are better at math than women, which measures their perception of the external gender stereotype. These two measurements allows us to further test for heterogeneity in updating.

In addition, we investigate for heterogeneity in updating behavior by gender. Their has been experimental evidence that group identity influences belief updating, where subjects are overconfident about own-group members' performance on an intelligence test (Cacault and Grieder, 2019). In the case of men, the effect of group identification and stereotype-activated bias work in the same direction and would result in $\beta_1 > \beta_0$. In the case of women, they work in opposite directions: group identification could mean that women place more weight on signals that favor women, but stereotype-activated bias might mean that less weight is placed on signals that favor women. This horse race could therefore go one of two ways: $\beta_1 > \beta_0$ or $\beta_1 < \beta_0$, and the result is of empirical interest. There is also experimental evidence that women tend to update more conservatively than men, and are less confident about their own performance in gender-incongruent tasks (Coffman et al., 2019), which might lead to further heterogeneity based on gender.

We also study whether confidence in one's own beliefs is correlated to updating mistakes. In the demographics questionnaire at the end of the experiment, we inform the subjects that there is a mathematical formula (Bayes' Rule) that calculates the correct chances of the events after each performance report. We then ask them how many out of the 6 posterior beliefs do they think are within 5% of what the mathematical formula would prescribe. This gives us a measure of their confidence in their own updating ability and belief accuracy.

Identification of (8) and (9) using OLS suffers from two possible sources of bias. First, possible measurement errors in the beliefs reported by subjects will cause downward bias on OLS estimates. Second, as Mobius et al. (2014) pointed out, if there exists individual heterogeneity in the responsiveness to signals, the OLS estimate of δ will be downward biased due to correlation between $\hat{\mu_{t-1}}$ and $\beta_{i1} - \beta_1$ and $\beta_{i0} - \beta_0$ in the error term. We use an instrumental variable strategy to address bias of OLS. We use the varying group compositions to induce exogenous variation in the priors by exploiting the fact that statistically speaking, a higher proportion of men in the 20-person group will result in a higher chance of the top scorer in the group being a man (and vice versa). Therefore, we construct an instrument for each subject's prior logit belief using the percentage of men in the 20-person group matched to the subject.

3 Experimental Design

The experiment consists of 2 stages, the *Pre-Study* and the *Main Study*. The details of each stage are explained below. We employ two separate subject pools for the pre-study and the main study, the *task workers* and the *evaluators*, respectively. In the pre-study, task workers complete a math test under time limit and a demographics questionnaire. Each subject receives a score for their performance on the math test, and this information is incorporated into the Main Study updating task.

The main study utilizes a 2×2 between-subject design. There are two signal accuracy treatments, a high informativeness treatment with 75% signal accuracy, and a low informativeness treatment with 56% accuracy. Within each signal accuracy treatment, we randomize subjects between a gender treatment and an ID treatment. In the gender treatment, subjects complete updating tasks about the math performance of men vs. women. The ID treatment serves as a control for baseline updating behavior in the absence of stereotypes, where subjects complete updating tasks about the math performance of people with odd ID numbers and even ID numbers (the ID numbers are randomly assigned and subjects are aware of this).

In the main study, each subject is matched with 2 randomly generated 20-person groups from the Pre-Study subject pool. Group A has the same number of men and women, and we elicit subjects' beliefs about the math performance of men vs. women (people with odd ID vs. people with even ID in the ID treatment) in Group A to establish a baseline measurement of their existing gender stereotypes. Group B has varying relative compositions, subjects complete updating tasks on this group. Specifically, they perform 6 rounds of belief updating tasks on a binary event, namely the top scorer in their matched group being a man (person with odd ID). We start by eliciting their prior beliefs, then provide them with a series of 6 independently and identically distributed signals, each with 75% accuracy. We collect their updated posterior after each round of signal provision.

The experiment is being conducted on Amazon Mechanical Turk (MTurk) and the surveys are programmed using the Qualtrics XM suite. We have finished data collection for the pre-study with a sample of 104 subjects, and a pilot experiment for the main study with 53 subjects and 75% signal accuracy.

3.1 Pre-Study

In the pre-study, subjects completed an online survey which included a timed math test and a demographics questionnaire. The math test included 20 questions and subjects were given 15 minutes to answers as many questions as possible. The questions vary in difficulty, and the time limit is designed to maximize variance of the scores. We created the questions based on sample questions from the ACT math test.[3] The questions cover algebra, geometry, trigonometry, and elementary probability theory. To prevent cheating, we changed the numbers in the ACT sample questions and computed the new answers ourselves, so that subjects could not directly find answers from searching on the internet. Calculators are allowed and they are encouraged to have a calculator as well as pen and paper ready before beginning the math questions. The score on the test is calculated as the number of correct answers.

After participants complete the math test, they are asked to complete a demographics questionnaire which collects information on their age, gender, race, ethnicity, education, and political party affiliation.

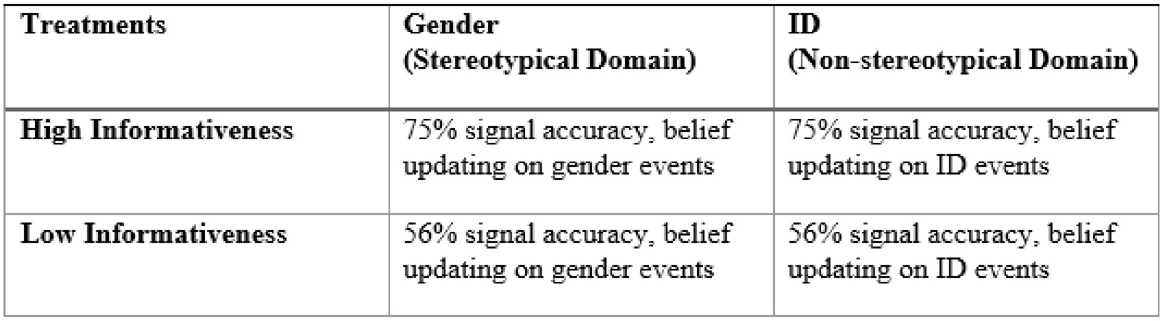
Subjects receive a flat participation fee of $1 for completing the survey and a bonus payment that equals to their score multiplied by $0.40 in order to incentivize effort on the math test. We recruited 136 MTurk workers for our subject pool. After excluding people who failed attention checks, we retain a sample of 103 subjects for the pre-study.

3.2 Main Study

In the main study, we investigate the updating behavior of subjects in the domain of stereotypical and non-stereotypical beliefs. We first elicit subjects' prior beliefs about an event that is either related or unrelated to the gender stereotype about math ability, then let subjects complete 6 rounds of updating tasks. In each updating round, we provide an informative signal and elicit posterior beliefs about the same event after the new information is presented.

Our main study utilizes a 2 × 2 treatment design. We randomize subjects between two treatments with differing signal informativeness. The high informativeness treatment provides signals that have 75% accuracy, whereas the low informativeness treatment

[3] We based our questions on the sample questions from these ACT math practice tests: `https://www.act.org/content/act/en/products-and-services/the-act/test-preparation/math-practice-test-questions.html?page=0&chapter=0`



Treatments	Gender (Stereotypical Domain)	ID (Non-stereotypical Domain)
High Informativeness	75% signal accuracy, belief updating on gender events	75% signal accuracy, belief updating on ID events
Low Informativeness	56% signal accuracy, belief updating on gender events	56% signal accuracy, belief updating on ID events

Figure 1: Treatment summary of the main study

provides signals that have 56% accuracy.[4] We vary the signal informativeness to study whether information content in signals is related to the level and pattern of updating mistakes. Within each signal informativeness treatment we randomize subjects between a gender treatment and an ID treatment. In the gender treatment, subjects update about the math test performance of men vs. women, a binary event in the domain of gender stereotype about math ability. In the ID treatment, subjects update about the math performance of subjects with a randomly assigned odd ID number vs. even ID number, a binary event unrelated to gender stereotypes. We can observe the baseline updating behavior in the absence of gender stereotypes in the ID treatment. Comparing updating behavior between these two treatments allows us to identify any systematic difference in updating patterns that is activated by gender stereotypes. Figure 1 provides a brief summary of the treatments in the main study.

Subjects receive a participation fee of $3 for completing the survey. They also have the chance to earn a bonus of $5, which is determined by one of the belief elicitation questions which is randomly selected at the end of the survey. The bonus is calculated using the BDM incentive scheme.

3.2.1 Events

Each subject is informed that they have been matched with two randomly selected groups (Group A and Group B) of 20 people from the pool of people (task workers) who com-

[4] More accurately, the signal accuracy is $\frac{5}{9}$ in the low informativeness treatment. We designed this so that it is easier to explain the signal accuracy and can help subjects better understand it. In the instructions we tell subjects that each performance report (signal) is produced by a randomly selected computer out of 9 computers. Out of these 9 computers, 5 are glitched and always produce an incorrect report, and the rest 4 computers always produce the true report.

pleted the math test, and that the test scores of everyone in their matched groups have been recorded. We also tell subjects that the survey has 2 parts. In part 1 they will be answering questions that ask them to guess the probability of certain events happening for Group A, in part 2 Group B. The composition of the groups are revealed to the subject. In the gender treatment, we reveal the number of men and women in the 20-person groups. In the ID treatment, we tell subjects that the people who completed the math test were assigned a random ID number, and we reveal the number of people with odd IDs and even IDs in their matched groups.

In the gender treatment, we elicit subjects' beliefs about the probability of the following events for Group A:

- $E_{A1,gender}$ = The top scorer in Group A is a man
- $E_{A2,gender}$ = The average score of the men in Group A is higher than the average score of the women in Group A

And we elicit prior and updated posterior beliefs about the following event for Group B:

- $E_{B1,gender}$ = The top scorer in Group B is a man

In the ID treatment, we elicit beliefs about the following events for Group A:

- $E_{A1,ID}$ = The top scorer in Group A has odd ID
- $E_{A2,ID}$ = The average score of the people with odd ID in Group A is higher than the average score of the people with even ID

And the prior and posterior beliefs about the following event for Group B:

- $E_{B1,ID}$ = The top scorer in Group B has odd ID

The main updating tasks are done on Group B events.

All of the events are binary, which allows us to capture with a single number the subjects' entire probability distribution over possible states of the world (e.g. the top scorer is a man or the top scorer is a woman). This greatly simplifies the belief elicitation procedure and allows for an incentive compatible way of belief elicitation. Another significant benefit of imposing binary events is that it enables clean identification of updating procedure used by subjects and a clear comparison with the null of Bayesian updating.

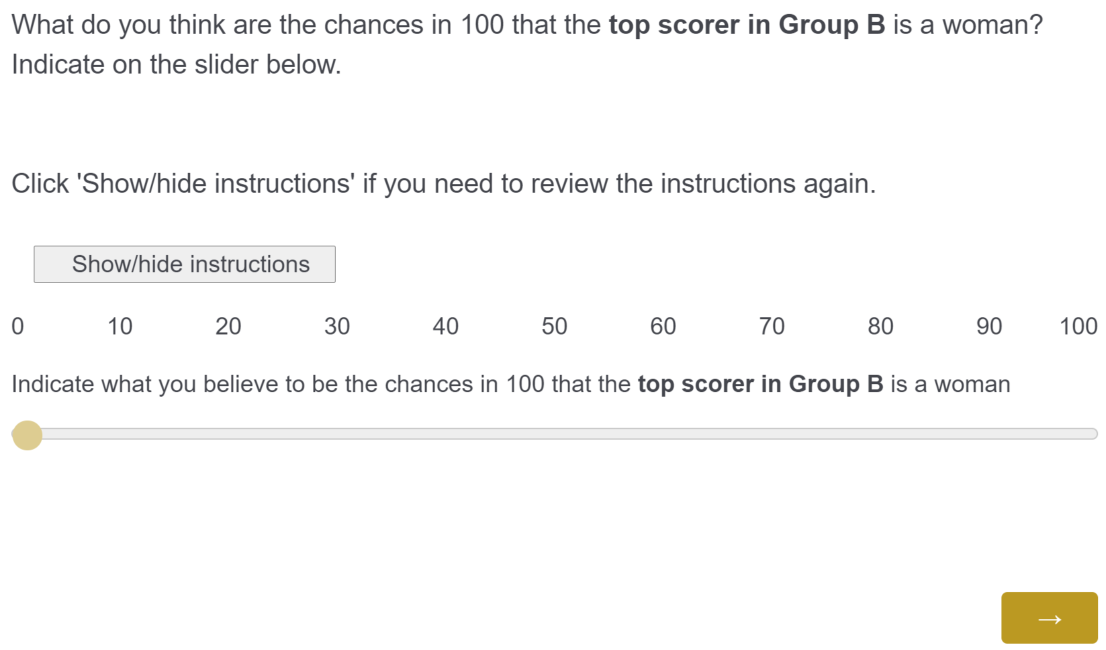
In Group A, the relative composition is fixed and always balanced, with 10 men (people with odd ID) and 10 women (people with even ID). In Group B, the relative composition varies. Varying the composition of Group B induces exogenous variation in the subjects' priors, since statistically speaking, a group with more men than women is more likely to have a man as the top scorer (and vice versa). We use the proportion of men in the group to construct an instrument for the prior on $E_{B1,gender}$ and employ an IV strategy to account for OLS bias due to possible individual heterogeneity in the responsiveness to signals. Having the same number of men and women in Group A allows us to observe subjects' prior beliefs without exogenous variation induced by changes in group composition, and they serve as a baseline measure of subjects' existing stereotypes. Suppose a subject's belief about the probability of $E_{A1,gender}$ is greater than 50%, then we classify the subject as having a stereotype favoring men, and vice versa. Eliciting beliefs about both the average performance and the top performance captures stereotypes about the mean and the right tail of the distribution.

We randomize the language of the gender event to avoid activating the implicit association between math and men, which might induce subjects to shift their beliefs in favor of men. Subjects in the gender treatment are randomized between two versions of each gender event. For example, the two versions of $E_{A1,gender}$ are "The top scorer in Group A is a man," and "The top scorer in Group A is a woman."

One concern with the gender treatment is that since subjects are seeing a series of belief elicitation questions about math performance by gender, they might realize the purpose of the experiment and report less truthfully. To make it harder for subjects in the gender treatment to guess the true purpose of the experiment, we include 3 belief elicitation tasks about the math performance by age in the two groups. Specifically, we tell subjects the number of people aged 35 and above and the number of people younger than 35 in Group A and B. After elicitation of $E_{A1,gender}$ and $E_{A2,gender}$, we elicit beliefs about the following two events about Group A: "The top scorer in Group A is aged 35 or above," and "The average score of the people aged 35 or above is higher than the average score of the people younger than 35 in Group A." After the 6 rounds of updating tasks about Group B, we elicit subjects' belief about the following event: "The top scorer in Group A is aged 35 or above".

3.2.2 Belief Elicitation

We use an adapted version of the Becker-DeGoot-Marschack (BDM) mechanism (Becker et al., 1964) as in (Grether, 1980; Karni, 2009) to elicit subjects' beliefs about the prob-



What do you think are the chances in 100 that the **top scorer in Group B** is a woman? Indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

Indicate what you believe to be the chances in 100 that the **top scorer in Group B** is a woman

Figure 2: User interface for eliciting the prior belief for event $E_{B1,gender}$ in group B

ability of events. Subjects are asked to report the chances out of 100 of an event (e.g. the top scorer in Group B is a man) happening in each of the belief elicitation questions. Subjects use a graphical slider to report their beliefs, which can be any whole number between 0 and 100, as shown in Figure 2.

At the end of the survey, one of the elicitation questions is randomly selected to determine the bonus payment. Azrieli et al. (2018) show that paying for one randomly selected question (they term it the Random Problem Selection mechanism, or RPS) is "essentially the only incentive compatible mechanism".

After a question is selected for payment, a random number N between 1 and 100 is first generated. If N is less than or equal to the subject's reported probability p in this selected question, the subject is paid by an "event lottery", and the subject receives the bonus if the event occurs. If N is greater than p, the subject is paid by a "number lottery" with an N in 100 chance of receiving the bonus. The idea of the number lottery is conveyed to the subject by telling them that a second number between 1 and 100 will be generated if they enter into the number lottery. If this second number is less than or equal to N, they receive the bonus, otherwise they receive $0.

The BDM procedure for eliciting event probability is favorable because it is incentive compatible for arbitrary risk preferences. Another commonly used belief elicitation method, the Quadratic Scoring Rule (QSR), is only incentive compatible under the assumption of risk neutrality in comparison. The only assumption needed to ensure incentive compatibility for this particular implementation of BDM is that subjects never choose dominated random payments[5] (Karni, 2009; Healy, 2020). Thus, BDM is incentive compatible under expected utility, rank-dependent utility, and prospect theory (Trautmann and van de Kuilen, 2015). In addition, the BDM procedure to elicit probabilities has been shown to outperform QSR (Hollard et al., 2010; Holt and Smith, 2016).

A possible alternative to the BDM method is the two-stage Synchronized Lottery Choice Menu mechanism (abbreviated as LC) proposed by Holt and Smith (2016). The LC method is theoretically equivalent to the BDM and satisfies incentive compatibility under the same assumptions. In the two-stage synchronized choice menu, subjects are sequentially presented with a menu of two alternatives, the event lottery and the number lottery, and are asked to pick which option they prefer. In the first stage menu, subjects make choices between the event lottery and a coarse grid of number lotteries.[6] After they make their choices, a second stage finer choice menu is presented at their switching point with finer intervals of the number lottery.[7] The switching point in the second stage menu is interpreted as the subjective probability of the event.

We opt not to use the two-stage synchronized lottery choice menu method because it significantly increases the task load as well as cognitive load of subjects. Compared to BDM, which only asks for subjects to report a single number, the LC method requires subjects to make 20 distinct choices for each belief elicitation question, which is a 20-fold increase on the task load. Given the large number of elicitation questions in our survey, the lottery choice menu method is simply not practical. It is also difficult to produce instructions that are easy to understand for subjects for the LC method in our context, and the choice menu could further confuse subjects, resulting in noisier data.

One potential downside to BDM is that its incentive information could be difficult to understand (Cason and Plott, 2014; Holt and Smith, 2016). To mitigate this, we tell subjects that they can maximize their chance of winning the bonus by answering as

[5] More accurately, the assumptions needed for incentive compatibility are that agents' preference relation over the set of simple acts and lotteries exhibit probabilistic sophistication and dominance (Machina and Schmeidler, 1995), and that they have no stakes in the realization of the event apart from the incentive payments from the elicitation. See Karni (2009) for details.

[6] E.g. 0% chance of winning \$5, 10% chance of winning \$5, 20% chance of winning \$5, etc.

[7] For example, if a subject prefers the event lottery to a 40% number lottery but prefers the 50% number lottery, then the second stage menu will present choices between the event lottery and number lotteries with 1% intervals, from 41% chance of winning \$5, 42% chance of winning \$5, ..., to 50% chance of winning \$5.

accurately as possible, and explicitly allows them to skip viewing the incentive information. Omitting the incentive information has been shown to improve truth-telling in the context of the Binarized Scoring Rule (BSR) (Danz et al., 2020). If a subject chooses to view the incentives, we show them a version of the BDM incentives modeled after those used in Holt and Smith (2016), which conveys the idea in terms of an event lottery and a number lottery, and tell subjects that whichever lottery that gives the highest chances of the payoff will be used. The full incentive information we use is provided in the Appendix B.

3.2.3 Signals and Updating

After subjects complete belief elicitation questions about Group A in part 1 that measure their baseline stereotypes, they proceed the part 2 in which they complete the main updating tasks about Group B. Subjects are informed that they are matched with Group B, which is a randomly selected 20-person group from the pool of people who previously completed the math test, and that they will be answering questions that ask them to predict the chances of certain events about this group happening. The composition of Group B is presented to the subject again. In the gender treatment, we reveal the number of men and women, as well as the number of people aged 35 or above and the number of people younger than 35 among the 20 people in Group B. In the ID treatment, we inform subjects that the people who completed the math test were each assigned a randomly generated ID number. We then reveal the number of people with odd ID numbers and the number of people with even ID numbers among the 20 people in their matched group.

After reminding the subjects of the composition of Group B, we elicit subjects' prior belief about the probability of event $E_{B1,gender}$, the top scorer in Group B is a man[8], in the gender treatment, and $E_{B1,ID}$, the top scorer in Group B has odd ID, in the ID treatment. After the prior belief elicitation, each subject is presented with a series of 6 i.i.d. noisy signals that have 75% accuracy in the high informativeness treatment and 56% accuracy in the low informativeness treatment. For simplicity, we will discuss the 75% signal accuracy treatment hereafter.

Each signal is binary and reports one of the two possible outcomes of the binary event (the states of the world). For example, in the gender treatment, the event of interest is "the top scorer in Group B is a man". Therefore, if the true state is that the top scorer

[8] We randomize the language in the elicitation questions in the gender treatment between two versions, the top scorer is a man, and the top scorer is a woman. For simplicity, we use one version of the elicitation language in this section.

in Group B is a man, then with 75% probability the signal says "the top scorer is Group B is a man", and with 25% probability the signal says "the top scorer in Group B is a woman". Each signal is independently drawn and therefore the signals are independent conditional on the true state. The signal generating process is explicitly explained to subjects and therefore is public information. The signal structure produces key variation in the type of signals subjects encounter, i.e. signals that confirm or contradict their stereotypes, which allows us to identify any asymmetry in updating behavior. The setup of binary states and binary signals offer the advantage of clean identification and is also easier to for subjects to understand.

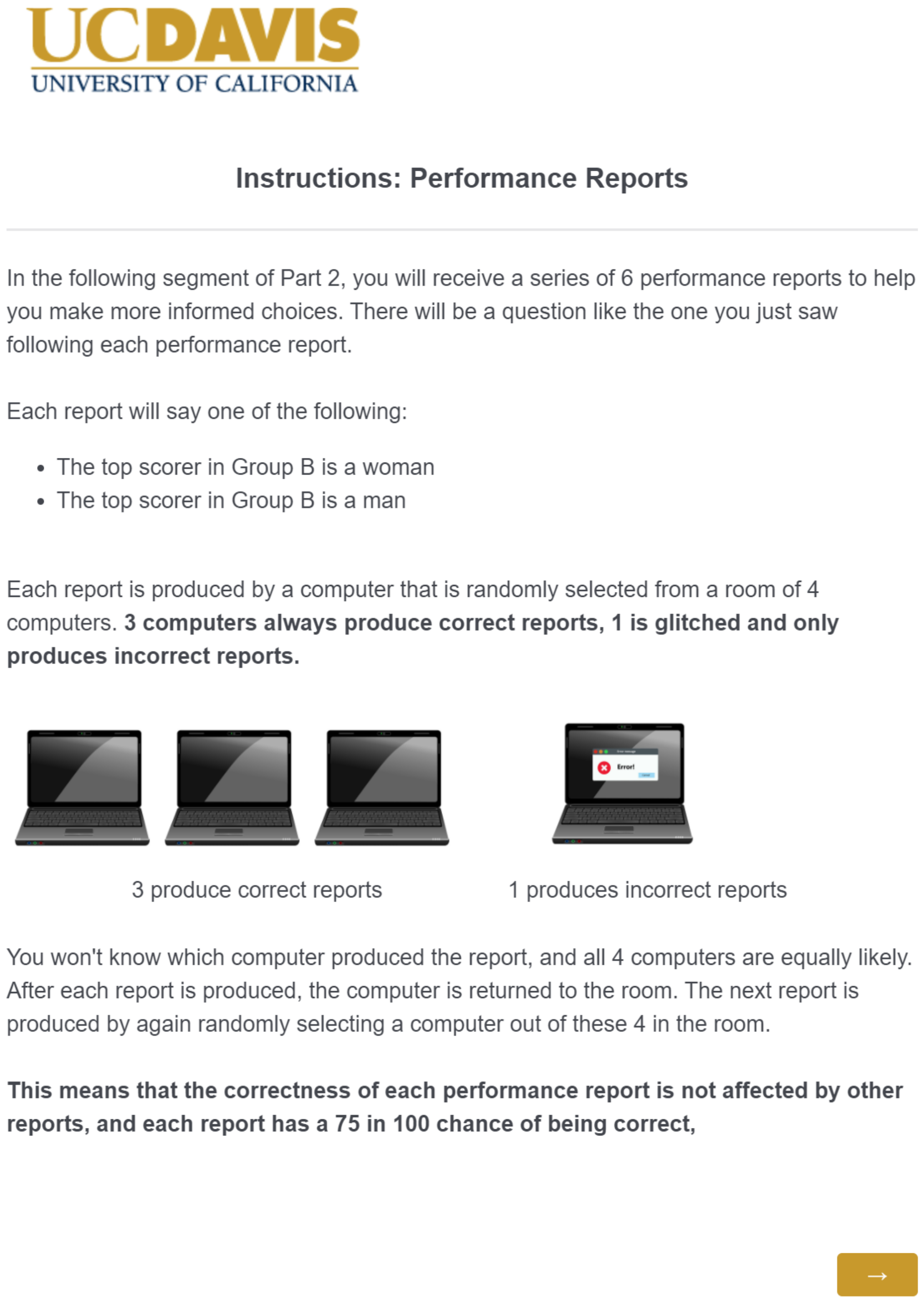
To ensure that we explain the signal structure in a way that is intuitive and easy to understand, we use the visual aid of computers, similar to the explanations in Mobius et al. (2014); Coutts (2019). We tell subjects that in the following segment in part 2, they will receives a series of 6 performance reports to help them make more informed choices. Each report is produced by a randomly selected computer from a room of 4 computers. 3 of these computers always produce correct reports, while 1 of them is glitched and always produce incorrect reports. Each computer is equally likely to be chosen, and a computer is randomly selected with replacement to produce each signal. See figure 3 for the full instructions for the signals.

We elicit the updated posterior belief of the same event after the subject observes each signal, In each posterior elicitation question, we present their reported belief from the previous elicitation question, as well as a summary of all the signals they have seen so far, in order to help facilitate accurate belief updating. Subjects also have access to information on the relative composition of their matched group in the form of a clickable button in every belief elicitation question.

4 Data

We have finished data collection for the pre-study and a pilot for the main experiment with 75%. In this section, we will provide a brief description of the pre-study sample and the main study pilot data.

In the pre-study sample, we have 136 unique subjects recruited from MTurk. After excluding subjects who failed our attention checks, we retain a sample of 104 subjects. Table 1 provides a tabulation of subjects by gender. We include the 103 subjects who identify as either men or women as our sample to generate the groups used in the main study. We will refer to these subjects as our pre-study final sample.



Instructions: Performance Reports

In the following segment of Part 2, you will receive a series of 6 performance reports to help you make more informed choices. There will be a question like the one you just saw following each performance report.

Each report will say one of the following:

- The top scorer in Group B is a woman
- The top scorer in Group B is a man

Each report is produced by a computer that is randomly selected from a room of 4 computers. **3 computers always produce correct reports, 1 is glitched and only produces incorrect reports.**

You won't know which computer produced the report, and all 4 computers are equally likely. After each report is produced, the computer is returned to the room. The next report is produced by again randomly selecting a computer out of these 4 in the room.

This means that the correctness of each performance report is not affected by other reports, and each report has a 75 in 100 chance of being correct,

Figure 3: Subjects' instructions for the signals (performance reports) in the gender treatment

Table 1: Number of Pre-Study Subjects by Gender

	Gender of subject Freq	 Pct
Other	1	0.96
Woman	38	36.54
Man	65	62.50
Total	104	100.00

We present the summary statistics of their math scores for our pre-study final sample in Table 2. The test is difficult as we see that the mean score of the entire sample is below 50%, with a large standard deviation. The mean score of women is slightly lower than men, although a two-sample t-test of mean score presented in Table 3 shows that there is no significant difference.

Table 2: Math Score Summary Statistics by Gender

	Mean	SD
Woman	9.08	4.64
Man	9.35	4.17
Total	9.25	4.33

Table 3: t-Test of Mean Score by Gender

	Difference
Math test score (out of 20)	-0.275
	(-0.31)
Observations	103

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

For the main study pilot, we recruited 63 subjects from MTurk. After excluding subjects who failed attention checks, we retain a sample of 50. In our sample, there are

26 subjects in the ID treatment and 24 in the gender treatment. Table 4 provides a tabulation of subject gender in each treatment. Overall, there are equal proportions of men and women in each treatment.

Table 4: Tabulation of Subject Gender by Treatment

	Gender	ID	Total
Woman	13	12	25
Man	11	14	25
Total	24	26	50
N	50		

One concern was whether subjects actually understood the instructions and completed the tasks as they were supposed to, given the nature of the belief updating tasks being more complex. As a sanity check, in the demographics survey at the end of the experiment, we ask subjects to provide instructions to a future participant on how best to complete the belief updating tasks. Overall, most subjects' responses reflect that they did in fact understand the instructions, the nature of the tasks, and the signal generating process. Further, their responses show that subjects do indeed update their beliefs on the particular events that are described in the experiment. In Appendix A we provided relevant excerpts from subjects' comments on this topic.

5 Preliminary Results

We calculate the correlation between actual posteriors and Bayesian predictions in each round of updating as a first step. The correlation coefficient is 0.67 in the first round, sits between 0.8 and 0.9 for rounds 2 to 5 while exhibiting a slight upward trend, and reaches 0.95 in the last round of updating. This could be a reflection of the high informativeness of the signals, where subjects see a large proportion of true signals and reach beliefs that are close to the true belief at the end of the 6 rounds of updating. The correlation coefficient is similar in magnitude to that reported in Coutts (2019) but higher than reported in Mobius et al. (2014).

Next, we graph the mean posterior beliefs of subjects in each round and compare to the mean Bayesian beliefs. Figure 4 shows the evolution of mean beliefs in each treatment. Overall the mean beliefs are quite close to the mean Bayesian beliefs. However, mean beliefs only tell a small part of the story as they do not show the actual updating pattern of subjects or heterogeneity.

Looking at mean belief revisions, we find some evidence of conservatism. Figure 5 shows that women and men have similar mean belief revisions in the ID treatment, and are conservative when compared to the mean Bayesian belief revisions. However, in the Gender treatment, women are more conservative than men, and the mean belief revisions of men are similar to mean Bayesian belief revisions.

Next, we provide some preliminary results from the main regression. We pool the data all 6 rounds of updating for the regressions due to the small sample size of the pilot. The small sample size results in standard errors that are quite large, and running the regressions separately for each round would not be very informative. Even in the pooled sample, the estimates are still not very precise, so the results shown here should be interpreted as suggestive patterns which will inform our design, rather than offering definitive conclusions. Following Coutts (2019) and Mobius et al. (2014), we also exclude wrong direction updates to avoid the possibility of results being driven by subjects who simply didn't understand instructions (there are about 15% of wrong direction updates in each round, although most subjects make at most one wrong direction update)[9]. Due to the fact that the two indicator function terms in the main regression (8) add up to 1, the regression should be run without a constant. We also cluster standard errors at the individual level.

Table 5 provides results from the main regression estimated separately for each treatment. The average updating behavior does not appear to differ between treatments. Overall, subjects appear to be more conservative than a perfect Bayesian. We next estimate the main regression by gender in each treatment and provide results in Table 6. Both women and men appear to be equally conservative in the ID treatment. However, women appear to be more conservative than men in the Gender Treatment, the stereotypical domain. There is some overlap between the 95% confidence intervals of the point estimates between men and women in the Gender Treatment, and although women are much more conservative than Bayesian, on average men appear to be similar to Bayesian. This confirms the pattern shown in Figure 5.

Next, we study our main hypothesis that people with a particular gender stereotype may be more resistant to updating when new information contradicts their stereotype

[9] 30 subjects did not make any wrong direction updates, 13 subjects made one such mistake, 6 subjects made 3 such mistakes, 2 subjects made 4 such mistakes, and only 1 subject made 5 or 6 wrong direction updates, respectively.

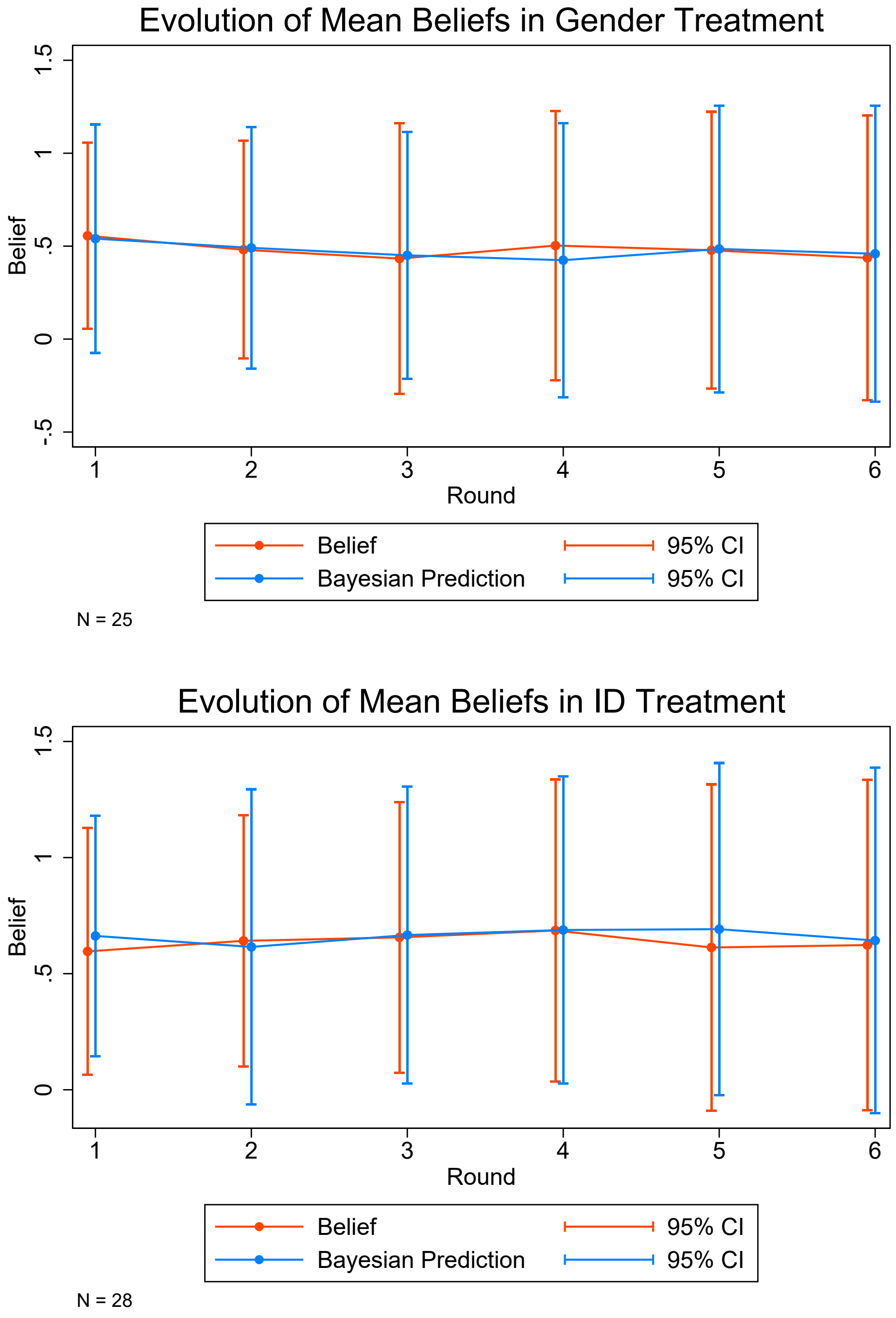


Figure 4: Comparison between mean posterior beliefs and Bayesian beliefs by treatment

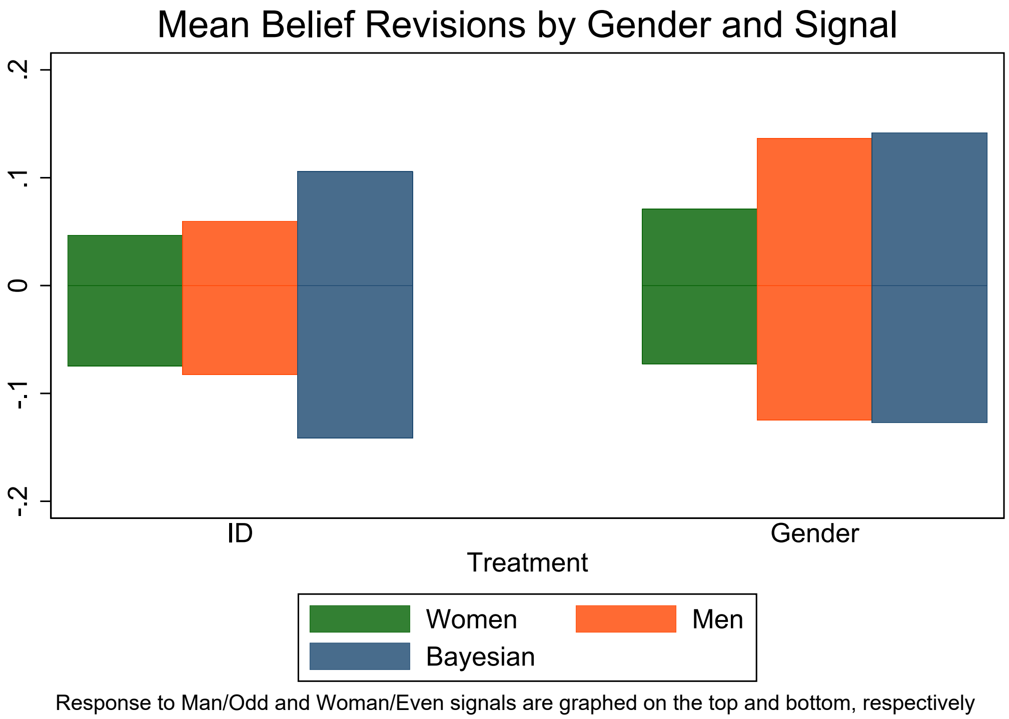


Figure 5: Mean belief revisions by treatment, subject gender, and signal type

Table 5: Pooled Main Regression by Treatment

	(1) Gender	(2) ID
δ	0.875	0.761
	[0.75,1.00]	[0.53,0.99]
β_1	0.676	0.718
	[0.43,0.92]	[0.41,1.03]
β_0	0.759	0.768
	[0.51,1.01]	[0.34,1.19]
Observations	105	121

95% confidence intervals in brackets

Table 6: Pooled Main Regression by Treatment and Gender

	Gender Treatment		ID Treatment	
	(1)	(2)	(3)	(4)
	Women	Men	Women	Men
δ	0.931	0.819	0.582	0.871
	[0.77,1.09]	[0.66,0.98]	[0.12,1.05]	[0.71,1.03]
β_1	0.400	1.054	0.876	0.638
	[0.17,0.63]	[0.67,1.44]	[0.34,1.41]	[0.25,1.03]
β_0	0.523	1.002	0.737	0.739
	[0.15,0.90]	[0.78,1.22]	[-0.05,1.53]	[0.30,1.18]
Observations	55	50	56	65

95% confidence intervals in brackets

as compared to when new information confirms their stereotype. We classify subjects in the Gender treatment as holding a strong gender stereotype about math favoring men if they report a belief that's equal to or above 60% for the event $E_{A2,gender}$ = The average score of the men in Group A is higher than the average score of the women in Group A[10]. We classify subjects as strongly favor women if their belief about this event is less than or equal to 40%. Similarly, we construct a comparison group in the ID treatment by classifying subjects as strongly favoring odd/even ID people in an analogous fashion. Table 7 provides a tabulation of subjects based on this classification in each treatment. There is a clear gender-math stereotype favoring men in the Gender Treatment, while the amount of subjects favoring odd and even ID people are the same, as we would expect in a non-stereotypical domain.

Table 7: Tabulation of Subjects Classified by Belief about Average Score in Group A

	Gender	ID	Total
Strongly Favor Women/Even	1	5	6
Moderate	14	18	32
Strongly Favor Men/Odd	9	3	12
Total	24	26	50
N	50		

In Table 8, we estimate the regression for subjects with strong stereotypes favoring men and all other subjects in the Gender Treatment, and those who strongly favor odd

[10] Group A has 10 men and 10 women in the Gender treatment, and 10 odd ID people and 10 even ID people in the ID treatment

ID and those who do not in the ID treatment. We would like to note that comparing people who strongly favor men and those who strongly favor women, and contrast the results against a similar comparison between people who favor odd ID / even ID would be a better approach. However, we are unable to take this approach since there is only 1 subject who strongly favor women, which is reflective of the small sample size of the pilot. In addition, due to the very low number of observations who hold strong priors in the ID treatment, the confidence intervals are very wide, and the corresponding results are not very informative. People who hold a strong stereotype favoring men seem to put higher weight on signals that say men are better than signals that say women are better, while there is less discrepancy in the weighting of signals for subjects in the ID treatment. However, due to the small sample size, confidence intervals are quite wide and we can only view the pattern described above as suggestive evidence.

Table 8: Pooled Main Regression by Treatment and Prior Stereotype

	Gender Treatment		ID Treatment	
	(1)	(2)	(3)	(4)
	Moderate	Favor Men	Moderate	Favor Odd
δ	0.880	0.673	0.797	0.459
	[0.72,1.04]	[0.26,1.09]	[0.53,1.07]	[0.28,0.64]
β_1	0.626	0.848	0.650	1.265
	[0.29,0.97]	[0.42,1.28]	[0.29,1.01]	[0.17,2.36]
β_0	0.815	0.494	0.739	0.581
	[0.45,1.18]	[0.15,0.84]	[0.24,1.23]	[-3.12,4.28]
Observations	71	34	108	13

95% confidence intervals in brackets

Another interesting result is when we look at heterogeneity by subject age. Table 9 shows regression results on heterogeneity by age. Subjects who are 45 or older appear to be more conservative in the Gender treatment while not in the ID treatment. Additionally, they appear to be more conservative when the signal received says women are better at math than when the signal says men are better. The 95% confidence intervals for estimates of β_0 for the two age groups are disjoint in the Gender treatment, although other confidence intervals do overlap substantially and prevent us from reaching definitive conclusions.

Table 9: Pooled Main Regression by Treatment and Age

	Gender Treatment		ID Treatment	
	(1)	(2)	(3)	(4)
	Below 45	45 or Older	Below 45	45 or Older
δ	0.836	0.806	0.747	0.833
	[0.69,0.98]	[0.42,1.20]	[0.44,1.05]	[0.64,1.03]
β_1	0.836	0.525	0.657	0.856
	[0.47,1.20]	[0.15,0.90]	[0.24,1.07]	[0.42,1.30]
β_0	0.964	0.365	0.889	0.530
	[0.68,1.25]	[0.05,0.68]	[0.24,1.54]	[0.28,0.78]
Observations	64	41	84	37

95% confidence intervals in brackets

6 Design Updates and Future Directions

6.1 Design Updates

There are several potential concerns with the current experiment we plan to address with some updates to the experimental design. First, even though we have included several belief elicitation questions about age as a distraction, it might still be relatively easy for subjects gender treatment to guess the fact that the true intent of the experiment is about gender stereotypes. This is due to the fact that there are only 3 questions about age, whereas there are 9 questions about gender. We are limited in our ability to add fillers and distraction questions because they drastically increase the length of the survey and amount of tasks subjects need to complete. This leads to the concern that since the gender stereotype about math is stigmatized, subjects who guessed the true intent of the study may self-censor their responses. To address this concern, we plan to test our hypothesis using the gender stereotype about sports knowledge, which is less stigmatized. In addition, we are moving to a design where we convey gender using names. Specifically, we ask participants in the pre-study to choose a username that is a typical name of their gender, and we tell the main study participants these usernames to convey the gender of the people they are updating on.

Another concern is that the actual updating of subjects may not be stable across time, as shown by some evidence in Coutts (2019) and Mobius et al. (2014), who use similar designs and the same structural framework. Instability of updating would mean that we cannot pool the sample across all updating rounds. We move our design to one round of updating on 6 different groups, instead of continuous updating on the same group. We vary the group composition across the 6 groups to induce some variation in the priors to capture the updating rule of subjects for a wide range of priors. This allows us to still pool the data across the 6 different groups to increase statistical power. We also plan to use smaller groups and elicit beliefs about the average score instead of the top score, as the prevailing gender-math stereotype is about the average math ability. This is also confirmed by subjects' responses to questions that ask them whether they think there exists a gender stereotype about average or top math ability, where 66% of subjects said they believe there exists a gender stereotype that the average math ability of men is higher than that of women, whereas almost 50% of subjects reported that they do not think or not sure there exists a gender stereotype that the top men are better at math than the top women.

In addition, we plan to incorporate a treatment condition in which subjects update beliefs in a gender stereotypical domain, but the stereotype in which favors women. In this domain, it is not clear how men will update due to two opposing forces: they could be motivated to update in accordance with the stereotype, but also to reach a conclusion that men are better. It is of interest to study whether belief updating in this domain have a similar pattern or the opposite asymmetry compared to the gender stereotypical domain that favors men. This provides a more robust evaluation of the effect of gender stereotypes on belief updating. Another possible update is to utilize a widely-held belief that is non-stereotypical as a control condition. The challenge lies in finding one such example that is practically implementable with quantifiable performance which allows belief updating on binary events.

6.2 Future Directions

Coutts (2019) hints at the fact that signal evaluation might depend on the sequence of prior signals received previously. This means that the signal evaluation function, even after conditioning on the prior and the new signal, can still depend on previous signals. As a future project, we plan to conduct the first systematic investigation of how the structure of signal history might affect how individuals evaluate new information. Here we describe a preliminary portable design. Subjects perform 2 rounds of belief updating on 4 similar events that are non-value relevant in order to remove the effect of motivated biases. One example task is to randomly selected an urn from 1 red urn and 3 blue urns, which contain 10 red balls and 10 blue balls each, respectively. Subjects first report their prior on the probability that the red urn is selected. Then, a ball is drawn from the selected urn and placed in a jar with 2 red balls and 2 blue balls each. In each round of updating, a ball is randomly drawn from the jar with replacement and the color of the ball is shown to the subject. Therefore, the signal is accurate with $\frac{3}{5}$ probability. Across the 4 events we could vary the proportion of the red jars to induce different priors. We can compare the updating behavior in the second round for subjects who receive different signals in the first round conditional on receiving the same signal in the second round, in order to understand the effect of prior signal on the signal evaluation function in the following period.

References

Albrecht, K., E. von Essen, J. Parys, and N. Szech (2013). Updating, self-confidence, and discrimination. *European Economic Review 60*, 144–169.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy 126*(4), 1472–1503.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964). Measuring utility by a single-response sequential method. *Behavioral Science 9*(3), 226–232.

Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *Review of Economic Studies 80*, 429–462.

Bénabou, R. (2015). The Economics of Motivated Beliefs. In *Conférence Jean-Jacques Laffont*, Volume 125, pp. 665–685.

Bénabou, R. and J. Tirole (2002). Self-Confidence and Personal Motivation. *Quarterly Journal of Economics 117*(3), 871–915,.

Bodenhausen, G. V. and R. S. Wyer (1985). Effects of Stereotypes on Decision Making and Information-Processing Strategies. *Journal of Personality and Social Psychology 48*(2), 267–282.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–773.

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.

Cacault, M. P. and M. Grieder (2019). How group identification distorts beliefs. *Journal of Economic Behavior and Organization 164*, 63–76.

Cason, T. N. and C. R. Plott (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy 122*(6), 1235–1270.

Charness, G. and C. Dave (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior 104*, 1–23.

Coffman, K., M. Collis, and L. Kulkarni (2019). Stereotypes and Belief Updating.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics 129*(4), 1625–1660.

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

Cvencek, D., A. N. Meltzoff, and A. G. Greenwald (2011). Math-Gender Stereotypes in Elementary School Children. *Child Development 82*(3), 766–779.

Danz, D., L. Vesterlund, and A. J. Wilson (2020). Belief elicitation: Limiting truth telling with information on incentives.

Dustan, A., K. Koutout, and G. Leo (2020). Beliefs about Beliefs about Gender.

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.

El-gamal, M. A. and D. M. Grether (1995). Are People Bayesian? Uncovering Behavioral Strategies. *The Quarterly Journal of Economics 90*(432), 1137–1145.

Epley, N. and T. Gilovich (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives 30*(3), 133–140.

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

Grether, D. M. (1980). Bayes Rule as a Descriptive Model: The Representativeness Heuristic. *The Quarterly Journal of Economics 95*(3), 537–557.

Grosjean, S. and B. Modell (2004). Confirmation bias. In *Cognitive Illusions: A handbook on fallacies and biases in thinking, judgement and memory*, Number January, Chapter 4, pp. 79–98. Psychology Press.

Healy, P. J. (2020). Explaining the BDM - or Any Random Binary Choice Elicitation Mechanism - to Subjects. Technical report.

Hollard, G., S. Massoni, and J.-C. Vergnaud (2010). Subjective beliefs formation and elicitation rules : experimental evidence.

Holt, C. A. and A. M. Smith (2016). Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics 8*(1), 110–139.

Hyde, J. S., S. M. Lindberg, M. C. Linn, A. B. Ellis, and C. C. Williams (2008). Gender Similarities Characterize Math Performance. *Science 321*(July), 494–495.

Hyde, J. S. and J. E. Mertz (2009). Gender, biology, and mathematics performance. *Proceedings of the National Academy of Sciences of the United States of America 106*(22), 8801–8807.

Johnston, L. (1996). Resisting change: Information-seeking and stereotype change. *European Journal of Social Psychology 26*(5), 799–825.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making 8*(4), 407–424.

Kahan, D. M., E. Peters, E. C. Dawson, and P. Slovic (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy 1*(1), 54–86.

Kahneman, D. and A. Tversky (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology 3*(3), 430–454.

Kahneman, D. and A. Tversky (1973). On the psychology of prediction. *Psychological Review 80*(4), 237–251.

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica 77*(2), 603–606.

Kovach, M. (2021). Conservative Updating.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin 480*(3), 480–498.

Kunda, Z. and L. Sinclair (1999). Motivated Reasoning With Stereotypes: Activation, Application, and Inhibition. *Psychological Inquiry 10*(1), 12–22.

Lueptow, L. B., L. Garovich, and M. B. Lueptow (1995). The persistence of gender stereotypes in the face of changing sex roles: Evidence contrary to the sociocultural model. *Ethology and Sociobiology 16*(6), 509–530.

Lueptow, L. B., L. Garovich-Szabo, and M. B. Lueptow (2001). Social change and the persistence of sex typing: 1974-1997. *Social Forces 80*(1), 1–36.

Machina, M. J. and D. Schmeidler (1995). Bayes without Bernoulli Simple Conditions for Probabilistically Sophisticated Choice. *Journal of Economic Theory 67*, 106–128.

Master, A. (2021). Gender Stereotypes Influence Children's STEM Motivation. *Child Development Perspectives 15*(3), 203–210.

Mobius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2014). Managing Self-Confidence.

Rabin, M. and J. L. Schrag (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics 114*(1), 37–82.

Taber, C. S. and M. Lodge (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science 50*(3), 755–769.

Trautmann, S. T. and G. van de Kuilen (2015). Belief Elicitation: A Horse Race among Truth Serums. *Economic Journal 125*(589), 2116–2135.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: heuristics and biases. *Science 185*(4157), 1124–1131.

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

Appendix A Subjects' Comments on How to Complete Belief Updating Tasks

"Trust the reports that come back, as they are consistent and really helped me guess my best on top scorers with even or odd."

"I think the performance reports will help you out with your percentages. At first you might be really unsure but having the performance reports will probably help your numbers."

"Just remember that only one computer is flawed, so the more reports you receive about a condition, the more likely it is true. "

"On the guessing tasks, use math to simplify things. If 3 out of 4 reports say a woman is a top scorer, based on the instructions, the chances of the reports being right are 75%. Use this knowledge to your advantage."

"The likelihood of a person being odd is the number of Odd candidates divided by the total. Then when you receive your status reports, you have to assume there is only a 25% chance it is wrong. So if the second, third or fourth agree the percent accuracy is 100 minus (0.25 x the number of reports). It is very unlikely the machine picked will be the wrong machine more than 2 times out of the 5 reports. "

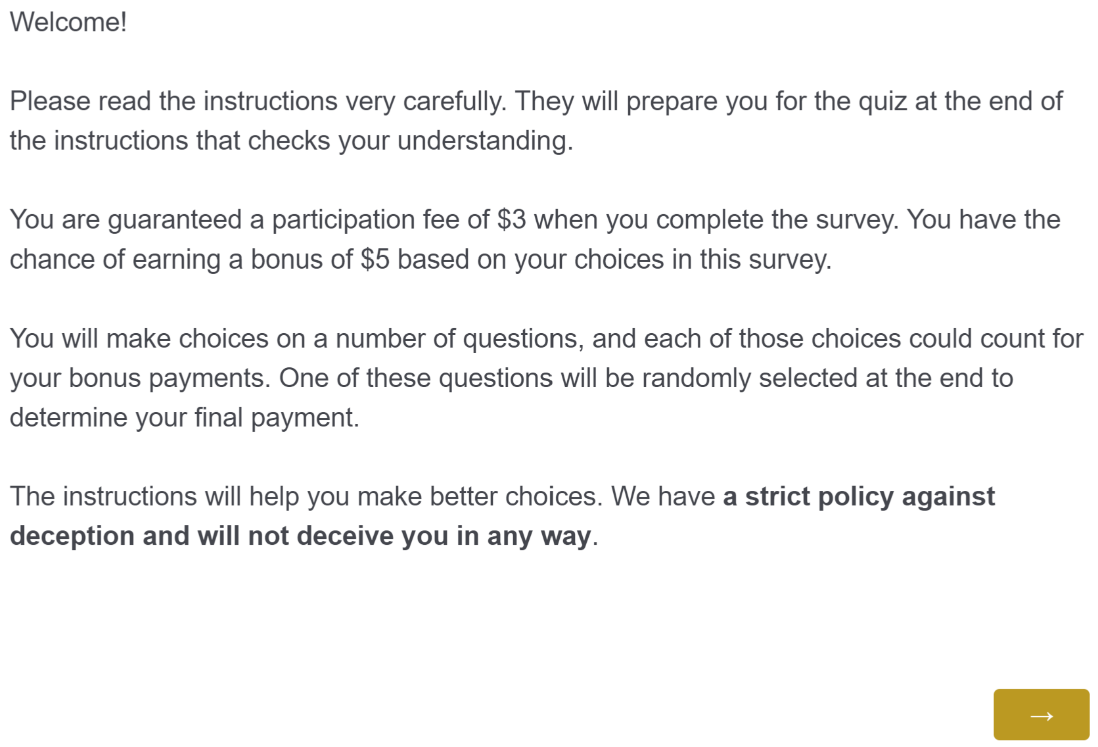
"I would advise you to use the data given to you to make the best educated guess you can. Use your understanding of statistics to help you gain a better understanding of the probabilities."

"Think about the probability based on the number of people with odd IDs and even IDs in the group to establish a baseline probability, then use the reports to adjust this probability."

"It is important to keep in mind that each performance report that the computers provide are selected independently of each other. This means that it is possible to get the same computer that submits incorrect reports multiple times in a row. Keep track of the entire tabulation and it may be worthwhile to be conservative in your initial estimations rather than swing from one side of the probability scale to the other after each result."

Appendix B Survey Instructions and User Interface

In this appendix, we provide the instructions to the subjects in the gender treatment of our main study pilot with the male version of elicitation language. We also show the graphical user interface of the elicitation questions, as well as the signals. The instructions with data from 2 randomly selected groups are presented below.



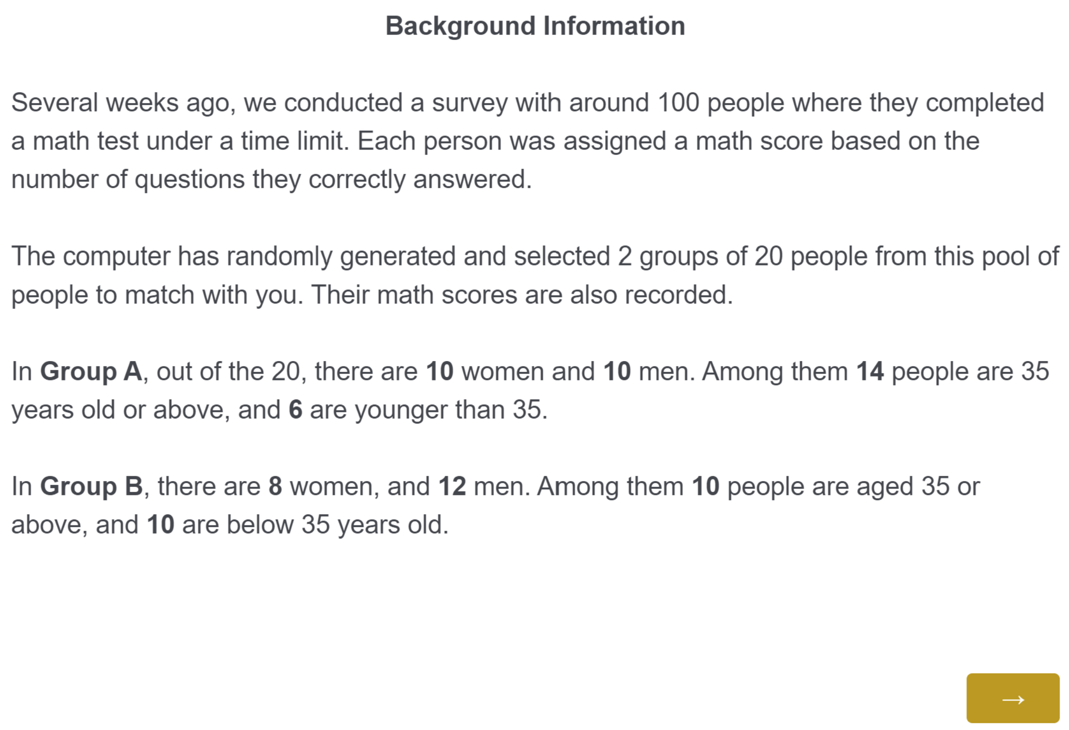
Welcome!

Please read the instructions very carefully. They will prepare you for the quiz at the end of the instructions that checks your understanding.

You are guaranteed a participation fee of $3 when you complete the survey. You have the chance of earning a bonus of $5 based on your choices in this survey.

You will make choices on a number of questions, and each of those choices could count for your bonus payments. One of these questions will be randomly selected at the end to determine your final payment.

The instructions will help you make better choices. We have **a strict policy against deception and will not deceive you in any way**.



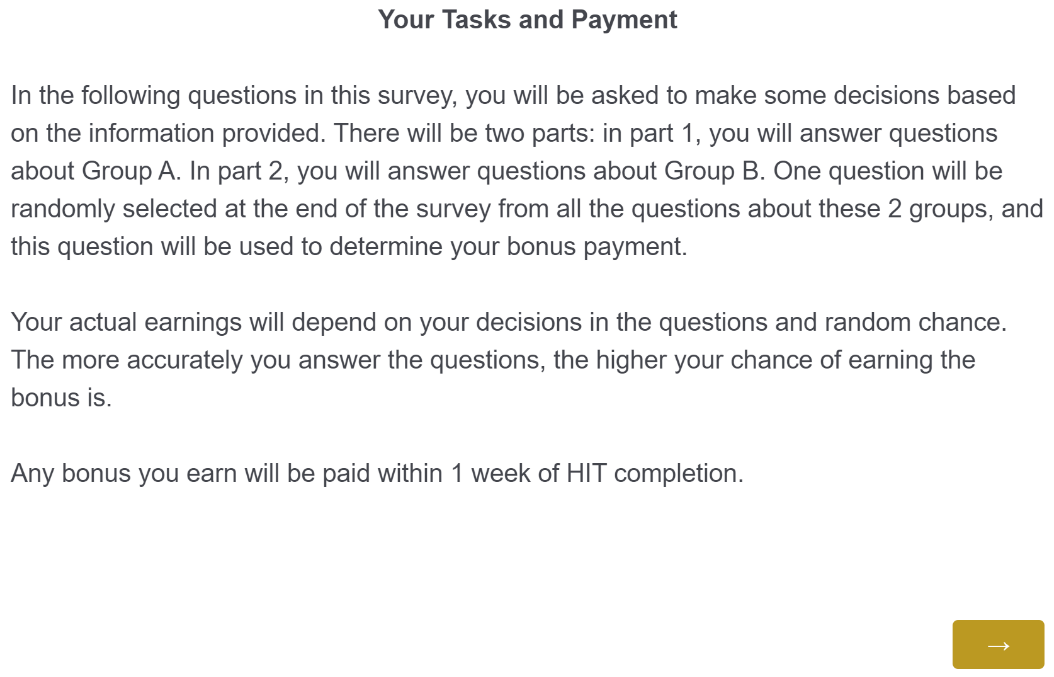
Background Information

Several weeks ago, we conducted a survey with around 100 people where they completed a math test under a time limit. Each person was assigned a math score based on the number of questions they correctly answered.

The computer has randomly generated and selected 2 groups of 20 people from this pool of people to match with you. Their math scores are also recorded.

In **Group A**, out of the 20, there are **10** women and **10** men. Among them **14** people are 35 years old or above, and **6** are younger than 35.

In **Group B**, there are **8** women, and **12** men. Among them **10** people are aged 35 or above, and **10** are below 35 years old.

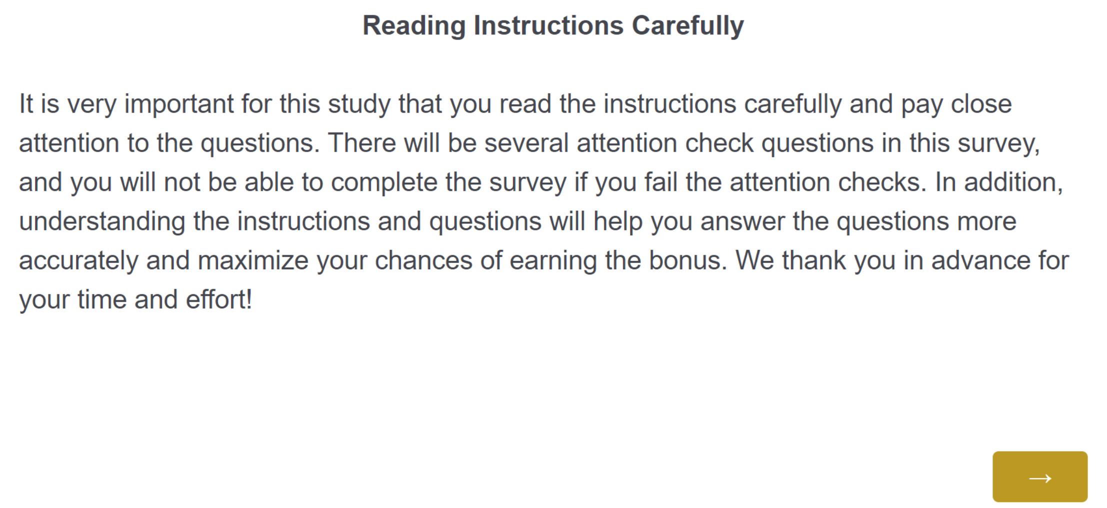


Your Tasks and Payment

In the following questions in this survey, you will be asked to make some decisions based on the information provided. There will be two parts: in part 1, you will answer questions about Group A. In part 2, you will answer questions about Group B. One question will be randomly selected at the end of the survey from all the questions about these 2 groups, and this question will be used to determine your bonus payment.

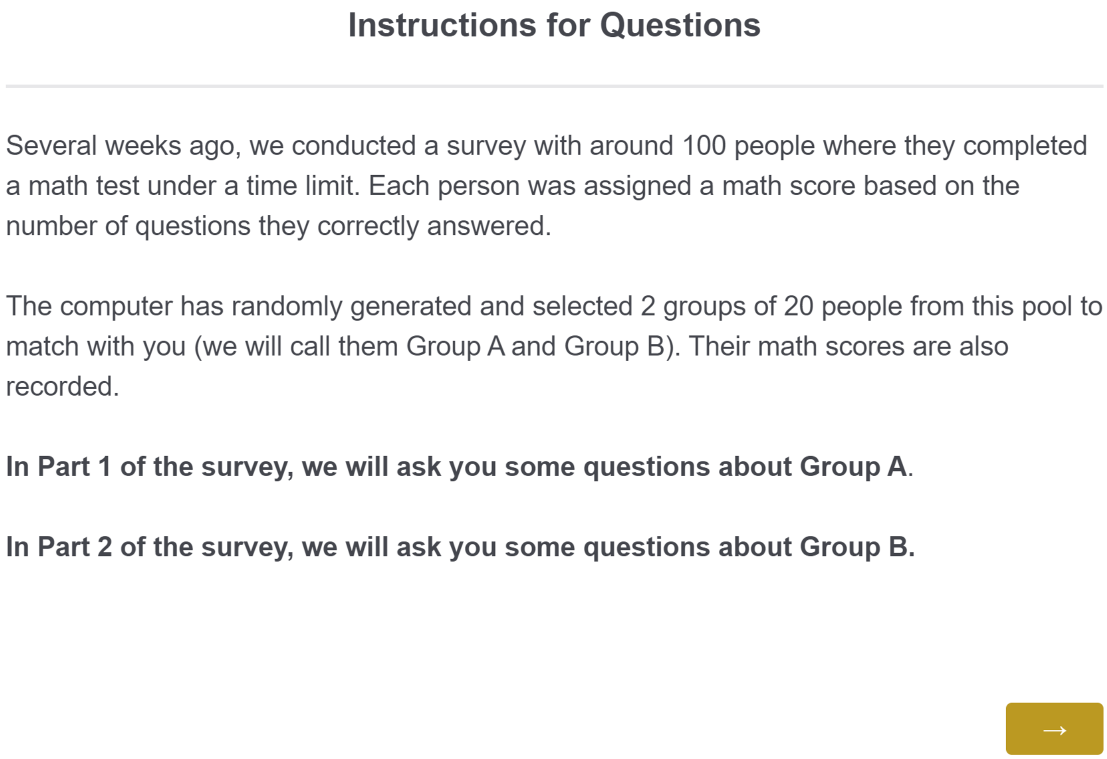
Your actual earnings will depend on your decisions in the questions and random chance. The more accurately you answer the questions, the higher your chance of earning the bonus is.

Any bonus you earn will be paid within 1 week of HIT completion.



Reading Instructions Carefully

It is very important for this study that you read the instructions carefully and pay close attention to the questions. There will be several attention check questions in this survey, and you will not be able to complete the survey if you fail the attention checks. In addition, understanding the instructions and questions will help you answer the questions more accurately and maximize your chances of earning the bonus. We thank you in advance for your time and effort!



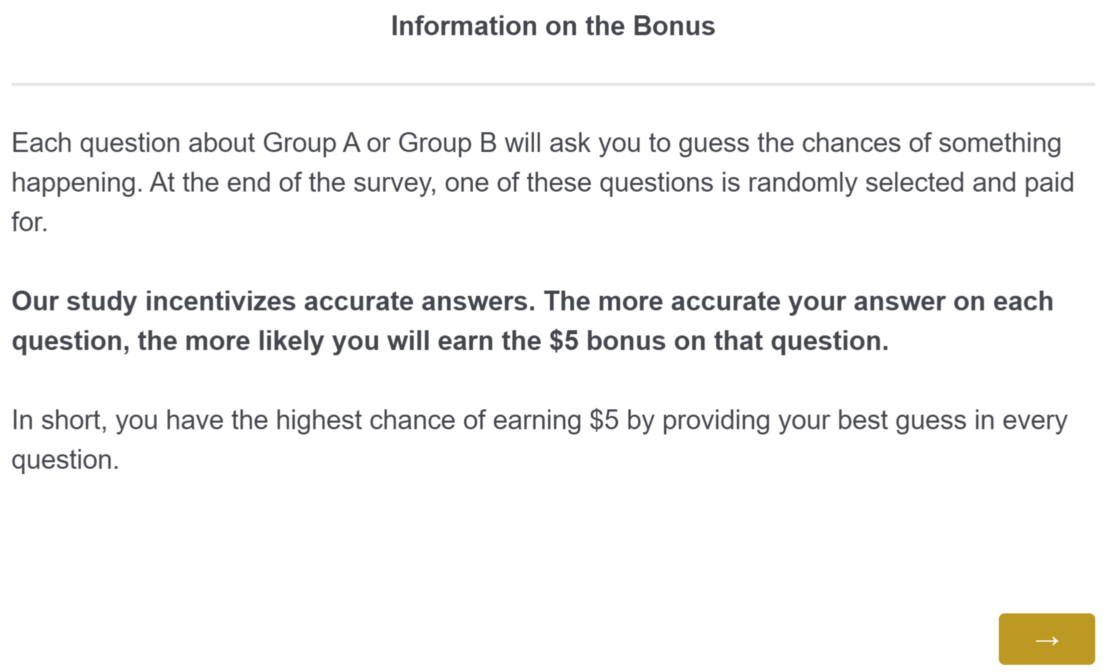
Instructions for Questions

Several weeks ago, we conducted a survey with around 100 people where they completed a math test under a time limit. Each person was assigned a math score based on the number of questions they correctly answered.

The computer has randomly generated and selected 2 groups of 20 people from this pool to match with you (we will call them Group A and Group B). Their math scores are also recorded.

In Part 1 of the survey, we will ask you some questions about Group A.

In Part 2 of the survey, we will ask you some questions about Group B.

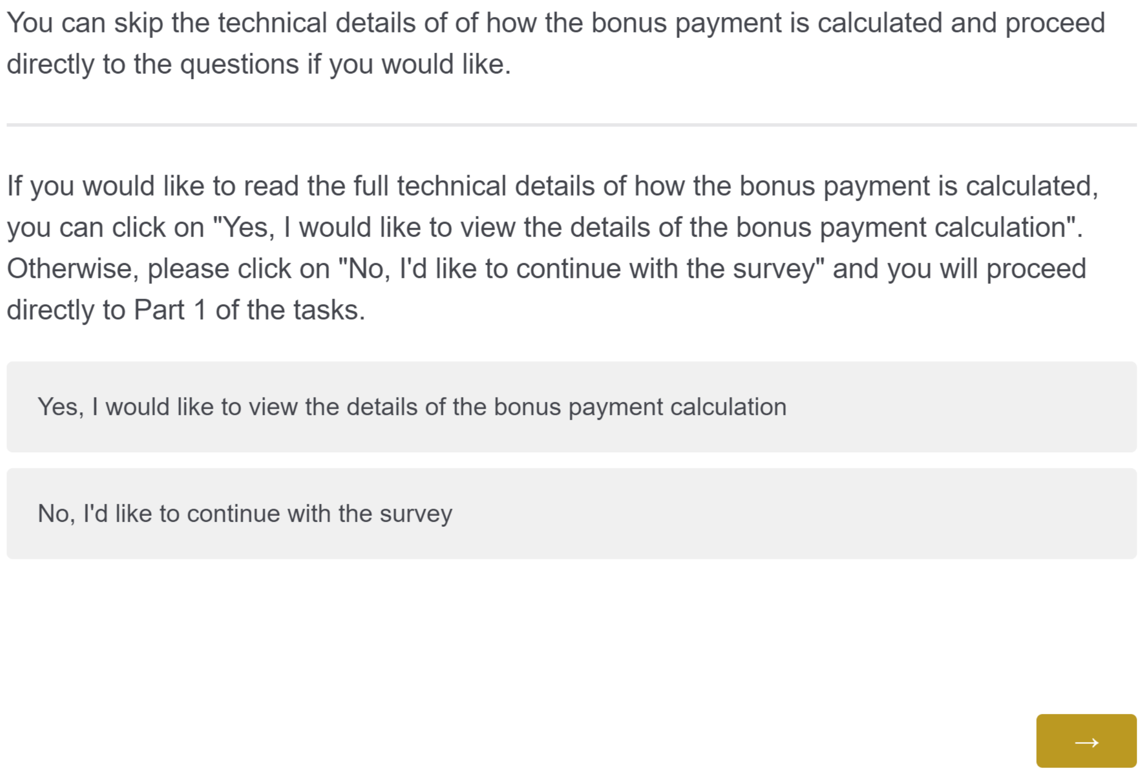


Information on the Bonus

Each question about Group A or Group B will ask you to guess the chances of something happening. At the end of the survey, one of these questions is randomly selected and paid for.

Our study incentivizes accurate answers. The more accurate your answer on each question, the more likely you will earn the $5 bonus on that question.

In short, you have the highest chance of earning $5 by providing your best guess in every question.



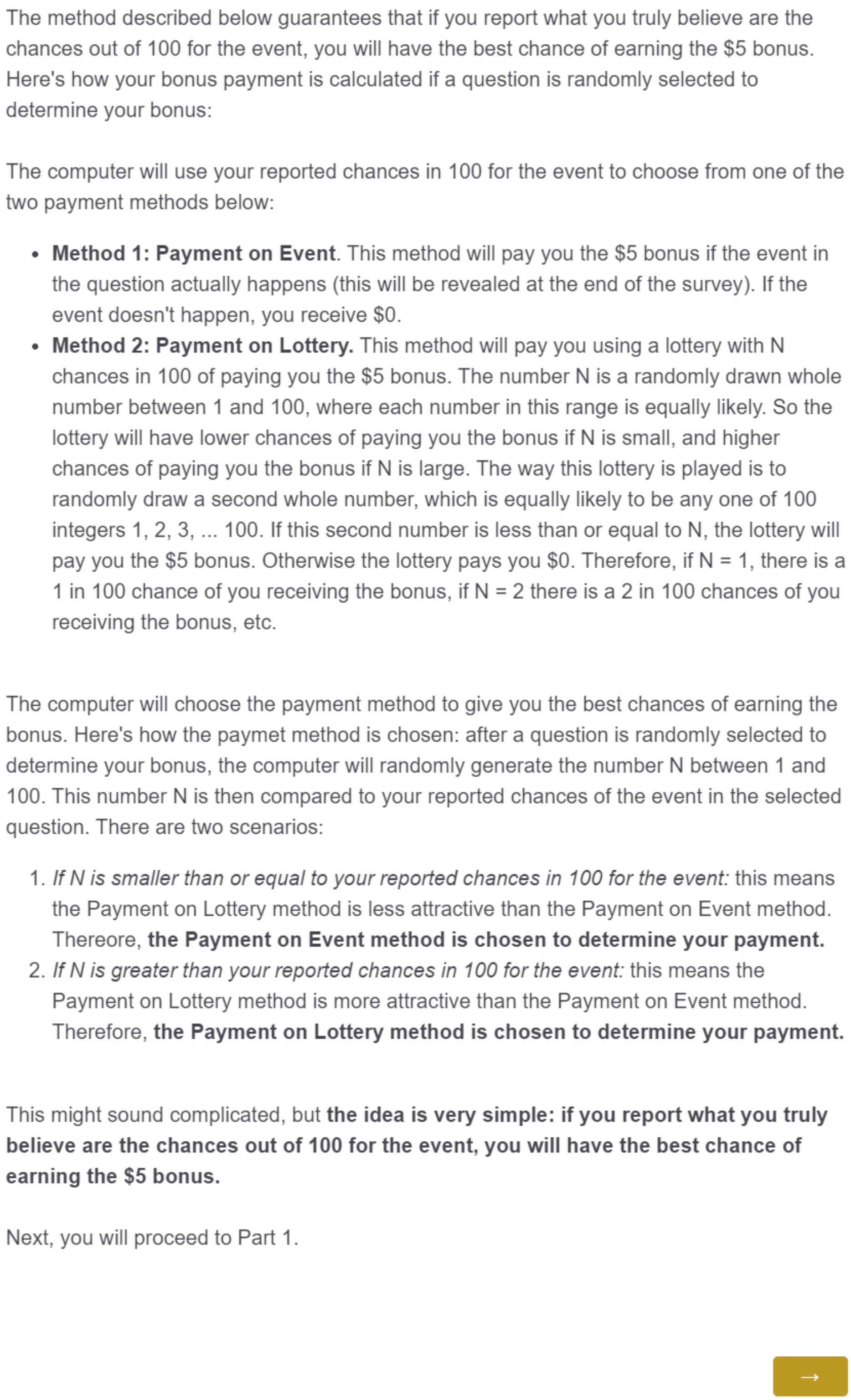
You can skip the technical details of of how the bonus payment is calculated and proceed directly to the questions if you would like.

If you would like to read the full technical details of how the bonus payment is calculated, you can click on "Yes, I would like to view the details of the bonus payment calculation". Otherwise, please click on "No, I'd like to continue with the survey" and you will proceed directly to Part 1 of the tasks.

Yes, I would like to view the details of the bonus payment calculation

No, I'd like to continue with the survey

If subjects choose to view the full incentive details, the following instructions for the incentives are shown. Otherwise they will proceed directly to Part 1.



The method described below guarantees that if you report what you truly believe are the chances out of 100 for the event, you will have the best chance of earning the $5 bonus. Here's how your bonus payment is calculated if a question is randomly selected to determine your bonus:

The computer will use your reported chances in 100 for the event to choose from one of the two payment methods below:

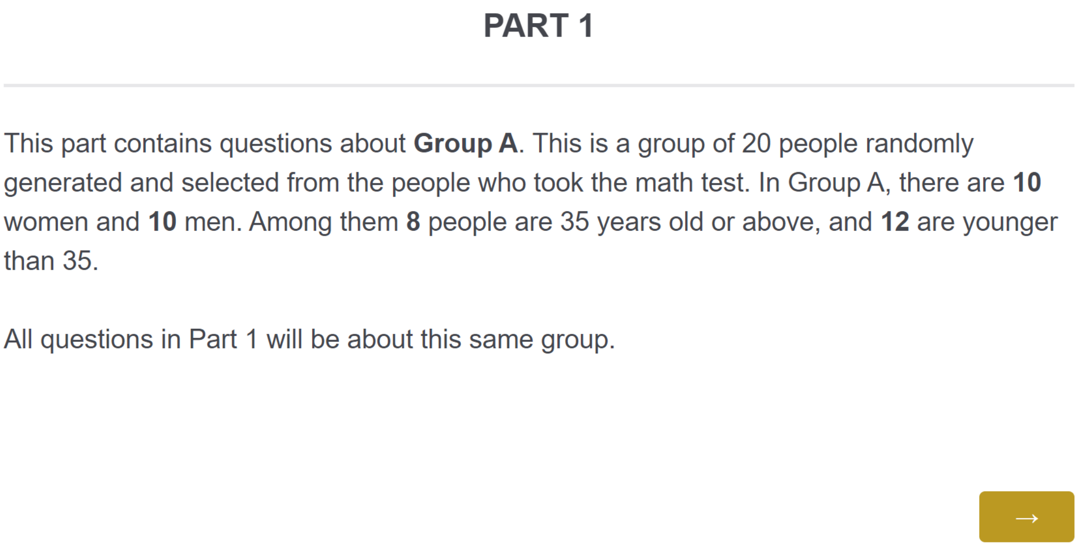
- **Method 1: Payment on Event**. This method will pay you the $5 bonus if the event in the question actually happens (this will be revealed at the end of the survey). If the event doesn't happen, you receive $0.
- **Method 2: Payment on Lottery.** This method will pay you using a lottery with N chances in 100 of paying you the $5 bonus. The number N is a randomly drawn whole number between 1 and 100, where each number in this range is equally likely. So the lottery will have lower chances of paying you the bonus if N is small, and higher chances of paying you the bonus if N is large. The way this lottery is played is to randomly draw a second whole number, which is equally likely to be any one of 100 integers 1, 2, 3, ... 100. If this second number is less than or equal to N, the lottery will pay you the $5 bonus. Otherwise the lottery pays you $0. Therefore, if N = 1, there is a 1 in 100 chance of you receiving the bonus, if N = 2 there is a 2 in 100 chances of you receiving the bonus, etc.

The computer will choose the payment method to give you the best chances of earning the bonus. Here's how the paymet method is chosen: after a question is randomly selected to determine your bonus, the computer will randomly generate the number N between 1 and 100. This number N is then compared to your reported chances of the event in the selected question. There are two scenarios:

1. *If N is smaller than or equal to your reported chances in 100 for the event:* this means the Payment on Lottery method is less attractive than the Payment on Event method. Thereore, **the Payment on Event method is chosen to determine your payment.**
2. *If N is greater than your reported chances in 100 for the event:* this means the Payment on Lottery method is more attractive than the Payment on Event method. Therefore, **the Payment on Lottery method is chosen to determine your payment.**

This might sound complicated, but **the idea is very simple: if you report what you truly believe are the chances out of 100 for the event, you will have the best chance of earning the $5 bonus.**

Next, you will proceed to Part 1.

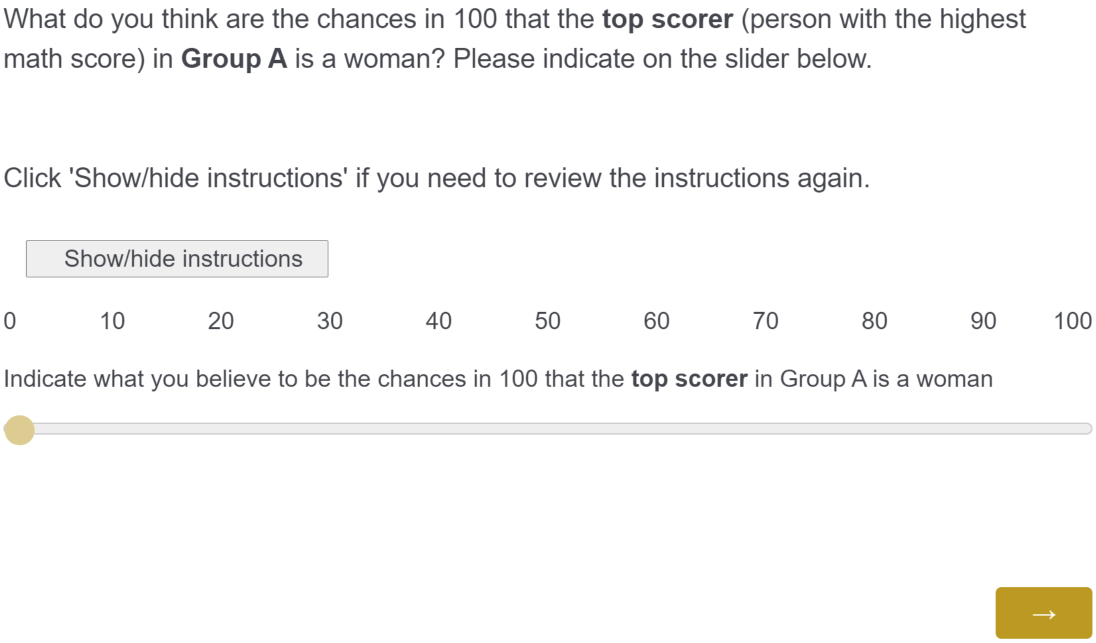


PART 1

This part contains questions about **Group A**. This is a group of 20 people randomly generated and selected from the people who took the math test. In Group A, there are **10** women and **10** men. Among them **8** people are 35 years old or above, and **12** are younger than 35.

All questions in Part 1 will be about this same group.

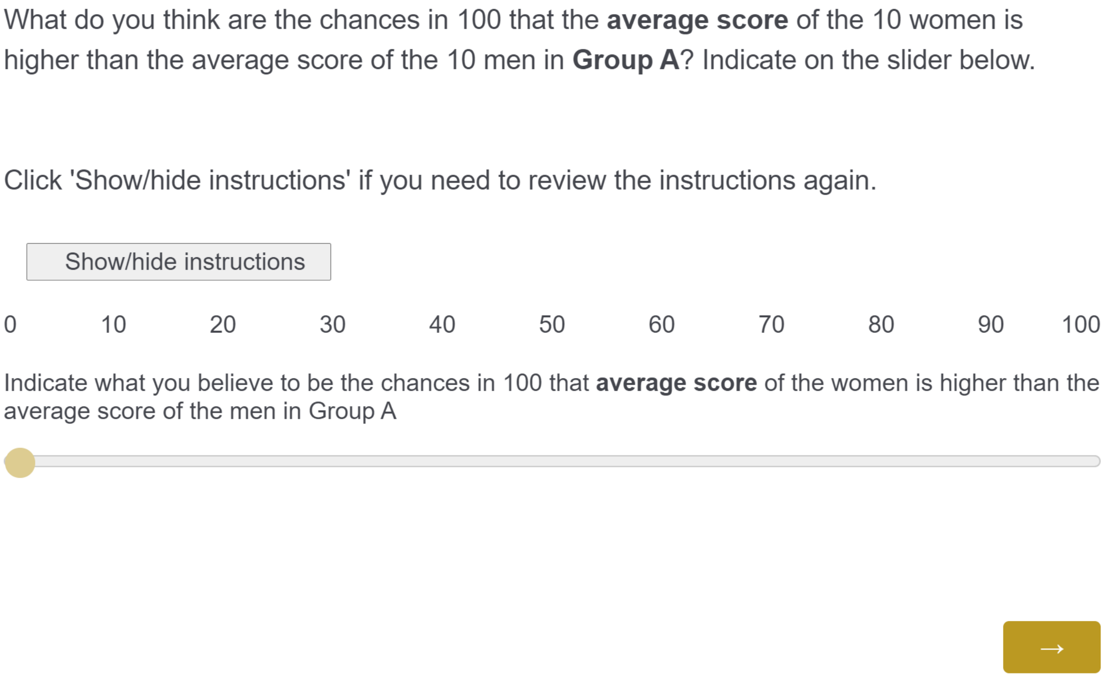
After the instructions for Part 1, subjects complete a quiz that test their knowledge of the background information about group compositions, etc. If they select a wrong answer, a warning is displayed and shows them the correct answer to reinforce understanding. After the quiz, the belief elicitation questions for Group A are shown. In each elicitation question, they can click a button to show them instructions which include the background information, group composition, and payment scheme.



What do you think are the chances in 100 that the **top scorer** (person with the highest math score) in **Group A** is a woman? Please indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

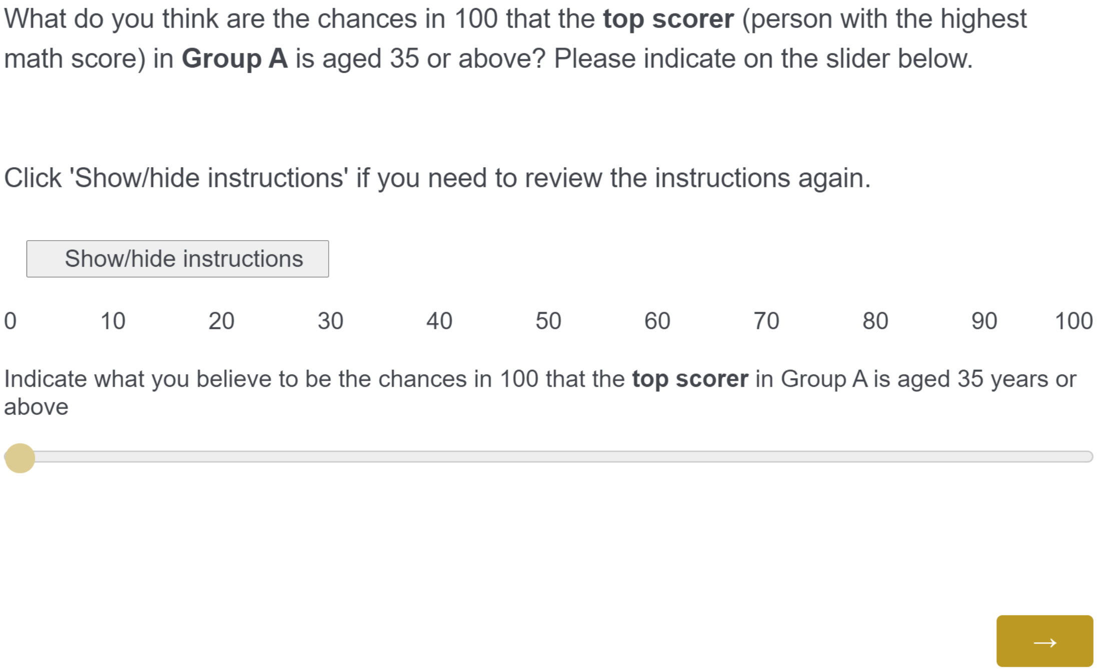
Indicate what you believe to be the chances in 100 that the **top scorer** in Group A is a woman



What do you think are the chances in 100 that the **average score** of the 10 women is higher than the average score of the 10 men in **Group A**? Indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

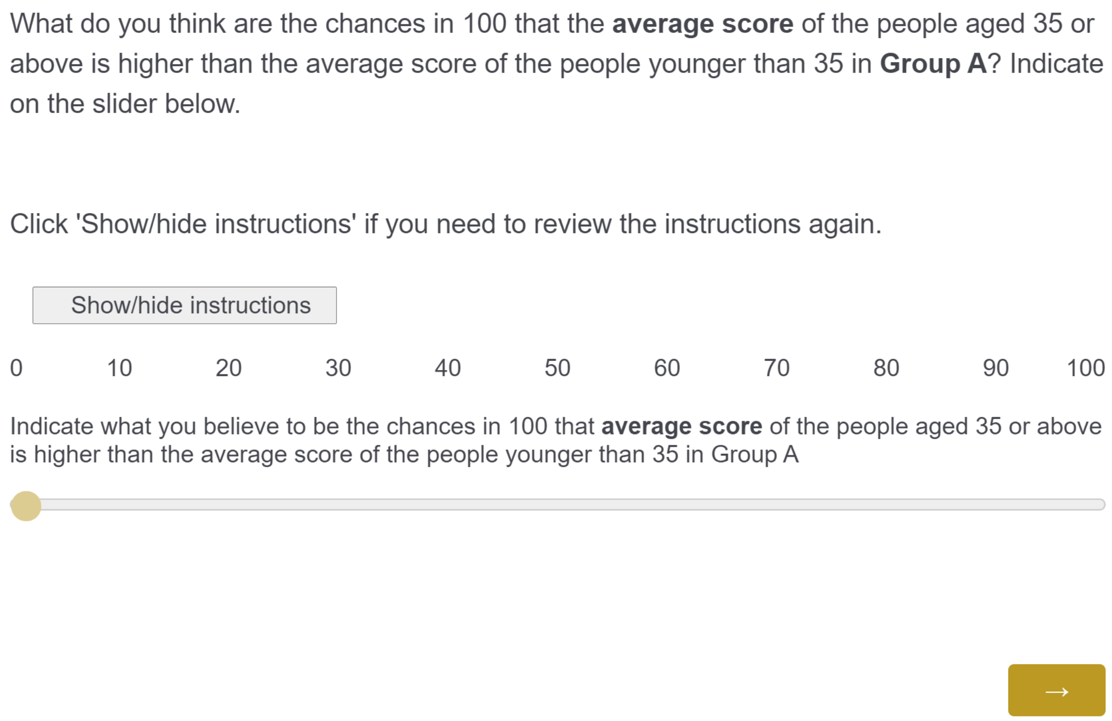
Indicate what you believe to be the chances in 100 that **average score** of the women is higher than the average score of the men in Group A



What do you think are the chances in 100 that the **top scorer** (person with the highest math score) in **Group A** is aged 35 or above? Please indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

Indicate what you believe to be the chances in 100 that the **top scorer** in Group A is aged 35 years or above

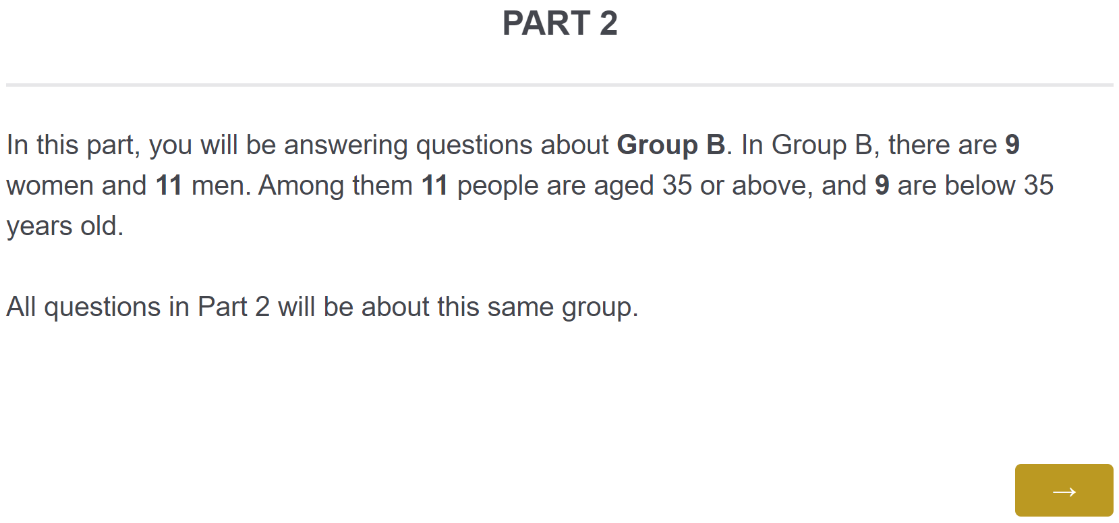


What do you think are the chances in 100 that the **average score** of the people aged 35 or above is higher than the average score of the people younger than 35 in **Group A**? Indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

Indicate what you believe to be the chances in 100 that **average score** of the people aged 35 or above is higher than the average score of the people younger than 35 in Group A

After elicitation questions about Group A, subjects receive instructions for Part 2.

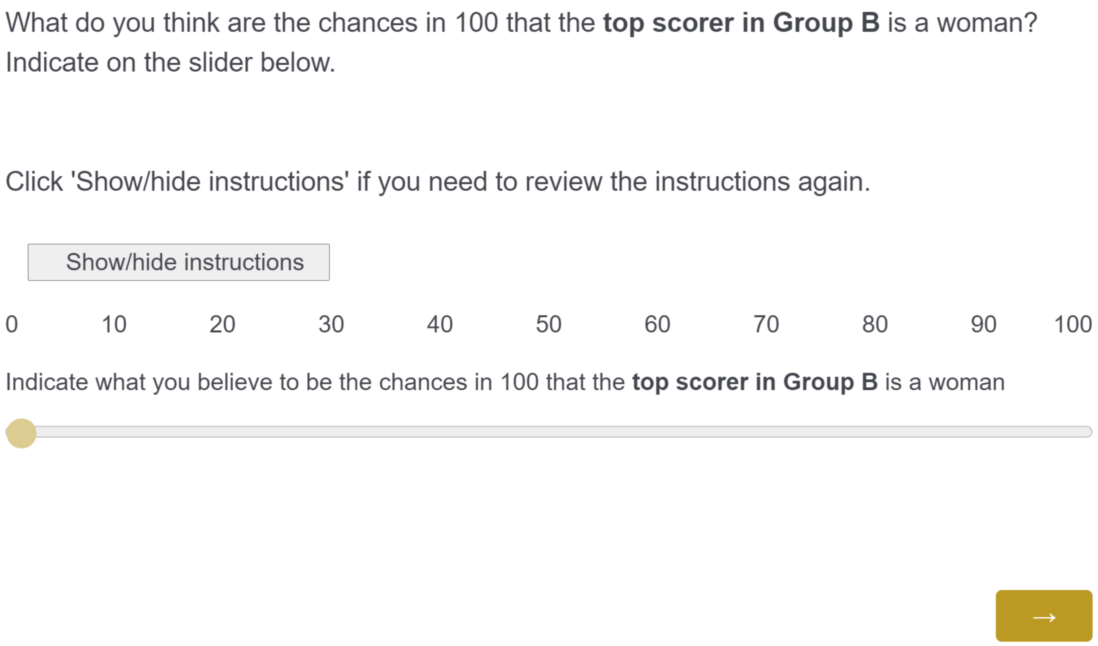


PART 2

In this part, you will be answering questions about **Group B**. In Group B, there are **9** women and **11** men. Among them **11** people are aged 35 or above, and **9** are below 35 years old.

All questions in Part 2 will be about this same group.

Subjects then complete the prior elicitation for our main updating task.

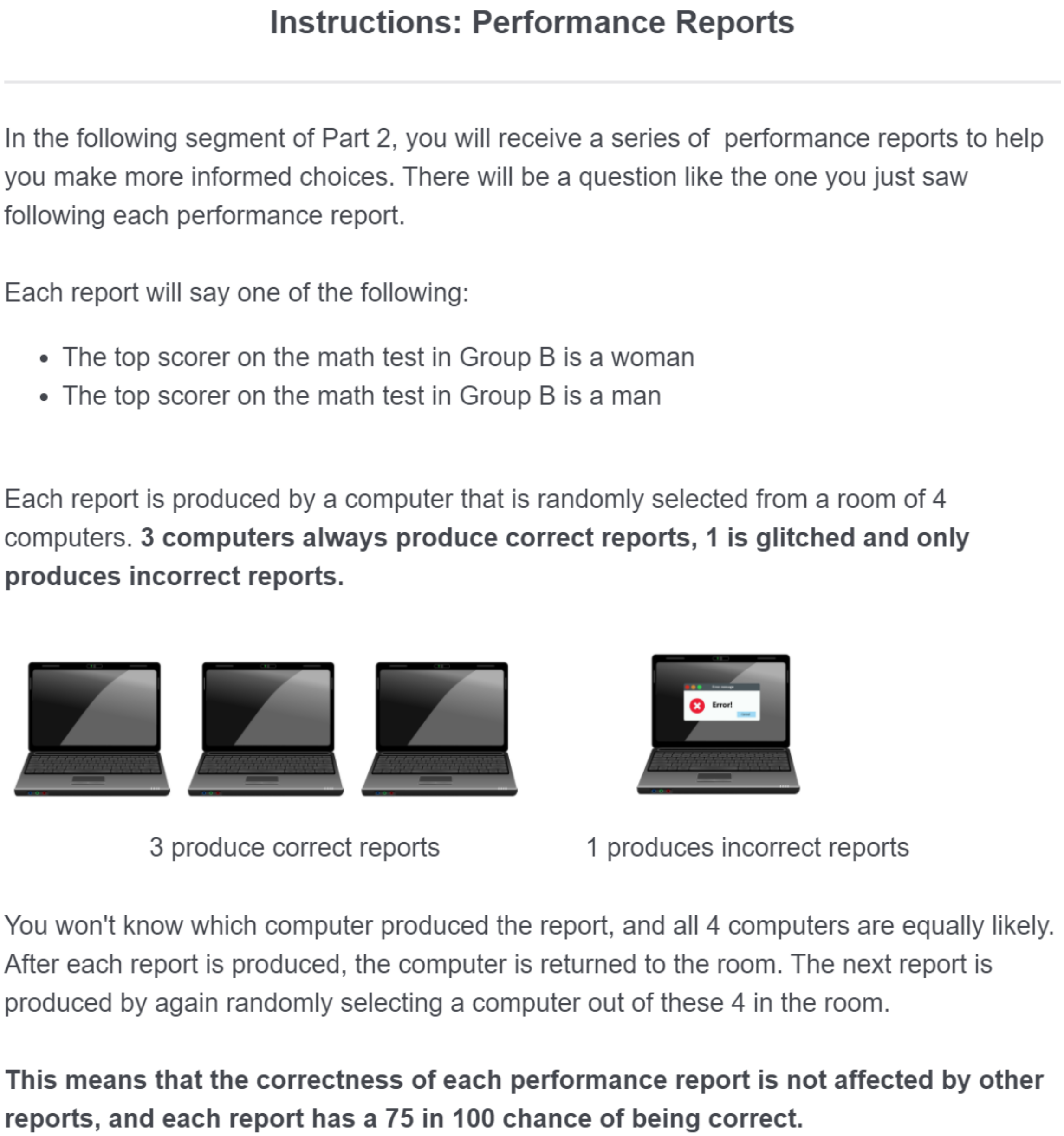


What do you think are the chances in 100 that the **top scorer in Group B** is a woman? Indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

Indicate what you believe to be the chances in 100 that the **top scorer in Group B** is a woman

After the prior elicitation, the instructions for the performance reports (signals) are shown to subjects.



Instructions: Performance Reports

In the following segment of Part 2, you will receive a series of performance reports to help you make more informed choices. There will be a question like the one you just saw following each performance report.

Each report will say one of the following:

- The top scorer on the math test in Group B is a woman
- The top scorer on the math test in Group B is a man

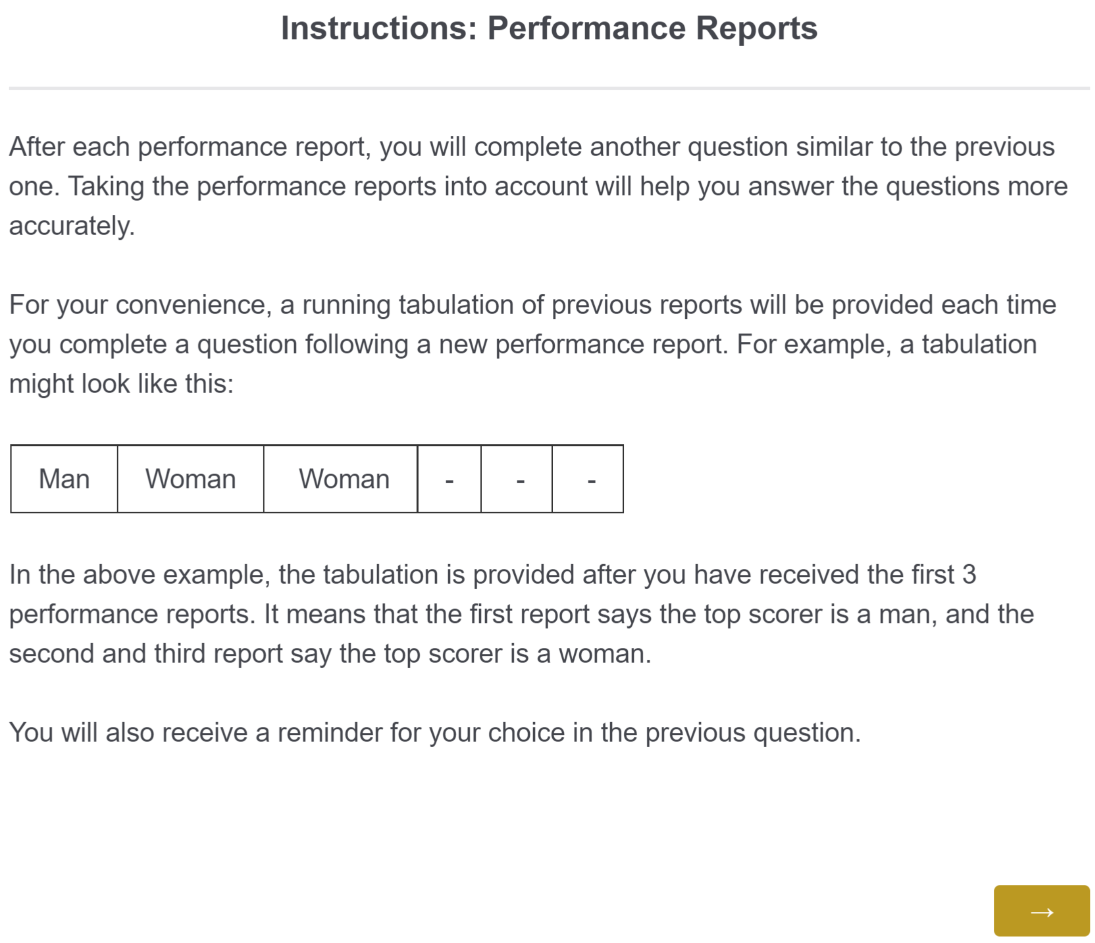
Each report is produced by a computer that is randomly selected from a room of 4 computers. **3 computers always produce correct reports, 1 is glitched and only produces incorrect reports.**

3 produce correct reports

1 produces incorrect reports

You won't know which computer produced the report, and all 4 computers are equally likely. After each report is produced, the computer is returned to the room. The next report is produced by again randomly selecting a computer out of these 4 in the room.

This means that the correctness of each performance report is not affected by other reports, and each report has a 75 in 100 chance of being correct.



Instructions: Performance Reports

After each performance report, you will complete another question similar to the previous one. Taking the performance reports into account will help you answer the questions more accurately.

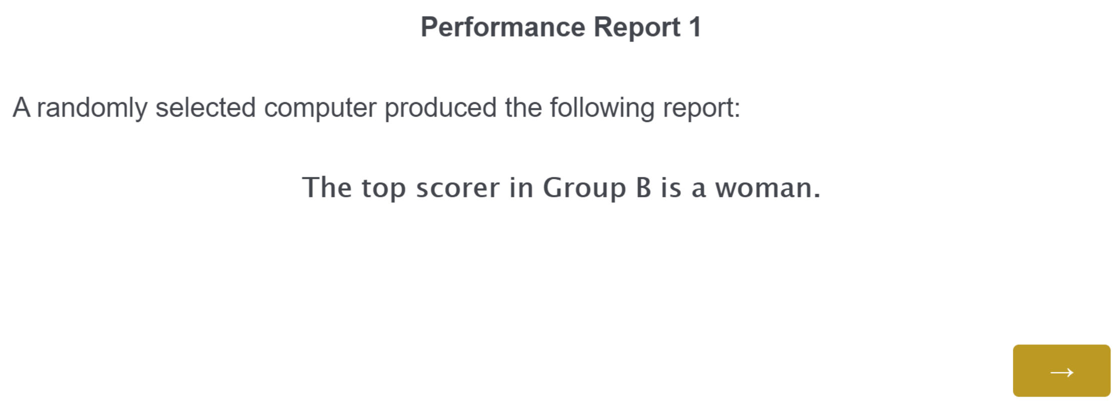
For your convenience, a running tabulation of previous reports will be provided each time you complete a question following a new performance report. For example, a tabulation might look like this:

Man	Woman	Woman	-	-	-

In the above example, the tabulation is provided after you have received the first 3 performance reports. It means that the first report says the top scorer is a man, and the second and third report say the top scorer is a woman.

You will also receive a reminder for your choice in the previous question.

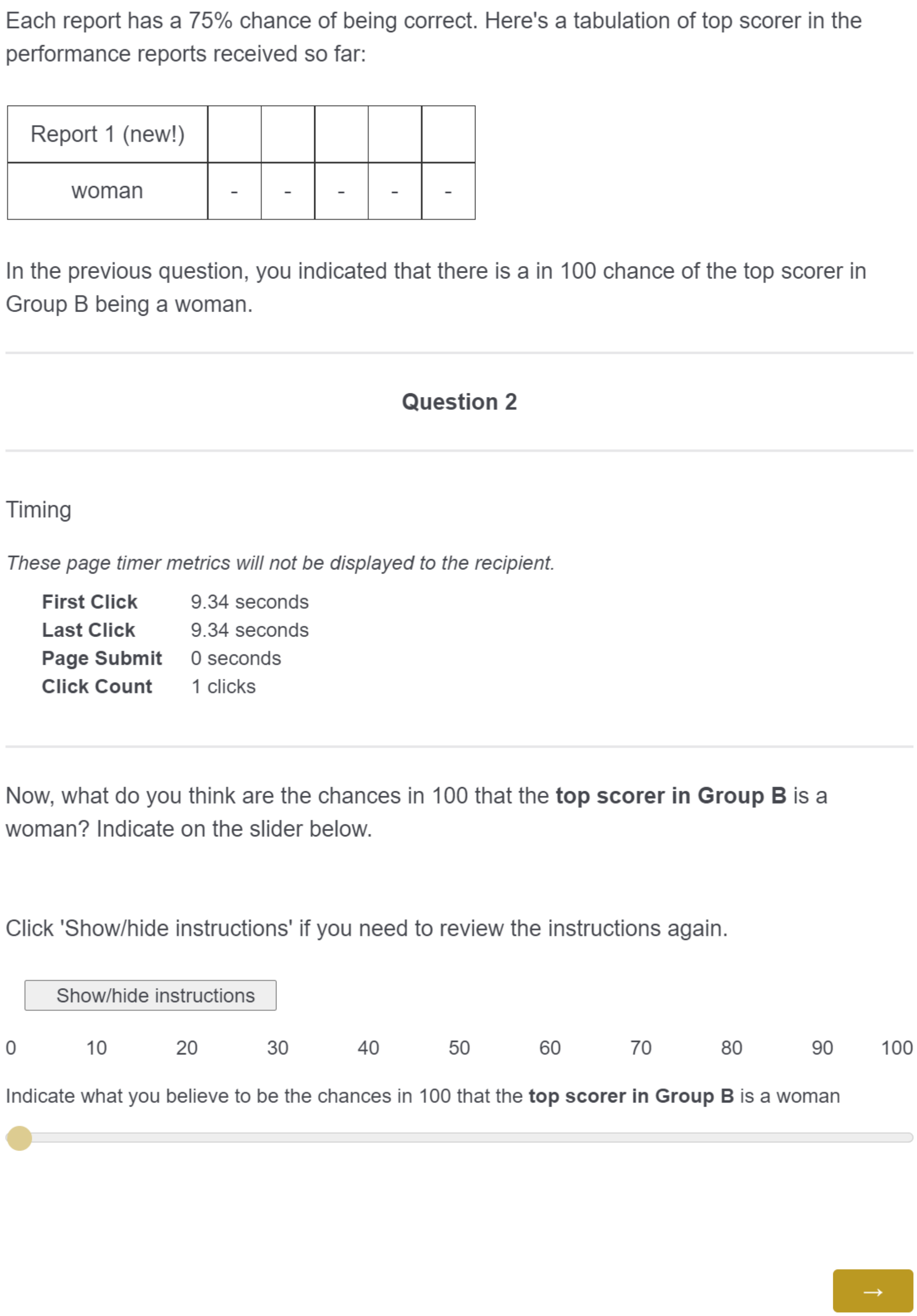
Subjects then proceed to receive the first performance report, and then complete the elicitation question for their first updated posterior. The same process is repeated 6 times, resulting in 6 rounds of belief updating. At the end of Part 2, we present a question that elicit subjects' belief about the probability of the top scorer in Group B being aged 35 or above. At the end of the survey, we present a demographics questionnaire and inform them of their actual bonus earnings. We also collect the time that subjects spend on each belief elicitation question.



Performance Report 1

A randomly selected computer produced the following report:

The top scorer in Group B is a woman.



Each report has a 75% chance of being correct. Here's a tabulation of top scorer in the performance reports received so far:

Report 1 (new!)					
woman	-	-	-	-	-

In the previous question, you indicated that there is a in 100 chance of the top scorer in Group B being a woman.

Question 2

Timing

These page timer metrics will not be displayed to the recipient.

First Click	9.34 seconds
Last Click	9.34 seconds
Page Submit	0 seconds
Click Count	1 clicks

Now, what do you think are the chances in 100 that the **top scorer in Group B** is a woman? Indicate on the slider below.

Click 'Show/hide instructions' if you need to review the instructions again.

Indicate what you believe to be the chances in 100 that the **top scorer in Group B** is a woman

BDM Incentives and Truth Telling

Che Sun

November 16, 2021

Abstract

Recent evidence has shown that incentive compatibility of belief elicitation mechanism doesn't always translate into the lab and revealing incentives might negatively impact truth telling. In this paper, I provide the first experimental evidence on the effect of incentive information on truth telling in the BDM mechanism. The results of the experiment will inform experimenters about necessary considerations in the practical implementation of the BDM mechanism and improve our understanding of its real-world performance. In the experiment, subjects are asked to report their belief about the probability of 4 events which has an objective probability. I elicit subjects' priors and have them perform two rounds of Bayesian updating tasks in each event. The experiment uses a between-subject design with 4 treatment arms: Full Information, No Information, Introspection with Payment, and Introspection without Payment. I investigate the percentage of truthful reports in each of the treatments and the deviations between actual and Bayesian posteriors to establish the effects of incentives.

1 Introduction

Eliciting beliefs of subjects is of great interest to experimental economics. Being able to observe people's beliefs is important for an experimenter to be able to make inference on their decisions and actions in games, for example, and it is central to experiments that study the formation and updating of beliefs. The current state of the art elicitation mechanisms include proper scoring rules such as the Quadratic Scoring Rule (QSR) that is based on Brier (1950) and the Binarized Scoring Rule (BSR) developed by Hossain and Okui (2013), as well as the BDM mechanism (Becker et al., 1964; Karni, 2009) which employs randomized payments.

The common property of these elicitation mechanisms is that they are incentive compatible, meaning an agent's dominant strategy is to always truthfully report their belief. In addition, BDM and BSR are incentive compatible for arbitrary risk preferences and BDM does not require subjects' preferences satisfy expected utility, which makes it particularly attractive. A number of recent experimental papers that study belief updating utilize the BDM elicitation procedure, including Mobius et al. (2014), Buser et al. (2018), Coutts (2019), and Coffman et al. (2021). There is also recent evidence that BDM outperforms QSR (Hollard et al., 2010; Trautmann and van de Kuilen, 2015; Holt and Smith, 2016), though to my knowledge no study has directly compared the performance of BSR with BDM.

The attractive theoretical properties of these state of the art belief elicitation mechanisms doesn't always fully translate to the real world, however, as emerging evidence shows. Cason and Plott (2014) document that many subjects misunderstand the incentives of BDM in a valuation elicitation task that measures the preference for an object with a known objective value, which results in systematic misreporting. Danz et al. (2020) study the effects of quantitative incentive information in the BSR mechanism, and find that the BSR incentives result in a larger proportion of misreporting than not revealing any incentive information at all. In addition, Abeler et al. (2019) conduct a meta analysis of 90 experiments in economics, psychology, and sociology and show that people's behavior are consistent with a preference for truth-telling and to be seen as truthful. Taken together, these recent evidence warrant an investigation into whether the incentive compatibility of the popular BDM elicitation method actually translates into the real world. To this end, I provide the first experimental evidence on the effect of incentive information on truth telling in the BDM mechanism. The results of the experiment will inform experimenters about necessary considerations in the practical implementation of the BDM mechanism and improve our understanding of its real world performance.

The BDM belief elicitation mechanism has several variations that are theoretically equivalent but differ in implementation. These include the direct elicitation and English Clock auction proposed by Karni (2009), and the two-stage Synchronized Lottery Choice Menu (LC) method proposed by Holt and Smith (2016). The LC method has been shown to to perform better than direct elicitation BDM, whereas the English Clock auction improves data quality by censoring naive respondents (Hao and Houser, 2012). However, they each have costs associated with their implementation. The LC method requires subjects to indicate their preferences over an event lottery and a number lottery in two consecutive choice menus for a total of 20 times, which is a dramatic increase in the cognitive and physical effort required for subjects compared to the direct elicitation method, where subjects only need to report one probability number. The English Clock auction cannot observe the true belief of subjects if the robot bidder exits the market first. Therefore, the direct elicitation method remains the most practical implementation of the BDM mechanism and this paper is primarily concerned with it. I leave the investigation of the effect of incentive information in the LC method to a future experiment.

In the following sections, I briefly review the theoretical properties of the BDM elicitation mechanism, and outline the experimental design.

2 Incentive Compatibility of BDM Mechanism

In this section, I provide a brief overview of the theoretical incentive compatibility of the BDM elicitation mechanism as outlined by (Karni, 2009). In the framework, an individual believes that the probability that an event E will occur is $\pi(E)$. A bet on an event E is a simple act that pays x if the event happens and y if the event doesn't happen, $x > y$, and denoted as $x_E y$. Denote also a lottery that pays x with probability p and y with probability $1 - p$ with $l(p, x, y)$.

Assumption 1 (Probabilistic Sophistication)**.** *A subject's preference relation $\succsim$ over the set of finite acts and lotteries D is said to exhibit probabilistic sophistication if it ranks acts or lotteries only on the basis of their probability distributions over outcomes (Machina and Schmeidler, 1995).*

This means that $\pi(E) = p \implies x_E y \sim l(p, x, y)$, and similarly $\pi(E) \geq p \implies x_E y \succsim l(p, x, y)$.

Assumption 2 (Dominance)**.** *A subject's preference relation over D exhibits dominance if $p \geq p'$ implies $l(p, x, y) \succsim l(p', x, y)$.*

In Machina and Schmeidler (1995), subjects exhibit first order stochastic dominance preference if roulette lottery R stochastically dominates roulette lottery R' implies that horse roulette lotteries $[R \text{ on } E] \succsim [R' \text{ on } E]$. Transplanted into the current simple framework, it is equivalent to the above.

Assumption 3 (No Stakes)**.** *The No Stakes assumption requires that subjects have no financial stakes in the realization of event E apart from the incentive payments on the belief elicitation.*

The experimenter is interested in the subjective probability $\pi(E)$ held by the subject. The elicitation procedure is as follows: the subject is asked to report their probabilistic belief of event E, $\mu \in [0, 1]$. A random number r is generated from the uniform distribution $[0, 1]$. If $r \leq \mu$, the mechanism awards the subject the event bet $x_E y$. If $r > \mu$, the subject is awarded the number lottery $l(r, x, y)$.

The above three assumptions ensure that truthfully reporting $\mu = \pi(E)$ is the subject's dominant strategy. Karni (2009)'s argument shows that truthful reporting guarantees that the subject always gets either the event bet or the number lottery with the highest probability of winning x. Suppose the subject reports a number $\mu > \pi(E)$. The chance of winning x is the same as reporting truthfully when $r \leq \pi(E)$ or $r \geq \mu$. However, if $\pi(E) < r < \mu$, the subject is forgoing the lottery $l(r, x, y)$ that gives a higher chance of winning x. The same argument applies when the subject reports $\mu < \pi(E)$. The above assumptions for incentive compatibility of BDM does not require that preferences satisfy expected utility. Additionally the incentive compatibility holds for arbitrary risk preferences.

3 Experimental Design

In the experiment, I utilize a between subject design with 4 treatments: Full Information, No Information, Introspection with Payment, and Introspection without Payment. The treatments vary the type and amount of incentive information that is provided to subjects, and the belief elicitation tasks and the events are held constant across treatments. In each of the treatments, there are 4 periods, and I elicit subjects' beliefs about 4 different events which have an objective prior in each of the periods. For each event, subjects are provided with a series of 2 informative signals, and their updated posteriors are elicit after each provision of signals. Therefore, there are 3 belief elicitations per period/event, and 12 belief elicitations in total per treatment. Subjects are paid for one randomly chosen event and one of the 3 beliefs in that chosen event in the Full Information and No Information treatments.

There are several theoretically equivalent versions of BDM belief elicitation, the most popular of which include one where subjects directly report their subjective probability of the event (Karni, 2009), and another termed the two-stage Synchronized Lottery Choice Menu (LC), where subjects indicate their preference between an event lottery and a number lottery (Holt and Smith, 2016). While the LC method has been shown to perform better than the direct elicitation BDM (Holt and Smith, 2016), it comes with additional cost to subject's time and cognitive effort (a 20-fold increase in the amount of choices subjects need to make). Therefore, in experiments where beliefs need to be elicited multiple times, the direct BDM elicitation is more practical and remains a popular choice of experimenters. In this experiment, I am primarily concerned with testing the direct elicitation BDM method where subjects report their probabilistic belief directly. I leave the investigation of the effects of incentive information in the LC method of BDM elicitation to additional treatments to be implemented in the future.

3.1 Events

In all of the events, subjects are presented with 5 urns that are red and blue, and asked to report the probability that a randomly selected urn is red. In the first event, there are one red and four blue urns, the second event two red and three blue urns, the third event three red and two blue urns, and the fourth event four red and one blue urn. For each event, subjects are informed that every single urn has equal chances of being selected. Therefore, I provide sufficient information to induce their subjective probabilities to be equal to the objective probabilities of selecting a red urn, i.e. 20% , 40%, 60%, and 80% in these four events, respectively. The only assumption needed is that subjects have a basic understanding of probability.

In every urn, there are 10 balls colored red and blue, with red urns containing more red balls and blue urns containing more blue balls. Subjects are informed of the composition of the red and blue urns, where each red urn contains 6 red balls and 4 blue balls, and each blue urn contains 4 red balls and 6 blue balls. In each event, after subjects report their priors, they are given two signals. Each signal consists of a draw (with replacement) from the selected urn, and subjects observe the color of the ball draw. Therefore, conditional on the true color of the selected urn, each signal is 60% accurate. After subjects observe each signal, their updated posterior belief is elicited. This amounts to two rounds of signal provision and belief updating in each period. At the end of each period, subjects are shown the implementation of the BDM procedure if each of their 3 belief reports were selected and their corresponding payment.

As Danz et al. (2020) and Schlag et al. (2015) pointed out, quantitative incentive information lead subjects to comprehend more complex tasks better, or they might lead them to exert more cognitive effort in the case of more effort-intensive tasks such as the aforementioned updating task. I compare the subjects' reported posterior beliefs against Bayesian posteriors, and contrast the deviation from Bayesian posteriors across treatments. This provides more insight into the effects of each treatment on subject performance in tasks that require more probabilistic sophistication and cognitive effort.

3.2 Treatments

In the Full Information treatment, subjects are provided with the full quantitative incentive information of the BDM mechanism. The incentives are framed as a choice between an event lottery and a number lottery. Subjects are informed that a numbered ball is randomly drawn from a box that contains 100 balls numbered 1 to 100. After one of their decisions is selected for payment, the number on the drawn ball is compared to their reported belief. If the number on the ball is smaller than or equal to their reported belief, they are paid if the selected urn is red. If the number on the ball is greater than their reported belief, they are entered into a lottery with chances (out of 100) of winning equal to the number on the ball. This number lottery is implemented by drawing another ball with replacement from the box and comparing its number to that on the initial ball. In addition, they are informed that the purpose of this procedure is to incentivize truth telling, and a mathematical proof is available upon request.

In the No Information treatment, subjects are not provided with any incentive information. They are simply informed that they will be paid on one randomly selected belief elicitation task, and that the payment algorithm is designed so that the more accurately and truthfully they report their beliefs, the higher chances of earning the payment they will have. In the Introspection with Payment treatment, subjects are simply asked to report their beliefs and encouraged to report truthfully and accurately. They are paid a flat fee for the belief elicitation tasks. The Introspection without Payment treatment is exactly equivalent to the Introspection with Payment treatment except for the fact that there are no monetary incentives for reporting beliefs, which is in line with the most common practices of belief elicitation in psychology experiments.

3.3 Sample

I plan to recruit subjects from undergraduate students at University of California, Davis, as well as workers on the online crowd-sourcing platform Prolific. The experiment will be conducted in person with undergraduates at UC Davis. The advantage of conducting the experiment in person is that actual physical urns and balls can be used, and they can be examined by subjects prior to the experiment, which can help earning subjects' trust in the experimenter as well as the algorithm. Additionally, the visual and physical aids will make it easier to explain the procedures to subjects.

Even though in-person experiments have attractive properties, it is not always practical and entails more costs than conducting experiments online. Therefore, I also conduct the experiment in a survey form using visual aids with Prolific workers. Gupta et al. (2021) show that Prolific workers produce less noisy data than the popular platform MTurk. Comparing results from the in-person and online experiments will provide insights on the effects of incentives and the resulting performance of the BDM mechanism in the most common real-world scenarios and subject populations of economics experiments.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for Truth-Telling. *Econometrica 87*(4), 1115–1153.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964). Measuring utility by a single-response sequential method. *Behavioral Science 9*(3), 226–232.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review 78*(1), 1–8.

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.

Cason, T. N. and C. R. Plott (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy 122*(6), 1235–1270.

Coffman, K. B., P. U. Araya, and B. Zafar (2021). A (Dynamic) Investigation of Stereotypes, Belief-Updating, and Behavior.

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

Danz, D., L. Vesterlund, and A. J. Wilson (2020). Belief elicitation: Limiting truth telling with information on incentives.

Gupta, N., L. Rigotti, and A. Wilson (2021). The Experimenters' Dilemma: Inferential Preferences over Populations.

Hao, L. and D. Houser (2012). Belief elicitation in the presence of naïve respondents: An experimental study. *Journal of Risk and Uncertainty 44*(2), 161–180.

Hollard, G., S. Massoni, and J.-C. Vergnaud (2010). Subjective beliefs formation and elicitation rules : experimental evidence.

Holt, C. A. and A. M. Smith (2016). Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics 8*(1), 110–139.

Hossain, T. and R. Okui (2013). The binarized scoring rule. *Review of Economic Studies 80*(3), 984–1001.

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica 77*(2), 603–606.

Machina, M. J. and D. Schmeidler (1995). Bayes without Bernoulli: Simple Conditions for Probabilistically Sophisticated Choice. *Journal of Economic Theory 67*, 106–128.

Mobius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2014). Managing Self-Confidence.

Schlag, K. H., J. Tremewan, and J. J. van der Weele (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics 18*(3), 457–490.

Trautmann, S. T. and G. van de Kuilen (2015). Belief Elicitation: A Horse Race among Truth Serums. *Economic Journal 125*(589), 2116–2135.