Second-Order Beliefs and Gender

Andrew Dustan[1], Kristine Koutout[2] and Greg Leo[1]

[1]Department of Economics, Vanderbilt University
[2]Graduate School of Business, Stanford University

May 23, 2022

Abstract

Beliefs about beliefs—second-order beliefs—about the differences between populations are important to understanding differences in outcomes between those populations. To study their potential impact, we develop an incentive-compatible experimental framework for eliciting beliefs (first-order) and beliefs about beliefs (second-order) about the differences in any measurable characteristic between any two populations. We implement the procedure to study beliefs about men's and women's performance on a math task and their choices in the ultimatum game. In the math task, 71% of participants believe that most *men* believe men outscore women. In contrast, 34% believe that most *women* believe men outscore women. Despite these differences in second-order beliefs with respect to math ability, we observe no such difference in first-order beliefs. On the other hand, we find no difference in participants' beliefs about men's and women's beliefs with respect to choices in the ultimatum game, which is consistent with our observation of no difference in first-order beliefs. These results have important labor market implications for the persistence of gender gaps.

Keywords: Higher-Order Beliefs, Gender, Experimental Methods

> "Leaders in the field— men and sometimes women— simply don't believe that women are as good at doing science."
>
> Alison Coil From a 2017 article in Wired Magazine

[1] Contact information: Leo [Corresponding], g.leo@vanderbilt.edu, Vanderbilt University Department of Economics 5201 West End Avenue Nashville, TN 37235. Dustan, andrew.dustan@vanderbilt.edu. Koutout, kkoutout@stanford.edu. We thank Andrea Moro, Federico Gutierrez, Ernesto Reuben, Nikos Nikoforakis, Enrique Fatas, Lise Vesterlund, and seminar attendees at BEEMA V, SEA 2019, ESA-North America 2019, The University of Toledo, and NYU Abu Dhabi for helpful feedback. Matthew Chambers and Maksim provided outstanding programming and technical support. This research was approved by the Institutional Review Board at Vanderbilt University on November 6, 2017, with approval number 171857 and on January 20, 2022, with approval number 220082.

1 Introduction

Do women *believe* that leaders in science, technology, engineering, and math (STEM) fields believe that women are bad at doing science? Such beliefs about beliefs—*second-order beliefs*—could drive women to sort out of STEM fields, leading to the observed gender gap in employment (Beede et al., 2011). Importantly, this belief-driven sorting could occur regardless of leaders' true beliefs about women's scientific abilities. When historically persistent beliefs about the differences between men and women—*first-order beliefs*—cause disparities, they may generate second-order beliefs that perpetuate those disparities even once first-order beliefs change.

In addition to the STEM/non-STEM employment gender gap, second-order beliefs could contribute to empirically documented differences in men's and women's outcomes in education (Lundberg, 2017) and wages (Blau and Kahn, 2017), among other outcomes of interest. Moreover, second-order beliefs about the differences between populations characterized by other dimensions, such as race/ethnicity, religious affiliation, or sexual orientation, may be important to understanding differences in outcomes between these groups. Despite their potential importance, second-order beliefs about population differences have rarely been studied and never, to our knowledge, directly measured.

In this paper, we introduce second-order beliefs about the differences between two populations as an important factor to consider in the study of unequal outcomes and provide researchers with an incentive-compatible experimental framework to measure them. We implement this procedure in a lab experiment to elicit beliefs about the relative ability and choices of women and men using two tasks that are commonly employed in the experimental literature studying gender differences. Specifically, we elicit beliefs about men's and women's performance on a timed math task (Niederle and Vesterlund, 2007; Reuben et al., 2014) and choices in the ultimatum game (Eckel and Grossman, 2001; Solnick, 2001).

In addition to selecting one task that measures ability and one choice task, our goal was to use one task in which there are no gender differences and one task in which there are gender differences, as documented by the literature. Previous studies indicate that women and men perform equally well on timed math tasks (Hyde et al., 1990; Niederle and Vesterlund, 2007; Reuben et al., 2014). In contrast, Eckel and Grossman (2001) find that women have a higher propensity to accept offers in the ultimatum game.[1]

We find an interesting contrast between first- and second-order beliefs about ability. There is no evidence that men's and women's first-order beliefs differ;[2]

[1] Solnick (2001) reports a result in the opposite direction, but it is not statistically significant.

[2] This result is consistent with other studies that find no gender differences in beliefs about men

however, both men and women believe that such differences may exist. In the math task, 71% of participants believe that most *men* believe that men outscore women. In contrast, only 34% believe that most *women* believe that men outscore women. Moreover, we find no evidence of significant differences between men and women in these second-order beliefs. In the ultimatum game, the finding holds that there is no difference in men's and women's beliefs, either first- or second-order; however, we also find no differences in participants' beliefs about men's beliefs and their beliefs about women's beliefs. We show that these results are robust to the inclusion of covariates and possible inattention.

In summary, even when men and women have similar first-order beliefs, second-order beliefs about men and women can vary substantially. These statistically and economically significant differences in beliefs about men's and women's beliefs may imply different incentives to acquire skills or to engage in the labor market. The varying results we obtain for second-order beliefs in our two tasks suggest that some characteristics are likely to be more susceptible than others to these differences in second-order beliefs. Overall, our results suggest that second-order beliefs are an important, yet relatively unexplored, mechanism that could perpetuate gender gaps regardless of differences in skills or first-order beliefs.

We make two contributions in this paper. First, we develop a generalized incentive-compatible experimental framework to serve as a template for eliciting first- and second-order beliefs about the differences between two populations regarding a measurable characteristic. No other paper to our knowledge has presented a methodology for measuring second-order beliefs about population differences. We discuss our design decisions extensively in Section 3 to facilitate the practitioner's careful choice of 1) the property of a participant's beliefs to target, 2) the function of two population-specific distributions that implies this property, and 3) the experimental protocol to most effectively elicit this function.

In brief, the belief elicitation procedure works as follows. First, we elicit first-order beliefs. In the math task, for example, participants are asked to reveal their belief about who correctly answered more math summations in a timed task—a randomly chosen man or a randomly chosen woman (and by how many summations). Participants' stated beliefs are then compared to a random draw from a sample of people who completed the math task. We use the Binarized Scoring Rule (BSR) (Hossain and Okui, 2013) to incentivize the truthful revelation of beliefs.

After the first-order belief elicitation, we ask participants to reveal what they believe a random man and a random woman chose *when asked the same question they just answered.* Participants are again rewarded based on how close their stated

and women such as Babcock et al. (2017), Bordalo et al. (2019), Moss-Racusin et al. (2012), and Reuben et al. (2014).

belief is to a realized outcome drawn from a sample of first-order beliefs using the BSR. In this intuitive way, participants reveal their second-order beliefs.

Our second contribution is providing the first empirical evidence that second-order beliefs could lead to unequal outcomes. Existing studies of behavioral responses to potential discrimination cannot distinguish whether second-order beliefs or beliefs about others' *preferences* drive those responses. Similar to the distinction between classical statistical (Arrow et al., 1973; Phelps, 1972) and taste-based (Becker, 1957) discrimination models, beliefs about beliefs about measurable characteristics (such as productivity) have different policy implications than beliefs about preferences.

When either second-order beliefs or beliefs about preferences are miscalibrated, policies to correct them must use information of some form, either about true first-order beliefs or about true preferences. For the policy to be effective, that information must be aimed at the correct primitive: second-order beliefs or beliefs about preferences. For example, if women are less likely to major in STEM fields because they believe that STEM professors do not like working with women, then treating second-order beliefs by saying "STEM professors believe women are equally capable in STEM fields as men" would be ineffective. Our results make the case for the explicit study of the role of beliefs about beliefs about population differences in generating unequal outcomes.

We next discuss the relationship of our work to the literature in Section 2. In Section 3, we detail the experimental framework and rationale for each of the design decisions, as well as reasons why practitioners might make different decisions based on their specific research question. In Section 4, we present the results of our implementation of the experimental framework to study beliefs (and beliefs about beliefs) about the differences between men's and women's performance on a math task and choices in the ultimatum game. We conclude in Section 5 with a discussion of the implications of our results and directions for future research.

2 Related Literature

Higher-order beliefs about the strategic sophistication of opponents have received substantial attention in the experimental literature, especially with regard to the "level-k" model (see Crawford et al., 2013 for a survey). The level-k model predicts game behavior based on a player's level of rationality. A level-2 player, for instance, is rational and believes that other players believe they are rational—a second-order belief. Kneeland's (2015) innovative study of strategic sophistication uses a player's chosen strategies in a series of "ring games" to measure lab participants' levels of rationality. She finds that 71% of participants make choices that rely on second- or higher-order beliefs.

More specifically, second-order beliefs with respect to individual actions have been studied in the experimental literature. The most closely related paper to ours in the literature on beliefs and strategic decision-making is Manski and Neri (2013), in which the authors elicit first- and second-order beliefs about actions in a 2×2 game to study consistency between actions and beliefs. The literature on guilt aversion also elicits second-order beliefs and correlates them to own actions (Bacharach et al., 2007; Bellemare et al., 2011; Charness and Dufwenberg, 2006; Dufwenberg and Gneezy, 2000; Guerra and Zizzo, 2004). These studies elicit reflective beliefs of the form: person A's beliefs about person B's beliefs about what person A will do. In contrast, we elicit beliefs of the form: person A's beliefs about person B's beliefs about differences between two populations in some measurable characteristic.

We are the first, to our knowledge, to propose second-order beliefs as a potential mechanism driving gender differences in outcomes. The closest literature to the study of these beliefs is the theoretical work on self-fulfilling prophecies in statistical discrimination models (see Fang and Moro, 2011 for a review). In these models, minority workers choose to invest less in human capital as a rational response to employers' beliefs that minority workers are less likely to invest. Workers' responses rely on their beliefs about employers' beliefs about the differences between minority and majority workers; however, empirical work on statistical discrimination has focused exclusively on employers' beliefs (e.g. Dianat et al., 2022; Altonji and Pierret, 2001; Ewens et al., 2014), not workers' second-order beliefs. This attention on first-order beliefs could be due to the assumption that beliefs are accurate, so changing first-order beliefs will lead to changes in second-order beliefs. Our results on the differences in first- and second-order beliefs, as well as recent work demonstrating inaccurate statistical discrimination (Bohren et al., 2019a), suggest this assumption may not hold in real markets.

A number of papers study mechanisms that could be consistent with second-order beliefs. For example, Alston (2019) shows that women in a lab experiment anticipate discrimination on a sports trivia task and are willing to pay to hide their gender from prospective "employers." Charness et al. (2020) similarly show in a lab experiment that men are twice as likely as women to choose to reveal their gender in a job market for a stereotypically male task. Also in the lab, Manian and Sheth (2021) show that participants believe that other participants are less likely to follow women's advice about how to play a game.

Outside the lab, Glover et al. (2017) find that minority workers in French grocery stores perform better under less-biased supervisors, where bias is measured using an implicit association test. In addition, several natural experiments show that information visible to job seekers, such as a diversity statement or explicit gender requirement, changes application behavior (Kuhn and Shen, 2021; Flory et al., 2021a,b). Lastly, Bohren et al. (2019b) find, in a field experiment using an online math forum, that women's positively-rated posts are favored over men's positively-rated posts, while women with no evaluation history experience discrimination. One explanation for this behavior is that people believe that other people believe women are less competent, so equally positive ratings for a man and a woman imply that the woman is more competent.

Coffman (2014) and Bordalo et al. (2019) study an idea closely related to second-order beliefs. In a series of lab experiments, Bordalo et al. test for the effects of "self-stereotyping" on confidence and behavior. Stereotypes such as "women are bad at math" are first-order beliefs about a measurable characteristic. In order to self-stereotype, a person must have beliefs about what those stereotypes are; therefore, when stereotypes can be classified as first-order beliefs, the person uses their *second-order beliefs* to self-stereotype.

Exploring another type of beliefs-based mechanism, Babcock et al. (2017) consider how the distribution of low-promotability tasks may impede women's career progression. They find that beliefs about willingness to accept these low-promotability tasks are a primary driver of their inequitable distribution in the lab. In a related thread of literature that studies beliefs about social norms, Bursztyn et al. (2020) measure and treat men's beliefs about other men's opinions about women working outside the home in Saudi Arabia.[3]

While we are the first to study second-order beliefs about population differences, *first*-order beliefs about population differences have been studied in the lab using various elicitation procedures.[4] Some of these procedures are indirect, where beliefs are inferred from actions. For instance, in Aguiar et al. (2009) participants choose whether they prefer to have a dictator allocation from a man or woman. Similarly, Castillo and Petrie (2010) and Fershtman and Gneezy (2001) infer beliefs about different races or ethnicities from contributions in a public goods game and choices in a trust game, respectively. Beliefs have also been elicited directly. Albrecht et al. (2013) use a price list to elicit beliefs about gender differences in a spatial reasoning task. Reuben et al. (2014) directly elicit expectations about men's and women's performance on a timed math task. Similarly, Schniter and Shields (2014) directly elicit expectations about the choices of young and old people in a trust game.

[3] We differentiate between beliefs about a belief, such as what proportion of women work outside the home, versus beliefs about a preference, such as whether a woman "should" be allowed to work outside the home.

[4] The goal of this paper is not to compare our method of eliciting first-order beliefs to other methods. Rather, we focus on the novelty of eliciting second-order beliefs about population differences.

3 Experimental Framework

In this section, we establish a framework for eliciting first- and second-order beliefs about the differences between two populations. The first step is precisely defining the beliefs of interest. That requires determining what properties of a participant's beliefs answer the research question.

The first-order belief we want to learn about in this experiment is whether a participant believes that, given a random draw from two populations, the characteristic of interest is most likely to be larger for the person drawn from population one or two. The second-order belief we want to learn about is whether a participant believes that random draws from population one/two are most likely to have first-order beliefs that favor one population or the other.

These beliefs may be relevant in a variety of scenarios. With respect to first-order beliefs, consider a professor who must choose between two students to advise—one is a man and the other a woman. The professor might care about who is most likely to be more productive. Then, for second-order beliefs, the woman student might care about whether it is most likely that the professor believes it is most likely that the man or the woman is more productive. In this case, note how the second-order belief can affect behavior. The student might choose to reach out to advisors based on characteristics that she believes are correlated with more favorable beliefs towards women, rather than factors such as compatibility in research interests. Importantly, this strategic scenario that potentially generates unequal outcomes could be caused purely by second-order beliefs, regardless of advisors' first-order beliefs.

Different decision-making models can motivate the targeting of other properties of first-order beliefs. For instance, the relevant property for a profit-maximizing employer would be the difference in *expected* productivity between a man and woman, rather than whether a man or woman is *most likely* to be more productive. While we have chosen a simple, binary property of first-order beliefs to study, our procedure may be adapted to study beliefs about the mean difference between two populations, various quantiles of that difference, or other relevant properties.

The second-order belief of interest also depends on the decision-making model. There are two choices that determine the nature of the elicited second-order belief. The choice of the property of the participant's first-order belief determines the relevant second-order belief distribution. Then, just as in first-order beliefs, the experimenter needs to choose the property of interest for the participant's second-order belief distribution. From the example above, choosing to target first-order beliefs about the difference in expected productivity between a man and woman worker, instead of whether the man or woman is most likely to be more productive, changes the relevant second-order belief distribution.

3.1 Operationalizing the Property of Interest

To operationalize our property of interest, whether a participant believes that the characteristic of interest is most likely to be larger for the person drawn from population one or two, let X_1 be the random variable measuring the characteristic of interest in population one and X_2 be the same in population two. Then, we want to learn if the participant believes that these distributions have the property $P(X_1 > X_2) \geq \frac{1}{2}$ or that $P(X_1 < X_2) \geq \frac{1}{2}$. Either condition implies that from a randomly selected pair, the most likely outcome is the person from group one (or respectively two) has a higher value in the measure of interest.

For second-order beliefs, we want to learn whether a participant believes that, when we take a random draw from the first-order belief distribution, the belief is most likely to favor population one or two. So, we want to learn if a participant believes that a random draw from population one (or two) is most likely to *believe* that $P(X_1 > X_2) \geq \frac{1}{2}$ or $P(X_1 < X_2) \geq \frac{1}{2}$.

The next step is to determine what function of a participant's subjective belief distributions reflects the property of interest. In the next subsection, we explain why we choose to elicit the median of the distribution of differences between population-specific distributions. We discuss the alternatives in some detail in Appendix A so the practitioner can make an informed decision based on their own property of interest.

3.2 The Median Difference

The property we describe, $P(X_1 > X_2) \geq \frac{1}{2}$ or $P(X_1 < X_2) \geq \frac{1}{2}$, is implied by a statement about the median of the distribution of *differences between the populations*, $X_1 - X_2$. When there exists a median strictly greater than zero:

$$P\left(X_1 - X_2 \geq Median(X_1 - X_2)\right) \geq \frac{1}{2} \Rightarrow P(X_1 - X_2 > 0) \geq \frac{1}{2} \Leftrightarrow P(X_1 > X_2) \geq \frac{1}{2}.$$

By the same argument, the existence of a median strictly below zero implies $P(X_1 < X_2) \geq \frac{1}{2}$. By eliciting the median of the population difference $X_1 - X_2$, we elicit the participant's first-order belief regarding which population (if any) is most likely to have a higher value in the measure of interest.

Now let Z_1 be the random variable measuring first-order beliefs in population one and Z_2 be the same in population two. Draws from Z_1 or Z_2 are draws of beliefs about the median of the population difference $X_1 - X_2$. A belief that $Median(Z_1) > 0$ implies a participant believes the probability a person from population one believes that $Median(X_1 - X_2) > 0$ is at least $\frac{1}{2}$. Thus, the belief that $Median(Z_1) > 0$ means the participant believes there is at least a $\frac{1}{2}$ probability that a randomly chosen person from population one believes that $P(X_1 > X_2) \geq \frac{1}{2}$.

While we choose to elicit similar properties of first- and second-order beliefs, it is possible (and reasonable) to choose different properties. For example, the practitioner interested in learning about managers' first-order beliefs regarding the expected differences in productivity may still want to learn about second-order beliefs about the most likely first-order belief. In this case, it would be appropriate to elicit the *mean* of the managers' first order beliefs, but the *median* of second-order beliefs.

3.3 Incentive Structure

When eliciting beliefs, the first priority in experiment design is incentivizing truthful revelation. We begin with a payment structure that is incentive-compatible for all expected utility maximizers and some non-expected utility maximizers. The Binarized Scoring Rule (BSR), generalized by Hossain and Okui (2013), works by taking any proper scoring rule (i.e., a payment rule that reaches its maximum under truthfulness) and binarizing it, so that participants are maximizing the probability of winning the "large" prize rather than maximizing the size of the prize. This change in objective makes the payment rule incentive-compatible for all risk preferences. Using "probability currency" to induce risk-neutral behavior has a long tradition in experimental economics (Smith, 1961; Roth and Malouf, 1979), and similar binary procedures for belief elicitation are discussed by Karni (2009), Schlag et al. (2013), and Qu (2012). Schlag et al. specifically discuss binary lotteries for eliciting medians.

The probabilistic structure of the BSR outperforms other payment rules, such as the popular Quadratic Scoring Rule (QSR) introduced by Brier (1950) (Hossain and Okui, 2013). The QSR incentivizes participants by varying the amount of money earned, rather than the probability of earning some fixed amount of money. That is, the closer a participant's predicted value is to the random realization, the more money they earn. This rule works for risk-neutral participants, but risk-averse participants would be incentivized to "hedge" their guess. Hossain and Okui show that participants in a lab experiment report more accurate beliefs under the BSR compared to the QSR when reporting probabilities, but the rules perform equally well in eliciting means, as theory predicts. In general, incentivized belief elicitation outperforms non-incentivized elicitation (Trautmann and van de Kuilen, 2015), particularly when there is a social cost to revealing beliefs as is the case with gendered beliefs (Babin, 2019).[5]

[5] Danz et al. (2020) have recently demonstrated that certain implementations of the binarized version of the quadratic scoring rule may be subject to a pull-to-center effect (participants' stated beliefs are biased towards 50%); however, we do not elicit a probability, our implementation is substantially different, and such an effect would only make our results more conservative. Charness

The BSR proceeds as follows: participants in the experiment win either prize A or prize B, with the value of A exceeding the value of B: $U(A) > U(B)$. We are interested in the random variable X. Participants report $\theta \in \Theta$ where θ is the participant's predicted value of some function of X. A loss function $l(x, \theta)$ returns the prediction error from a random realization of X and the participant's predicted value θ. The experimenter compares the prediction error to a random draw K from a uniform distribution $U(0, \overline{K})$. If the prediction error is less than K, the participant wins prize A. Otherwise, the participant wins the lesser prize B. The form of the loss function determines which function of X participants should report. For example, the BSR would elicit the mean by binarizing the QSR loss function $(x-\theta)^2$. Other payment rules elicit the mode or quantiles.

Since only two prizes are involved, and the size of the smaller prize is the complement of the larger prize, the BSR procedure for some loss function $l(x, \theta)$ can be summarized by probability of winning the large prize A which is given formally as:

$$P(A) = 1 - \frac{l(x, \theta)}{\overline{K}}$$

As discussed in the previous subsection, we are interested in the median of participants' subjective distributions. The loss function for the median is $|x - \theta|$, so in our experiment

$$P(A) = 1 - \frac{|x - \theta|}{\overline{K}} \tag{1}$$

For the first-order belief, x is defined as a draw from the distribution of differences in the population: $X_1 - X_2$. For the second-order belief elicitation, x is defined as a draw from the distribution of beliefs in the relevant population: Z_1 or Z_2.

A sufficient assumption on the utility function for the incentive-compatibility of the BSR to hold is *monotonicity with respect to stochastic dominance* (Hossain and Okui, 2013), originally defined by Machina and Schmeidler (1992). Although the monotonicity assumption is not satisfied by the expected utility functions in prospect theory (Kahneman and Tversky, 1979 and 1983), *Theorem 4* of Hossain and Okui extends the incentive-compatibility of the BSR to account for this type of preferences. The incentive-compatibility of the BSR holds in this case when the participant treats the large prize as a gain and the small prize as a loss. We have

et al. (2020) conclude that introspection does no worse than complex belief elicitation methods. Like Danz et al., Charness et al. focus on the elicitation of probabilities that have an obvious objective value, whereas we elicit a median (not a probability) about an unknown (to the participants) distribution.

followed the advice of Hossain and Okui in setting the small prize to zero.[6]

While our procedure is incentive-compatible for many decision theories discussed in the literature, there is a possibility that the formal incentive compatibility does not extend to *some* decision process used by one of our participants. This has the potential to impair our interpretation of the elicited value as the median; however, note that for any random variable X, $P(X > 0) \geq \frac{1}{2}$ is implied by *any* quantile of X below 50% being larger than zero since the quantile function is increasing on $(0, 1)$. Thus, for some hypothesized decision theory to impair our interpretation of the ordinal information we collect, it would have to lead participants to report a quantile of their subjective belief *above* 50%.

3.4 Generating Samples for Incentives

In order to pay participants using the BSR, we need a sample from which to draw realizations. We pay participants for their first-order belief elicitation by sampling the measure of interest from populations one and two, X_1 and X_2. Then, we pay participants for their second-order belief elicitation by sampling from the first-order beliefs of populations one and two, Z_1 and Z_2. Therefore, we need two samples: one measuring the characteristic of interest and the other measuring first-order beliefs.[7]

The sample measuring the characteristic of interest can be generated as part of the experiment or taken from an existing data source (e.g., past experiments or administrative data). For example, the publicly available population distributions of SAT scores by gender can be sampled to incentivize elicitation of beliefs about the differences in men's and women's SAT performances. If the experimenter generates the data themselves, a single participant can be treated as a random draw from the population. Large samples are not needed—the measurement of one person from each population is sufficient.[8]

The characteristics of interest in this experiment are choices in the ultimatum game and scores on a timed math task. In the ultimatum game, called "Task 1" in the experiment, Player 1 is endowed with \$10 and must decide how much to offer Player 2. Player 2 decides whether to accept Player 1's offer, or to reject, in which case both participants receive nothing. We use the strategy method to elicit participants' choices as both Player 1 and Player 2. Our measure of interest is Player 2's minimum acceptable offer (MAO), the smallest amount Player 1 could propose

[6] When losing the lottery, the subjects still leave with their show-up fee of \$5; however, we believe the participants treat this as endowed wealth at the time of assessing the lotteries. The instructions re-enforce this by emphasizing that a loss in the lottery leads to zero gain.

[7] Note that we cannot measure both in the same sample, since we need the former to pay participants in the latter.

[8] While only one data point from each distribution is needed to incentivize belief elicitation, that data point must be truly random from the perspective of the participants.

such that the participant would accept. Online Appendix A shows the instructions for the strategy-style ultimatum game.

Any differences between men's and women's MAOs (their *willingness to accept*) can be interpreted in multiple ways. First, since any amount above \$0 generates a higher payoff than rejecting, a participant interested only in maximizing earnings accepts any offer above \$0. A higher MAO indicates that the participant is motivated by more than earnings and may be interested in fairness, inequality aversion, competitiveness, etc. Since the ultimatum game has the structure of a take-it-or-leave-it offer in negotiation, differences in MAO can also be interpreted in that context. For instance, women's higher propensity to accept offers in Eckel & Grossman (2001) could be due to social norms dictating that women should be more cooperative or less demanding.[9]

In Task 2, the math task, participants add sets of five two-digit numbers for five minutes. Participants are paid \$0.50 for each correct sum. Online Appendix B shows the instructions for the math task. Previous work (Niederle and Vesterlund, 2007; Reuben et al., 2014) uses timed arithmetic tasks because women and men perform equally well on them (see also Hyde et al., 1990). Despite this, people believe that men score higher than women in math tasks (Reuben et al., 2014).

Unlike the sample measuring the characteristic of interest, the sample measuring first-order beliefs should be collected using the belief elicitation procedure detailed here. The measurement of second-order beliefs relies on the recursive nature of our procedure (a belief about a belief is measured in the same terms as the original belief) to help participants understand the procedure. In other words, to intuitively define second-order beliefs, we need to be able to tell participants that other participants who we are asking about *answered the same questions they just did.*

Like the sample measuring the characteristic of interest, the sample of first-order beliefs can be as small as one person from each population. For example, in this experiment, the measurement of the characteristics of interest in one man and one woman would be sufficient to elicit first-order beliefs. Likewise, the elicitation of first-order beliefs from one man and one woman would be sufficient to incentivize the elicitation of second-order beliefs. To the participant, it does not matter if the random draw used to incentivize them is from a sample of 1 or from a sample of 1,000 because the sample itself is a random draw from the population.

3.5 Belief Elicitation

The belief elicitation procedure begins with the first-order belief elicitations about the characteristics of interest. We elicit participants' first-order beliefs by asking

[9] Solnick (2001) finds that women have higher MAOs, but the difference is not statistically significant at traditional levels.

them to report who they believe performed "better"[10]—a randomly drawn person from population one or a randomly drawn person from population two—and by how much. For the math task in our experiment, we ask who answered more summations correctly and, for the ultimatum game, who chose the higher MAO. We randomize which of the two first-order beliefs we elicit first. Participants report their beliefs by moving a slider like the one presented in Figure 1. The sequence of probabilities reported in the accompanying table are determined by equation (1), which converts the loss function for the median to a binary form and gives the probability of winning the large prize for any report θ and outcome x.

Our implementation of the BSR communicates all relevant incentive information without teaching participants the complex payment rule. The interactive slider allows participants to observe, for every possible stated belief, their probability of winning in every realization of the random draw. Presenting this summary of the payment rule, rather than the payment rule itself, also eliminates the need for specialized mathematical knowledge, thereby increasing the range of potential applications. Simplicity in the belief elicitation procedure is particularly important in our experimental framework because we want to elicit second-order beliefs. To incentivize truthful revelation of second-order beliefs, participants must *believe* that other participants are incentivized to tell the truth about their first-order beliefs.

The slider's starting position is always the center, reporting that the man and woman scored equally in the task. Participants move the slider to the right if they believe the randomly selected man scored higher on the math task (or chose a higher MAO) and to the left if they believe the randomly selected woman scored higher (or chose a higher MAO). When the participant moves the slider, the table updates at each point of the support to show the associated sequence of probabilities of winning the large prize based on each possible realization of the random draw. Participants are told in the instructions that the procedure is designed such that it is optimal to report their best guess about the median.

Implementing equation (1), the binarized form of the loss function for the median, requires a choice for $\overline{K}$. Recall that $\overline{K}$ is the maximum on the uniform distribution from which we take a draw to compare to the evaluated loss function. That means $\overline{K}$ determines the size of the support over which participants can express their beliefs. There are trade-offs in the selection of $\overline{K}$. The wider the support, the flatter the slope of the objective function, weakening the incentive to be precise; however, a narrow range for $\overline{K}$ might truncate the choices of participants with more extreme beliefs. We choose to elicit beliefs over a 21 point support for both tasks: gender neutrality at zero and ten points on either side. This support matches the natural

[10] We put "better" in quotation marks because in tasks like the ultimatum game, it is unclear whether a higher or lower MAO is better. This language is not used in the experiment.

maximum of the ultimatum game, in which the largest difference is between a MAO of \$10 and \$0. Since there is no natural maximum for the math task, the choice might constrain our participants, so we label the endpoints as “10+”.[11]

After eliciting participants’ first-order beliefs, the belief elicitation procedure continues by informing participants that people from populations one and two answered the same questions they just did.[12] We elicit second-order beliefs by asking participants to report what they believe a randomly drawn person from population one (and two) reported when *they* answered those questions. In our experiment, we ask participants what they believe a randomly chosen man from a previous session reported and, likewise, what a randomly chosen woman reported as her first-order belief for each characteristic. That is, we elicit four second-order beliefs— one for each gender/characteristic pair. We randomize the sequence in which these four second-order beliefs are elicited. Participants report their beliefs using a slider like the one in Figure 2. Screenshots of our experiment are available in Online Appendix B.

While we collect cardinal information about participants’ median beliefs, the median was chosen only because it has an ordinal interpretation about underlying probabilities. The additional cardinal information may be interesting, but the cardinal results confound two factors: the magnitude of participants’ beliefs about population differences and participants’ beliefs about absolute levels of characteristics in the populations.

To illustrate this point, consider a participant who reports that their median belief is that a randomly selected man answers two more summations correctly than a randomly selected woman. The interpretation of those “two more summations” differs based on whether the participant believes people answer five summations total on average or twenty summations. Moreover, it is unclear how the additional quantitative results would be more informative than ordinal results. For example, knowing that people believe that men believe men outscore women on a simple math task may inform our understanding of the employment gap in STEM fields, but knowing specifically how many more math summations they are believed to outscore women by on this one particular task would not. Thus, in the results section we focus on the ordinal information provided by the median beliefs.

[11] We observe very few results at the endpoints. See Appendix Figures B1 and B2.

[12] The procedure we use precludes randomizing the order of the first- and second-order tasks. Even if we were to randomize the order of the elicitation of first- and second-order beliefs, we would need to give participants information on the first-order belief elicitation in order to ask them about their second-order beliefs. In other words, we must expose participants to the “object” of the belief elicitation (first-order beliefs) prior to eliciting the belief, so randomizing the order would likely be negated through exposure anyway.

3.6 Salience of Gender

We elect to make gender salient in our procedure, rather than try to disguise our intentions. Experimenters often obfuscate the purpose of an experiment about gender to avoid confounding factors such as an experimenter demand effect or social costs associated with revealing gendered beliefs. For example, one concern with our procedure is that *most* of the possible choices involve expressing some difference between men and women. This could create a demand effect, leading participants with neutral beliefs to express differences. On the other hand, revealing beliefs that "favor" one gender over another could impose some social cost on participants. This cost would bias results towards zero. Instead of attempting to design our experiment to neutralize these biases, we rely on our relatively strong and carefully designed monetary incentives to ensure that our results indicate true patterns in participants' beliefs.

Obfuscating gender is particularly untenable in our experiment because we want to elicit second-order beliefs. To elicit true second-order beliefs in our framework, it is vital that participants clearly understand that they are revealing their beliefs about men and women *and believe that other people clearly understood* that they were revealing their beliefs about men and women. When gender is obfuscated, this requirement becomes more burdensome, since participants must also believe that other participants saw and interpreted the signal of gender in the same way they did.

Even supposing that participants all interpret the signal of gender identically, obfuscating gender in both the first- and second-order belief elicitation means that participants reporting their second-order belief would need to deduce both the gender of the person in the first-order belief elicitation and the implied gender difference that person is asked about. There are two common methods of obfuscating gender: using physical identity (e.g. Reuben et al., 2014) or names (e.g. Bertrand and Mullainathan, 2004), rather than directly saying "a man" or "a woman." Consider, for example, if we used names in our experiment. Our first-order belief elicitation would have to ask each participant their beliefs about the difference in math scores of some named woman and some named man. In our second-order elicitation, we would have to show the name of the person whose first-order belief is relevant *and* the pair of names about whom that participant reported their first-order beliefs.[13]

[13] We could also have made gender directly salient in the first-order belief elicitation by asking about "a woman" and "a man" but then obfuscated gender in the second-order belief elicitation by showing the name of the person whose first-order belief is relevant. However, we find it unlikely that this obfuscation would be successful, since we would still have to make clear that this participant was asked their first-order beliefs about (a non-obfuscated) woman and man.

3.7 Implementation

We implemented this experiment at the Vanderbilt University Experimental Economics Lab (VUEEL) from November 2017 to January 2018 and online on Prolific in February 2022. In-person participants were recruited using the ORSEE system (Greiner, 2015). We stratified the sampling of in-person participants by gender and of online participants by gender and STEM versus non-STEM major. For both the in-person and online experiment, we recruited people who identified as either a man or a woman.[14]

No one participated in more than one session of the experiment. In-person participants had participated in at maximum one experiment (unrelated to this one) 7 months prior to the start of this experiment in the VUEEL.[15] Online participants had a wide range of experience in previous experiments.

The belief elicitation data come from a total of 354 participants, 175 of whom are men and 179 of whom are women. Table 1 lists the sample sizes by gender for the samples used to incentivize belief elicitations as well as the in-person and online samples that generate our belief elicitation data.[16] Note that all participants in the samples used to incentivize belief elicitations participated in-person at the VUEEL, as was explained to both the online and in-person participants.

Participants in the in-person sessions received paper copies of the instructions used to measure the characteristics of interest, but completed the experiment on laptops using the oTree software (Chen et al., 2016). All instructions were read aloud by the experimenter. Participants had to pause at two points in the experiment to wait for all other participants to complete the previous section so the experimenter could read the instructions to everyone at the same time (after the example and after the first-order belief elicitation). After the belief elicitations, the experiment concluded with a demographic survey. Each session lasted approximately 30 minutes.[17]

Participants in the online sessions were shown PDFs of the instructions used to measure the characteristics of interest, which they could access during the belief elicitations. Like the in-person experiment, the experiment was programmed with the oTree software (Chen et al., 2016). In lieu of the experimenter reading the instructions aloud, audio files reading the text were provided on each page of instructions. Instruction pages had a timer the length of the audio file to ensure participants had

[14] Our ORSEE registration page and Prolific both allow people to identify as non-binary.

[15] We know that 19 participants, 8 men and 11 women, from the first experiment participated in this experiment (the second in the VUEEL); however, our ORSEE data is not connected to our experimental data, so we do not know which of the 157 participants are the 19 who participated in the first experiment.

[16] Recall that we need only one draw from each population for each sample, but we collect slightly more.

[17] The time from the actual start of the experiment to when all participants completed the six belief elicitations and demographic survey was typically 15 to 20 minutes.

sufficient time to review the instructions. Outside of the instruction pages, participants were allowed to proceed at their own pace, but we recorded the amount of time they spent on each page.[18] The average time to complete the experiment was about 13 minutes. See Online Appendix B for screenshots of the full experiment.

Table 2 shows the descriptive statistics of our combined sample by gender. We observe that men are slightly more likely to major in STEM fields and, in the online sessions, men have participated in more previous experiments than women.[19] In Appendix Table B1, we show that online participants are more likely to major in STEM fields and identify as white, have less educated parents, and are older.

In-person participants received \$5 for participating in the experiment and could earn the "large" prize of \$15 from the belief elicitations. Online participants received \$3 for participating and could earn \$5 from the belief elicitations. In both, one decision out of the six was chosen at random at the end of the experiment to determine payment.[20] In-person participants earned \$18.09 on average and online participants earned \$7.11 on average, including the participation fee.

4 Results

We present the experimental results for the math task, followed by the ultimatum task and an intra-participant comparison of beliefs. We do not have predefined hypotheses about these belief distributions. One way to develop such hypotheses would be to use "common knowledge" arguments; however, the beliefs underlying those common knowledge arguments are precisely what we are seeking to measure. We describe the data instead.

4.1 Math Task

Beliefs pertaining to the math task are summarized in Table 3 and Figure 3. Panel A of Table 3 presents first-order beliefs. Column 1 shows that most participants believe that there is *some* difference in men's and women's performance on the math task (81%, SE = 2.1%), with 46% (SE = 2.6%) believing that men outscore women. Testing for a difference in the gender-specific proportions reported in Columns 2 and 3, we cannot reject at conventional significance levels that men and women have the same probability of believing that men outscore women (50% for men vs.

[18] Online Appendix C shows that our main results are robust to removing online participants who may have been inattentive, based on the amount of time they spent on the belief elicitations.

[19] Note that doing many experiments on Prolific is different from doing many experiments in a typical in-person lab. First, the experiments are shorter (one- and two-minute experiments are prevalent). Second, participants on Prolific can do many (although not an unlimited number) experiments each day.

[20] There were two first-order belief elicitations and four second-order belief elicitations.

42% for women, $p = 0.114$). The finding of a statistically insignificant gender gap in man-favoring beliefs is robust to controlling for participant covariates in a linear probability model, as reported in Columns 1 and 2 of Appendix Table B2.

Similarly, using a Wilcoxon rank-sum test, we cannot reject that men's and women's first-order belief distributions are identical ($p = 0.127$).[21] This result is robust to the inclusion of participant covariates, operationalized by estimating proportional odds (ordered logit) models with the ternary belief as the dependent variable, as reported in columns 1 and 2 of Appendix Table B3.[22] We note, however, that we cannot rule out a range of differences in first-order beliefs, including both positive and negative differences. For example, the 95% confidence interval for the men-women gap in the proportion believing that men outscored women is $[-2\%, 19\%]$.

Although we find no evidence that first-order beliefs differ by gender, participants believe that such differences in first-order beliefs exist. Using the Wilcoxon signed-rank test, we reject equality of distributions of second-order beliefs about men's beliefs and women's beliefs regarding math performance ($p < 0.001$).[23] Furthermore, as reported in Column 1 of Table 3, Panel B, 71% (SE=2.4%) of participants believe that most men believe men outscore women, while Column 1 of Panel C shows that only 34% (SE=2.5%) of participants believe this about women's first-order beliefs. A test of difference in proportions rejects that they are equal ($p < 0.001$).[24]

As with the first-order beliefs, we cannot reject that men's and women's second-order belief distributions are identical, with respect to either men's or women's first-order beliefs. We cannot reject that men and women have the same probability of believing that most men believe men outscore women (68% for men vs. 74% for women, $p = 0.234$) or that most women believe men outscore women (30% for men vs. 38% for women, $p = 0.127$). Controlling for covariates in Columns 3 through 6 of Appendix Table B2 does not affect the finding regarding second-order beliefs about men, but it does result in a marginally significant 9.2 p.p. ($p = 0.074$)

[21] The Wilcoxon rank-sum test for equality of first-order belief distributions uses the ternary distributions illustrated in Figure 3. Recall that we collect cardinal information, even though our outcome of interest is ternary. The Wilcoxon rank-sum test for equality of the first-order *cardinal* distributions gives $p = 0.173$. The cardinal distributions for all elicitations are in Appendix Figures B1 and B2.

[22] When participant gender is the lone covariate in the proportional odds model, the p-value for the test that its coefficient equals zero is numerically identical to the p-value from the Wilcoxon rank-sum test. Adding covariates generalizes the Wilcoxon rank-sum test by allowing for inclusion of further covariates.

[23] The Wilcoxon signed-rank test is used to account for intra-participant dependence. The Wilcoxon signed-rank test also rejects equality of cardinal second-order belief distributions ($p < 0.001$).

[24] This result will be explored further, including demonstrating robustness to participant covariates, in Section 4.3.

gender difference in second-order beliefs about women.[25] This result suggests that, conditional on observable participant characteristics, women may be more likely than men to believe most women report that men outscore women in the math task.

Examining the ternary belief second-order belief distributions, we again cannot reject that men and women hold the same beliefs with respect to either men's ($p = 0.352$) or women's ($p = 0.415$) first-order beliefs.[26] This finding is robust to the inclusion of covariates (Appendix Table B3, Columns 3 through 6).

While this experiment lacks the statistical power to make definitive statements about whether participants' second-order beliefs are correctly calibrated, some conclusions are possible.[27] The 95% confidence set for the median of men's first-order ternary belief distribution includes both "no difference between man and woman" and "man outscores woman," but excludes "woman outscores man." The same is true for women's first-order beliefs. Thus, only a reported second-order belief (about either a man's or a woman's reported first-order belief) of "woman outscores man" can be classified as miscalibrated. Participants are much more likely to report this miscalibrated second-order belief about women (42%, $SE = 2.6\%$) than about men (14%, $SE = 1.8\%$), a difference that is statistically significant ($p < 0.001$). To the extent that we can detect miscalibrated second-order beliefs in the data, it seems that the difference in gender-specific second-order beliefs is driven by participants wrongly believing that most women's first-order beliefs "favor" women.

To summarize, we are unable to reject equality in first- and second-order belief distributions between men and women, but have strong evidence that participants *believe* men and women hold different first-order beliefs. In particular, most participants believe that most men believe men outscore women, while they do not believe this about women.

4.2 Ultimatum Task

Beliefs about the ultimatum task are summarized in Table 4 and Figure 4. First-order beliefs are reported in Panel A of Table 4. As shown in Column 1, most participants believe that men choose a higher MAO than women (62%, $SE = 2.6\%$). Similar to the math task, we cannot reject that the proportions of men and women

[25] This estimate is close in magnitude to the estimate that does not control for covariates (7.7 p.p.).

[26] We again use the Wilcoxon rank-sum test for equality of distributions, comparing the gender-specific ternary distributions. The null hypothesis of no differences in the gender-specific cardinal second-order belief distributions cannot be rejected for beliefs about women ($p = 0.590$), while there is marginal evidence for differences in cardinal second-order belief distributions about men ($p = 0.060$).

[27] Here, calibration refers to how well second-order beliefs match median first-order beliefs in the population for which beliefs were being elicited—participants in the Vanderbilt experimental sessions.

believing that men report a higher MAO are equal ($p = 0.351$).[28] This finding is robust to the inclusion of covariates in Appendix Table B4, Columns 1 and 2. Nor can we reject that the distributions of men's and women's first-order beliefs are the same ($p = 0.456$), which is again robust to including covariates in Appendix Table B5.[29] As in the math task, we cannot rule out positive or negative differences in beliefs: the 95% confidence interval for the men-women gap in the proportion believing that men had a higher MAO is $[-15\%, 5\%]$. Interestingly, the point estimates suggest that men are 4.8 percentage points (SE=5.1 p.p.) *less likely* than women to hold the "stereotypical" belief that men have a higher MAO than women, although this difference is not statistically significant at conventional levels.

In contrast to the math task, we cannot reject that participants believe men and women hold the same first-order beliefs about gender differences in MAO ($p = 0.182$).[30] Column 1 of Panel C shows that 61% (SE=2.6%) of participants believe that most women believe men choose a higher MAO, which is higher than the 58% (SE=2.6%) of participants (Panel B) believing this about men's first-order beliefs, but not statistically significantly so ($p = 0.346$).

Again, we cannot reject that men and women hold similar second-order beliefs about either men's or women's first-order beliefs. We cannot reject that men and women have the same probability of believing that most men believe men report a higher MAO (60% for men vs. 55% for women, $p = 0.372$) or that most women believe men report a higher MAO (59% for men vs. 63% for women, $p = 0.545$). These findings are robust to controlling for covariates in Columns 3 through 6 of Appendix Table B4.

Similarly, we cannot reject the null hypothesis that men and women have identical ternary second-order belief distributions about men ($p = 0.565$) or about women ($p = 0.348$).[31] After adding covariates in Columns 3 through 6 of Appendix Table B5, the results with respect to second-order beliefs about men are unchanged, but gender differences in second-order beliefs about women are now marginally significant. That is, after accounting for observable gender differences, woman participants may believe women's beliefs are more man-favoring than man participants do.

The 95% confidence sets for the medians of men's and women's first-order ternary

[28] The tests used for the ultimatum task analysis are the same as those used for the math task: tests for differences in proportions when comparing proportions across genders, Wilcoxon rank-sum tests when testing for differences in ternary distributions between genders, and Wilcoxon signed-rank tests when testing for within-participant differences in second-order beliefs with respect to different genders.

[29] The test for differences in the cardinal distributions does find marginal evidence of a difference ($p = 0.063$), but given the concerns with interpreting the cardinal measures and the lack of evidence for differences between the ternary distributions, we do not interpret this finding further.

[30] We can, however, reject equality of the cardinal second-order belief distributions ($p = 0.014$).

[31] We also fail to reject equality of the cardinal second-order belief distributions about men ($p = 0.634$) or about women ($p = 0.417$).

belief distributions include only "man has higher MAO than woman," meaning that all other second-order beliefs are miscalibrated. Second-order beliefs about both genders are often miscalibrated: 42% (SE=2.6%) of second-order beliefs about men are miscalibrated, as are 39% (SE=2.6%) about women. This difference in the rate of miscalibration between genders is statistically insignificant ($p = 0.359$).

Similar to the math task, we do not have consistent evidence of gender differences in either first- or second-order beliefs about the ultimatum task. Furthermore, and different from the math task, we have no evidence that participants believe that men and women differ in their first-order beliefs.

4.3 Intra-participant Beliefs

In this section, we consider two types of intra-participants beliefs. First, we describe the extent to which participants' second-order beliefs mirror their own first-order beliefs. This analysis is useful in understanding whether people form second-order beliefs about others solely by considering their own beliefs.[32] If it were true that second-order beliefs simply mirror first-order beliefs, then we would have no need to elicit those second-order beliefs. We could infer them from first-order beliefs.

We first compare a participant's reported first-order belief to their second-order belief about a person of the same gender. Column 1 of Table 5 shows that, while the majority of participants believe that other participants of their same gender believe the same as themselves (60%, $SE = 2.6\%$ for the math task and 66%, $SE = 2.5\%$ for the ultimatum task), these proportions are far from 1 and are quite similar for men and women. That is, second-order beliefs are informed, but not solely determined, by a participant's own first-order beliefs.

Next, we consider the beliefs that favor men. Tables 6 and 7 report results from estimating a linear probability model with an indicator for reporting a second-order belief that favors men as the dependent variable. The regressors of interest are participant gender and its interaction with an indicator for holding a first-order belief that favors men.

For the math task, Column 1 of Table 6 shows that, conditional on *not* holding a man-favoring first-order belief, men are 16 p.p. less likely than women to believe that men's beliefs are man-favoring. For both man and woman participants, holding a man-favoring first-order belief predicts a higher probability of holding man-favoring second-order beliefs. This relationship is larger for men (35 p.p., $SE = 6.6$ p.p) than for women (18 p.p., $SE = 6.3$ p.p.), and this gender difference is marginally significant ($p = 0.064$). Taken together, the coefficients imply that participants with man-favoring first-order beliefs are about equally likely to report man-favoring

[32] The direction of causality could also be reversed – people's beliefs about other people's beliefs could inform their own first-order beliefs, but this directionality is less intuitive.

second-order beliefs about men. These findings hold when controlling for covariates in Column 2.

Column 3 reports analogous results for second-order beliefs about women. Here, there is no apparent gender difference in second-order beliefs among participants who do not hold man-favoring first-order beliefs. Man-favoring first-order beliefs predict a higher probability of reporting man-favoring second-order beliefs among woman participants (31 p.p., $SE = 7.2$ p.p.) and man participants (17 p.p., $SE = 6.9$ p.p.). These findings are robust to including covariates in Column 4. These belief-by-gender interaction terms are not statistically different from each other ($p = 0.153$), but they mirror the patterns in the results regarding second-order beliefs about men. That is, in both cases, holding man-favoring first-order beliefs increases the probability of holding man-favoring second-order beliefs, to a larger extent when that second-order belief is about one's own gender.

Repeating this exercise for the ultimatum task in Table 7, we see no apparent gender differences in second-order beliefs about either gender among participants who did not report man-favoring first-order beliefs. Again, holding man-favoring first-order beliefs predicts a higher probability of reporting man-favoring second-order beliefs about men and women. Gender differences in this belief-by-gender interaction are muted compared to those in the math task, and we cannot reject equal coefficients for second-order beliefs about men ($p = 0.272$) or about women ($p = 0.927$).

While we are not aware of comparable experiments against which to benchmark the finding of partial intra-participant concordance between first- and second-order beliefs, we note parallels to two papers in the social norms literature. Bursztyn et al. (2020) find a positive correlation between one's own preferences for women working outside the home and reported beliefs about the proportion of others holding that same preference, but own preferences explain only 4% of variation in these beliefs.[33] Heap et al. (2020) separately elicit social preferences and perception of social norms regarding behavior in a dictator game. There is significant intra-participant disagreement between social preferences and norms. Crucially, perceived norms predict distribution decisions better than preferences do, indicating that reports of own preferences capture something fundamentally different than reports about others' preferences.[34]

The final type of intra-participant belief we consider is the relationship between participants' second-order beliefs. Since beliefs about beliefs may affect many types of decision-making, the difference between a person's beliefs about men and women

[33] The latter finding is based on our own analysis of the paper's replication data.

[34] We return to the relationship between the present experiment and social norms in the discussion section.

could drive sorting behavior, such as occupational sorting or sorting into STEM and non-STEM majors. A woman who believes that men's beliefs generally favor men more than women's beliefs do would be rational to seek environments with more same-gender decision-makers. We explore this relationship using a linear probability model with an indicator for holding the man-favoring second-order belief as the dependent variable, where each observation is a participant-by-task pair. That is, each participant now contributes two observations: one for their second-order belief about men and one for their second-order belief about women. The regressors of interest are indicators for participant gender, the gender about whom the second-order belief is being reported, and their interaction. Standard errors are clustered at the participant level.

Column 1 of Table 8 shows that women believe that men are 36 p.p. (SE=4.5 p.p.) less likely, compared to women, to believe that men outperform women in the math task. Column 3 finds no statistically significant differences in participants' beliefs about men's and women's beliefs on the ultimatum task (-7 p.p., $SE = 4.8$ p.p.). For both tasks, there is no evidence of gender differences in the differences between second-order beliefs about men versus women. Columns 2 and 4 of Table 8 show that these results are robust to the inclusion of covariates, with both stable coefficients and standard errors.[35]

5 Discussion

We establish an experimental framework for measuring both first- and second-order beliefs about the difference in some measurable characteristic between two populations. We implement the procedure in the lab to measure beliefs about the differences between men and women in their performance on a math task and choices in the ultimatum game. Our results are interesting, but intuitive. While men and women exhibit no statistically distinguishable differences in their first-order beliefs, in some cases people *believe* that such differences exist.

Because the experimental design intentionally made gender transparent in all tasks, it is worthwhile to consider alternative explanations for these results and ways that such explanations may be explored. This procedure incentivizes participants to report what they believe another person *reported* as their first-order belief and interprets that elicitation as the participant's second-order belief. Participants who believe there are social costs, experimenter demand effects, or any other biasing

[35] Because all participants report second-order beliefs about both men and women, there is zero correlation between participant covariates and the indicator for gender of the reported second-order belief. Thus, the identical point estimates are expected. Similarly, models including individual-level fixed effects yield identical points estimates for both the second-order belief indicator and its interaction with participant gender.

factors should account for them when reporting their second-order belief. This argument relies on participants being rational enough to consider the incentives of other participants. We believe participants are sophisticated enough to account for the full range of incentives affecting other participants;[36] therefore, a conservative interpretation of our most compelling results would be "participants believe that men and women *reveal* different first-order beliefs" rather than "*have* different beliefs."

Without an experimental design that treats the specific factors that may cause held and reported beliefs to diverge, it is difficult to give a definitive appraisal of this alternative explanation. We emphasize, however, that the present design and implementation have taken some steps to minimize social cost (responses are known to be unobserved by other participants and anonymized in the data) and to overcome psychic costs to revealing gendered beliefs using carefully-designed monetary incentives. Recent research has found experimenter demand effects to play at most a small role in experiments (Mummolo and Peterson, 2019; De Quidt et al., 2018), but this is likely context-specific and, in any case, the relevant question for measuring second-order beliefs is the extent to which participants believe that first-order belief reporters faced such effects.[37] Future work could focus specifically on quantifying the influence of factors that may drive a wedge between held and revealed beliefs.[38]

The potential implications for real-world markets of the discordant beliefs we observe are far-reaching. Consider a woman who believes that men managers believe men to be more productive than women in STEM fields. She may pay some economic cost to be matched with a woman manager rather than a man manager, even though there may be, in fact, no difference in men and women managers' beliefs. These second-order beliefs could contribute to observed gender differences in outcomes like the employment gap in STEM, regardless of differences in first-order beliefs or skills. Beyond the labor market, these second-order beliefs may have important implications in human capital, healthcare, marriage, and fertility decisions.

A number of avenues are open for future work. First, and foremost, is showing whether second-order beliefs affect market behavior. Koutout (2022) takes the first step by showing that workers' second-order beliefs about managers' gendered beliefs affect their job application behavior, but there are many more markets and popula-

[36] This belief is consistent with the results of Kneeland (2015), who finds that a large majority of subjects are at least second-order rational.

[37] The importance of the held versus revealed belief distinction depends on whether the relevant economic action is holding the belief or acting on it. In particular, if the barrier to revealing one's true beliefs is an intrinsic psychic cost, and this cost is borne by the decision-maker both in the lab and in economically-relevant contexts, then the revealed belief is in fact the relevant one to elicit.

[38] To explore the experimenter demand effect, for example, the researcher could follow De Quidt et al. (2018) by introducing a treatment arm that explicitly induces demand in the instructions for the first-order task. Likewise, a treatment arm might vary the observability of reported beliefs in the first-order task. The researcher could then estimate the effect of such "first-order treatments" on second-order beliefs reported by participants regarding treated and untreated populations.

tions to be studied. Second is understanding how second-order beliefs are formed. One promising theory applies the stereotype model in Bordalo et al. (2019). In this model, second-order beliefs are an exaggeration of first-order beliefs.

Researchers can also use our procedure to elicit a broad class of social norms and others' beliefs about what those norms are, although such norms are only a subset of the objects that can be elicited. Acemoglu and Jackson (2017) define descriptive social norms as "the distribution of anticipated payoff-relevant behavior"– what people believe others will *do*.[39] This is a first-order belief about the action of others in a population. In this sense, the first-order belief elicitation about the ultimatum task in this paper measures descriptive social norms regarding a gendered action. The second-order belief elicitation thus measures beliefs about a descriptive social norm.

Relatedly, this procedure can be used to elicit beliefs about injunctive social norms, such as those addressed by Bursztyn et al. (2020). Injunctive norms differ from descriptive norms because they relate to what people *should* do rather than what they *will* do (Cialdini et al., 1990). There is a subtle difference in the procedures for eliciting descriptive and injunctive norms and beliefs about them. Injunctive norms are not first-order beliefs, but they can be elicited directly. For example, Bursztyn et al. (2020) ask men whether they think women should work outside the home. Beliefs about the injunctive norm are simply first-order beliefs about responses to this direct elicitation. Thus, while beliefs about descriptive norms can be measured using second-order belief elicitations, beliefs about injunctive norms only require a first-order elicitation.

While we have focused on gender in this paper, the procedure is sufficiently general to study differences about other types of populations. The experimental framework can be used to elicit beliefs about differences by races/ethnicities, religious beliefs, sexual orientation, STEM/non-STEM workers, and political affiliation. Only small samples from the populations of interest are required to incentivize first- and second-order belief elicitation, enabling the study of beliefs about much smaller and difficult to recruit populations than was previously practical. Second-order beliefs likely play a role in how all of these populations interact with each other, so our experimental framework provides a general tool that can be adapted to study beliefs in most contexts.

[39] This is consistent with the definition of a social custom in Akerlof (1980), which is a behavior "whose utility to the agent performing it in some way depends on the beliefs or actions of other members of the community."

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

References

Acemoglu, D. and Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295.

Aguiar, F., Brañas-Garza, P., Cobo-Reyes, R., Jimenez, N., and Miller, L. M. (2009). Are women expected to be more generous? *Experimental Economics*, 12(1):93–98.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The quarterly journal of economics*, 94(4):749–775.

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

Alston, M. (2019). The (perceived) cost of being female: An experimental investigation of strategic responses to discrimination. *Working paper.*

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The quarterly journal of economics*, 116(1):313–350.

Arrow, K. et al. (1973). The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33.

Babcock, L., Recalde, M. P., Vesterlund, L., and Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714–47.

Babin, J. J. (2019). Detecting group gender stereotypes: Opinion-mining vs. incentivized coordination games. *Journal of Economic Perspectives*, 45(1):21–42.

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63(4):349–388.

Becker, G. S. (1957). *The economics of discrimination: an economic view of racial discrimination.* University of Chicago.

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B., and Doms, M. E. (2011). Women in stem: A gender gap to innovation. *Economics and Statistics Administration Issue Brief*, (04-11).

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2019a). Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019b). The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Brier, G. W. (1950). The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Atmospheric Sciences*, 7(4):283–290.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American economic review*, 110(10):2997–3029.

Castillo, M. and Petrie, R. (2010). Discrimination in the lab: Does information trump appearance? *Games and Economic Behavior*, 68(1):50–59.

Charness, G., Cobo-Reyes, R., Meraglia, S., and Sánchez, Á. (2020). Anticipated discrimination, choices, and performance: Experimental evidence. *European Economic Review*, page 103473.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

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

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015.

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

Crawford, V. P., Costa-Gomes, M. A., and Iriberri, N. (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62.

Danz, D., Vesterlund, L., and Wilson, A. J. (2020). Belief elicitation: Limiting truth telling with information on incentives. Technical report, National Bureau of Economic Research.

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

Dianat, A., Echenique, F., and Yariv, L. (2022). Statistical discrimination and affirmative action in the lab. *Games and Economic Behavior*, 132:41–58.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182.

Eckel, C. C. and Grossman, P. J. (2001). Chivalry and solidarity in ultimatum games. *Economic inquiry*, 39(2):171–188.

Ewens, M., Tomlin, B., and Wang, L. C. (2014). Statistical discrimination or prejudice? a large sample field experiment. *Review of Economics and Statistics*, 96(1):119–134.

Fang, H. and Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. *Handbook of social economics*, 1:133–200.

Fershtman, C. and Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics*, 116(1):351–377.

Flory, J., Leibbrandt, A., Rott, C., and Stoddard, O. (2021a). Signals from on high and the power of growth mindset: A natural field experiment in attracting minorities to high-profile positions.

Flory, J. A., Leibbrandt, A., Rott, C., and Stoddard, O. (2021b). Increasing workplace diversity evidence from a recruiting experiment at a fortune 500 company. *Journal of Human Resources*, 56(1):73–92.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, 55(1):25–30.

Heap, S. P. H., Matakos, K., and Weber, N. S. (2020). Non-selfish behaviour: Are social preferences or social norms revealed in distribution decisions?

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

Hyde, J. S., Fennema, E., and Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin*, 107(2):139.

Kahneman, D. and Tversky, A. (1979). On the interpretation of intuitive probability: A reply to jonathan cohen.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606.

Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica*, 83(5):2065–2079.

Koutout, K. (2022). Gendered beliefs and the job application decision: Evidence from a large-scale field and lab experiment. *Available at SSRN 4035946*.

Kuhn, P. J. and Shen, K. (2021). What happens when employers can no longer discriminate in job ads? Technical report, National Bureau of Economic Research.

Lundberg, S. J. (2017). Father absence and the educational gender gap.

Machina, M. J. and Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica: Journal of the Econometric Society*, pages 745–780.

Manian, S. and Sheth, K. (2021). Follow my lead: Assertive cheap talk and the gender gap. *Management Science*, 67(11):6880–6896.

Manski, C. F. and Neri, C. (2013). First-and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior*, 81:232–254.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., and Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479.

Mummolo, J. and Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113(2):517–529.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122(3):1067–1101.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661.

Qu, X. (2012). A mechanism for eliciting a probability distribution. *Economics Letters*, 115(3):399–400.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.

Roth, A. E. and Malouf, M. W. (1979). Game-theoretic models and the role of information in bargaining. *Psychological review*, 86(6):574.

Schlag, K. H. et al. (2013). Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality.

Schniter, E. and Shields, T. W. (2014). Ageism, honesty, and trust. *Journal of Behavioral and Experimental Economics*, 51:19–29.

Smith, C. A. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(1):1–25.

Solnick, S. J. (2001). Gender differences in the ultimatum game. *Economic Inquiry*, 39(2):189–200.

Trautmann, S. T. and van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589):2116–2135.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293.

Tables

Table 1: Sample sizes for incentive and belief elicitation samples, by gender

	(1)	(2)	(3)	(4)
			Belief	Belief
		First-order	elicitation	elicitation
	Task	beliefs only	(in-person)	(online)
Woman	12	4	77	102
Man	10	4	80	95

Notes: Column headers denote the samples. We measure the characteristics of interest in the "task" sample, which is used in incentivizing the first-order belief elicitations. The "first-order beliefs only" sample is used to incentivize second-order belief elicitations for participants in the full "belief elicitation" samples, which provide the data analyzed in the experiment. The "in-person" belief elicitation sample was collected in the Vanderbilt University Experimental Economics Lab, while the "online" belief elicitation sample was collected on Prolific.

Table 2: Participant characteristics

	(1) Men	(2) Women	(3) Difference
STEM major	0.39	0.30	0.10^{*}
	(0.49)	(0.46)	(0.05)
White	0.63	0.58	0.05
	(0.48)	(0.49)	(0.05)
English first language	0.78	0.77	0.02
	(0.41)	(0.42)	(0.04)
Age	23.67	23.88	-0.20
	(7.37)	(7.06)	(0.77)
Mother has graduate degree	0.23	0.27	-0.05
	(0.42)	(0.45)	(0.05)
Mother has bachelor's degree	0.39	0.36	0.03
	(0.49)	(0.48)	(0.05)
Mother has HS or associate's	0.29	0.30	-0.01
	(0.46)	(0.46)	(0.05)
Father has graduate degree	0.34	0.28	0.05
	(0.47)	(0.45)	(0.05)
Father has bachelor's degree	0.31	0.27	0.04
	(0.46)	(0.44)	(0.05)
Father has HS or associate's	0.25	0.32	-0.08
	(0.43)	(0.47)	(0.05)
Online participant	0.54	0.57	-0.03
	(0.50)	(0.50)	(0.05)
Previous experiments (Online participants)	341.59	269.03	72.56^{**}
	(293.16)	(199.44)	(35.98)
Observations	175	179	354

Notes: Gender-specific means and standard deviations (in parentheses) are in columns (1) and (2). Column (3) gives the difference in means and its standard error (in parentheses). "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. "Previous experiments" is a variable we have for online participants only and is equal to the number of experiments the participant completed on Prolific prior to our experiment. * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Belief elicitation results for math task

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. First-Order Beliefs				
W>M	0.345	0.314	0.374	-0.060
	(0.025)	(0.035)	(0.036)	(0.050)
W=M	0.195	0.183	0.207	-0.024
	(0.021)	(0.029)	(0.030)	(0.042)
W<M	0.460	0.503	0.419	0.084
	(0.026)	(0.038)	(0.037)	(0.053)
Panel B. Second-Order Beliefs, about Men				
W>M	0.136	0.131	0.140	-0.008
	(0.018)	(0.026)	(0.026)	(0.036)
W=M	0.155	0.189	0.123	0.066
	(0.019)	(0.030)	(0.025)	(0.038)
W<M	0.709	0.680	0.737	-0.057
	(0.024)	(0.035)	(0.033)	(0.048)
Panel C. Second-Order Beliefs, about Women				
W>M	0.421	0.423	0.419	0.004
	(0.026)	(0.037)	(0.037)	(0.052)
W=M	0.237	0.274	0.201	0.073
	(0.023)	(0.034)	(0.030)	(0.045)
W<M	0.342	0.303	0.380	-0.077
	(0.025)	(0.035)	(0.036)	(0.050)
Observations	354	175	179	354

Notes: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows "W>M", "W=M", and "W<M" report the proportion of participants in the math task who believe that the woman scores higher, the woman scores the same, the woman scores lower compared to the man.

Table 4: Belief elicitation results for ultimatum task

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. First-Order Beliefs				
W>M	0.155	0.154	0.156	-0.002
	(0.019)	(0.027)	(0.027)	(0.039)
W=M	0.220	0.246	0.196	0.050
	(0.022)	(0.033)	(0.030)	(0.044)
W<M	0.624	0.600	0.648	-0.048
	(0.026)	(0.037)	(0.036)	(0.051)
Panel B. Second-Order Beliefs, about Men				
W>M	0.212	0.217	0.207	0.010
	(0.022)	(0.031)	(0.030)	(0.043)
W=M	0.212	0.183	0.240	-0.057
	(0.022)	(0.029)	(0.032)	(0.043)
W<M	0.576	0.600	0.553	0.047
	(0.026)	(0.037)	(0.037)	(0.052)
Panel C. Second-Order Beliefs, about Women				
W>M	0.178	0.211	0.145	0.066
	(0.020)	(0.031)	(0.026)	(0.041)
W=M	0.212	0.194	0.229	-0.035
	(0.022)	(0.030)	(0.031)	(0.043)
W<M	0.610	0.594	0.626	-0.031
	(0.026)	(0.037)	(0.036)	(0.052)
Observations	354	175	179	354

Notes: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows "W>M", "W=M", and "W<M" report the proportion of participants who believe that the woman chooses a higher MAO, the woman chooses the same MAO, the woman chooses the lower MAO compared to the man in the ultimatum task.

Table 5: Proportion of participants with same-gender second-order beliefs matching their own first-order beliefs, by task

	(1) All	(2) Men	(3) Women
Math	0.605	0.629	0.581
	(0.026)	(0.037)	(0.037)
Ultimatum	0.658	0.640	0.676
	(0.025)	(0.036)	(0.035)
Observations	354	179	175

Notes: Gender-specific proportions of participants whose ternary second-order belief about their own gender is equal to their own ternary first-order belief and standard errors (in parentheses) are in columns (2) and (3).

Table 6: Intra-participant correlations in beliefs that favor men for math task

	(1) Second-Order, about Men, M > W	(2) Second-Order, about Men, M > W	(3) Second-Order, about Women, M > W	(4) Second-Order, about Women, M > W
First-order belief, M > W × Woman	0.177*** (0.063)	0.161** (0.063)	0.310*** (0.072)	0.331*** (0.073)
First-order belief, M > W × Man	0.347*** (0.066)	0.340*** (0.068)	0.168** (0.069)	0.189*** (0.070)
Man	-0.158** (0.071)	-0.180** (0.071)	-0.032 (0.062)	-0.047 (0.062)
STEM major		0.025 (0.051)		0.046 (0.054)
Previous experiments		-0.000 (0.000)		0.000 (0.000)
White		0.050 (0.054)		0.039 (0.054)
English first language		0.081 (0.059)		0.053 (0.060)
Age		0.002 (0.004)		-0.003 (0.004)
Mother has graduate degree		-0.032 (0.124)		-0.180 (0.134)
Mother has bachelor's degree		0.011 (0.119)		-0.246** (0.125)
Mother has HS or associate's		0.195* (0.108)		-0.222* (0.117)
Father has graduate degree		0.073 (0.103)		0.208* (0.106)
Father has bachelor's degree		0.064 (0.099)		0.137 (0.097)
Father has HS or associate's		-0.062 (0.089)		0.146 (0.090)
Online participant		-0.095 (0.068)		0.077 (0.069)
Constant	0.663*** (0.047)	0.516*** (0.156)	0.250*** (0.043)	0.263 (0.175)
Observations	354	354	354	354
p-value: = first-order belief coefficients	0.0635	0.0531	0.153	0.164

Notes: Coefficients and robust standard errors (in parentheses) are from linear regression models with the binary dependent variable indicated in the column heading. "First-order belief, M > W" is a binary variable equal to one if the participant believes the man outperforms the woman in the math task. "Man" and "Woman" are binary variables equal to one if the participant reports being that gender. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "Previous experiments" is equal to the number of experiments the participant completed on Prolific prior to our experiment and is equal to zero for in-person experiments. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 7: Intra-participant correlations in beliefs that favor men for ultimatum task

	(1) Second-Order, about Men, M > W	(2) Second-Order, about Men, M > W	(3) Second-Order, about Women, M > W	(4) Second-Order, about Women, M > W
First-order belief, M > W × Woman	0.217***	0.214***	0.476***	0.423***
	(0.077)	(0.080)	(0.070)	(0.073)
First-order belief, M > W × Man	0.333***	0.303***	0.467***	0.450***
	(0.073)	(0.077)	(0.069)	(0.071)
Man	-0.013	-0.005	-0.003	-0.061
	(0.086)	(0.089)	(0.081)	(0.081)
STEM major		-0.039		0.066
		(0.057)		(0.048)
Previous experiments		0.000		0.000
		(0.000)		(0.000)
White		0.076		0.084
		(0.061)		(0.052)
English first language		0.087		0.067
		(0.068)		(0.057)
Age		-0.002		-0.008**
		(0.004)		(0.003)
Mother has graduate degree		0.087		-0.400***
		(0.131)		(0.111)
Mother has bachelor's degree		0.111		-0.313***
		(0.124)		(0.104)
Mother has HS or associate's		0.100		-0.185*
		(0.116)		(0.097)
Father has graduate degree		-0.099		0.229**
		(0.106)		(0.102)
Father has bachelor's degree		-0.172*		0.179*
		(0.100)		(0.095)
Father has HS or associate's		-0.221**		0.156*
		(0.094)		(0.089)
Online participant		0.023		-0.076
		(0.072)		(0.067)
Constant	0.413***	0.378**	0.317***	0.573***
	(0.062)	(0.172)	(0.059)	(0.152)
Observations	354	354	354	354
p-value: = first-order belief coefficients	0.272	0.419	0.927	0.788

Notes: Coefficients and robust standard errors (in parentheses) are from linear regression models with the binary dependent variable indicated in the column heading. "First-order belief, M > W " is a binary variable equal to one if the participant believes the man chose a higher MAO than the woman in the ultimatum task. "Man" and "Woman" are binary variables equal to one if the participant reports being that gender. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "Previous experiments" is equal to the number of experiments the participant completed on Prolific prior to our experiment and is equal to zero for in-person experiments. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * p<0.10, ** p<0.05, *** p<0.01.

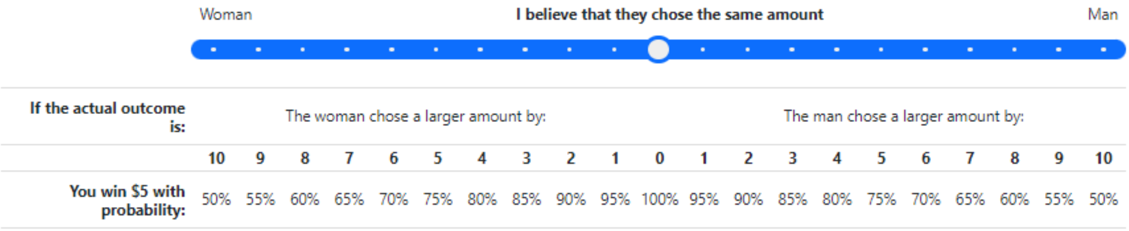
Table 8: Intra-participant difference in second-order beliefs about men and second-order beliefs about women

	(1) Second-Order, Math, M > W	(2) Second-Order, Math, M > W	(3) Second-Order, Ultimatum, M > W	(4) Second-Order, Ultimatum, M > W
Man participant	-0.077	-0.096*	-0.031	-0.063
	(0.050)	(0.051)	(0.052)	(0.051)
Belief about men	0.358***	0.358***	-0.073	-0.073
	(0.045)	(0.046)	(0.048)	(0.048)
Belief about men × Man participant	0.020	0.020	0.078	0.078
	(0.065)	(0.066)	(0.066)	(0.066)
STEM major		0.037		0.005
		(0.041)		(0.042)
Previous experiments		0.000		0.000
		(0.000)		(0.000)
White		0.059		0.143***
		(0.041)		(0.043)
English first language		0.063		0.101**
		(0.046)		(0.049)
Age		-0.001		-0.007**
		(0.003)		(0.003)
Mother has graduate degree		-0.100		-0.171*
		(0.089)		(0.091)
Mother has bachelor's degree		-0.062		-0.108
		(0.082)		(0.086)
Mother has HS or associate's		0.001		-0.025
		(0.076)		(0.084)
Father has graduate degree		0.118		0.091
		(0.075)		(0.077)
Father has bachelor's degree		0.072		0.055
		(0.072)		(0.077)
Father has HS or associate's		0.040		-0.050
		(0.064)		(0.071)
Online participant		-0.072		-0.071
		(0.051)		(0.056)
Constant	0.380***	0.321***	0.626***	0.733***
	(0.036)	(0.123)	(0.036)	(0.128)
Observations	708	708	708	708

Notes: Coefficients and standard errors (in parentheses), clustered at the participant level, are from linear regression models with the binary dependent variable indicated in the column heading. The unit of observation is the participant-by-second-order belief pair, so that each participant contributes two observations: one second-order belief about men and one second-order belief about women. "Man participant" is a binary variable equal to one if the participant reports being a man. "Belief about men" is a binary variable equal to one if the relevant second-order belief is about men. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "Previous experiments" is equal to the number of experiments the participant completed on Prolific prior to our experiment and is equal to zero for in-person experiments. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * p<0.10, ** p<0.05, *** p<0.01.

Figures

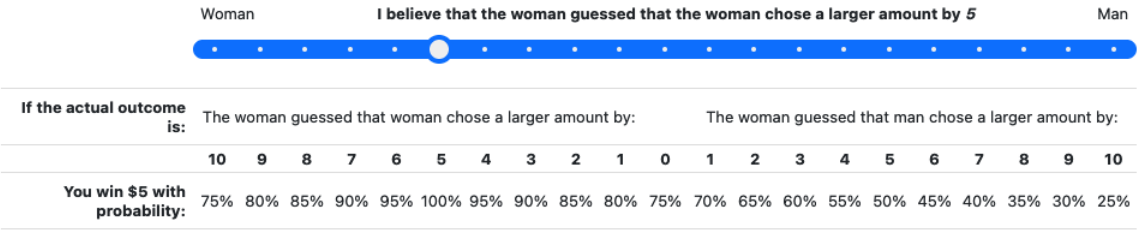
Figure 1: Example of slider interface used for first-order belief elicitation



I believe that they chose the same amount

If the actual outcome is:	The woman chose a larger amount by:											The man chose a larger amount by:									
	10	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%

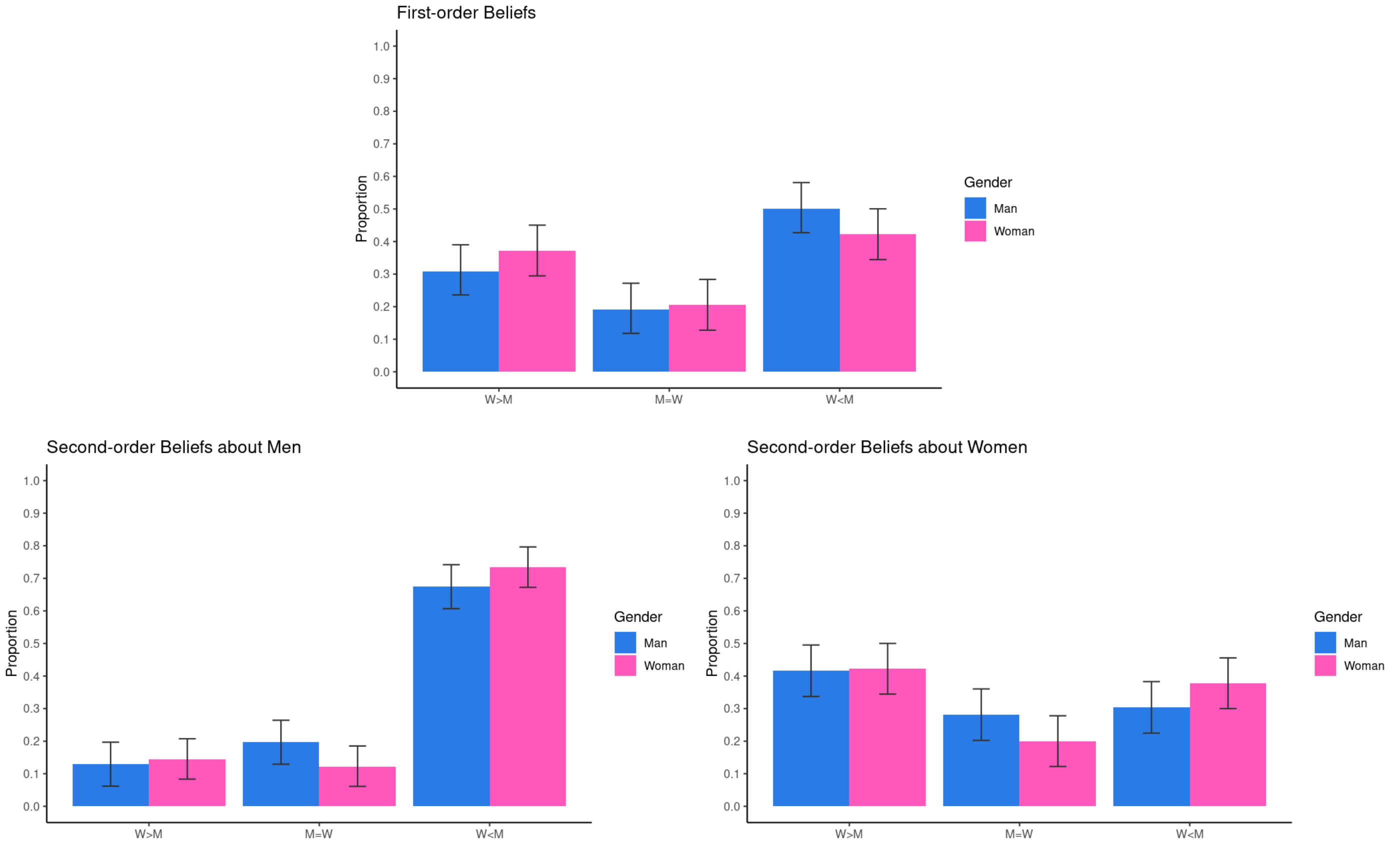
Figure 2: Example of slider interface used for second-order belief elicitation



I believe that the woman guessed that the woman chose a larger amount by *5*

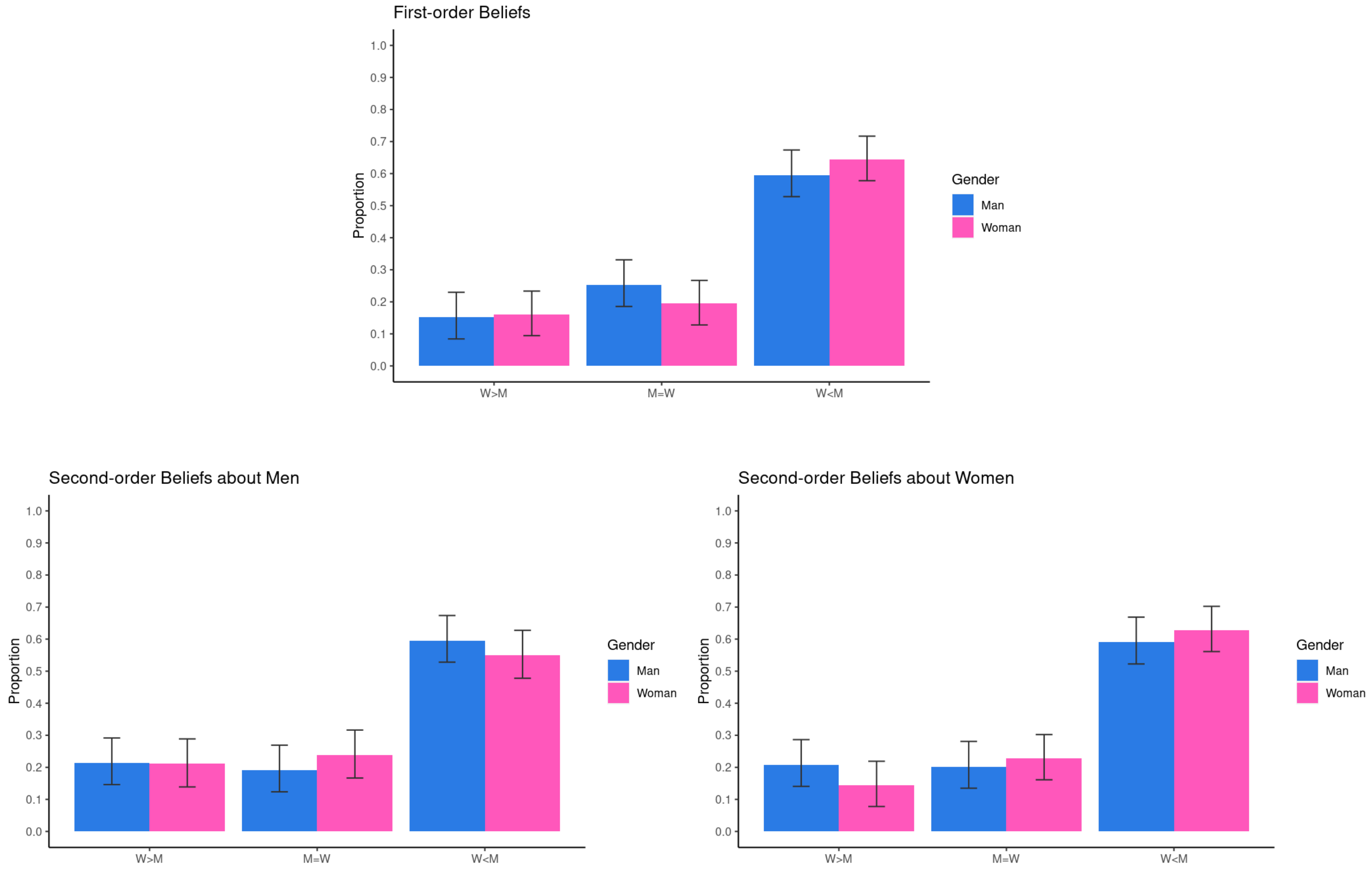
If the actual outcome is:	The woman guessed that woman chose a larger amount by:											The woman guessed that man chose a larger amount by:									
	10	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
You win $5 with probability:	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%	45%	40%	35%	30%	25%

Figure 3: Belief elicitations about the math task



Notes: From left to right on each graph, the bars represent the proportion of participants who believe (or believe that others believe) that the woman scores higher, the woman scores the same, the woman scores lower compared to the man in the math task.

Figure 4: Belief elicitations about the ultimatum task



Notes: From left to right on each graph, the bars represent the proportion of participants who believe (or believe that others believe) that the woman chooses higher MAO, the woman chooses the same, the woman chooses lower MAO compared to the man in the ultimatum task.

Appendix A: Alternative Approaches

We choose to elicit medians because they offer precise information about the property we are interested in—whether $P(X_1 > X_2) \geq \frac{1}{2}$ or $P(X_1 < X_2) \geq \frac{1}{2}$. We next consider alternative functions of the participant's subjective belief distributions that could also elicit this information and discuss why we did not choose them for this experiment. Practitioners with different properties of interest may find these alternative functions more appropriate.

Eliciting Probabilities

One alternative approach would be to directly elicit the probabilities of interest: $P(X_1 > X_2)$ and $P(X_2 > X_1)$. These probabilities are means of binary distributions equal to 1 when the event occurs and equal to 0 otherwise, where the events are $x_1 > x_2$ or $x_2 > x_1$. The BSR can elicit a mean as well as a median by using the appropriate loss function, ensuring that we could robustly elicit these probabilities. In fact, eliciting probabilities provides cardinal information about the participants' beliefs that is unobserved in our procedure. The cost of this additional information is an additional task for each belief elicited.

We choose not to take this approach because it requires two belief elicitations for each comparison of interest to determine which event is more likely. To determine whether $P(X_1 > X_2) \geq \frac{1}{2}$ or $P(X_1 < X_2) \geq \frac{1}{2}$ using the elicitation of probabilities would require that we elicit both $P(X_1 > X_2)$ and $P(X_1 < X_2)$. Since the outcome $x_1 = x_2$ is possible, the complement of $P(X_1 > X_2)$ is $P(X_1 \leq X_2)$, *not* $P(X_1 < X_2)$. While the cardinal information may be interesting, we argue that the precise probabilities of each event are not important enough in this experiment to justify the additional cognitive and time costs to participants from doubling the number of elicitations. Furthermore, since we use a random task payment procedure, doubling the number of tasks would also dilute the incentives.

Eliciting Modes of a Ternary Distribution

Another approach to determining which of a set of mutually independent outcomes is most likely is simply to ask participants which event they would like to condition their payment on. That is, ask participants to choose which outcome they think is most likely: $x_1 > x_2$, $x_1 < x_2$, or $x_1 = x_2$. This procedure is proper for eliciting the mode of a ternary distribution.

While the incentives of this procedure are clear and simple, participants with symmetric beliefs may nonetheless be incentivized to choose $x_1 > x_2$ or $x_1 < x_2$ instead of $x_1 = x_2$. Consider a continuous distribution that is identical for X_1 and X_2. Even though $X_1 = X_2$, it is sub-optimal to bet on the outcome $x_1 = x_2$ since $P(x_1 = x_2) = 0$. This also applies when X_1 and X_2 are discrete, but the probability of equality is sufficiently low.

Under this payment structure, participants in our experiment who believe that men and women perform equally well on the math task would be incentivized to choose one of the non-gender-neutral outcomes simply because there are many more ways for two people to have a different math score than there are for two people to have the same math score. Therefore, we would not be able to distinguish gender-neutral participants.

In contrast, using the median procedure, a participant with symmetric beliefs is incentivized to select zero as their median belief regardless of their belief about the probability that the two randomly chosen subjects score identically. Participants with symmetric beliefs and participants whose beliefs are substantially asymmetric can always be differentiated.

Eliciting Population Medians

We elicit the median of a distribution of differences. An alternative approach would be to elicit the medians of each distribution separately and take the difference. In other words, there are two possibly relevant quantities involving medians: the median of the differences and the difference in the medians.

Eliciting the medians of X_1 and X_2 does not provide us the relevant information to assess our property of interest: whether $P(X_1 > X_2) \geq \frac{1}{2}$ or $P(X_1 < X_2) \geq \frac{1}{2}$. Specifically, $Median(X_1) > Median(X_2)$ does not imply that $Median(X_1 - X_2) > 0$. Consider the data in Table A1: $Median(X_1) > Median(X_2)$ since $Median(X_1) = 3$ and $Median(X_2) = 2$; however, $Median(X_1 - X_2) = -1$ implying that $P(X_2 > X_1) > \frac{1}{2}$.

Table A1: Example distributions illustrating that $Median(X_1) > Median(X_2)$ does not imply $P(X_1 > X_2) \geq \frac{1}{2}$

	Value		
X_1	0	3	4
X_2	1	2	5

Appendix B: Supplemental Figures and Tables

Table B1: Participant characteristics, in-person vs. online

	(1) In-person	(2) Online	(3) Difference
Man	0.51	0.48	0.03
	(0.50)	(0.50)	(0.05)
STEM major	0.24	0.43	-0.18***
	(0.43)	(0.50)	(0.05)
White	0.47	0.71	-0.24***
	(0.50)	(0.45)	(0.05)
English first language	0.80	0.75	0.05
	(0.40)	(0.43)	(0.04)
Age	20.66	26.26	-5.61***
	(3.39)	(8.39)	(0.66)
Mother has graduate degree	0.38	0.15	0.22***
	(0.49)	(0.36)	(0.05)
Mother has bachelor's degree	0.44	0.32	0.12**
	(0.50)	(0.47)	(0.05)
Mother has HS or associate's	0.17	0.40	-0.22***
	(0.38)	(0.49)	(0.05)
Father has graduate degree	0.52	0.14	0.38***
	(0.50)	(0.35)	(0.05)
Father has bachelor's degree	0.34	0.25	0.09*
	(0.47)	(0.43)	(0.05)
Father has HS or associate's	0.12	0.42	-0.30***
	(0.33)	(0.49)	(0.04)
Observations	157	197	354

Notes: Means and standard deviations (in parentheses) are in columns (1) and (2). Column (3) gives the difference in means and its standard error (in parentheses). "Man" is a binary variable equal to one if the participant reports being a man. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table B2: Gender differences in beliefs that favor men for math task

	(1)	(2)	(3)	(4)	(5)	(6)
	First-Order, M > W	First-Order, M > W	Second-Order, about Men, M > W	Second-Order, about Men, M > W	Second-Order, about Women, M > W	Second-Order, about Women, M > W
Man	0.084	0.074	-0.057	-0.080	-0.077	-0.092*
	(0.053)	(0.054)	(0.048)	(0.050)	(0.050)	(0.052)
STEM major		0.009		0.021		0.054
		(0.057)		(0.053)		(0.056)
Previous experiments		0.000		0.000		0.000
		(0.000)		(0.000)		(0.000)
White		0.055		0.063		0.054
		(0.058)		(0.055)		(0.056)
English first language		-0.018		0.077		0.049
		(0.066)		(0.062)		(0.060)
Age		0.001		0.002		-0.003
		(0.005)		(0.004)		(0.004)
Mother has graduate degree		0.025		-0.028		-0.171
		(0.129)		(0.127)		(0.129)
Mother has bachelor's degree		0.212*		0.076		-0.200*
		(0.123)		(0.122)		(0.121)
Mother has HS or associate's		0.061		0.209*		-0.206*
		(0.119)		(0.113)		(0.113)
Father has graduate degree		-0.097		0.073		0.163
		(0.113)		(0.102)		(0.110)
Father has bachelor's degree		-0.118		0.050		0.093
		(0.109)		(0.100)		(0.103)
Father has HS or associate's		-0.011		-0.056		0.136
		(0.103)		(0.091)		(0.094)
Online participant		-0.248***		-0.151**		0.007
		(0.073)		(0.071)		(0.072)
Constant	0.419***	0.454***	0.737***	0.575***	0.380***	0.425**
	(0.037)	(0.168)	(0.033)	(0.157)	(0.036)	(0.172)
Observations	354	354	354	354	354	354

Notes: Coefficients and robust standard errors (in parentheses) are from linear regression models with binary dependent variable indicated in the column heading. “Man” is a binary variable equal to one if the participant reports being a man. “STEM major” is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. “Previous experiments” is a variable we have for online participants only and is equal to the number of experiments the participant completed on Prolific prior to our experiment. “White” is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. “English first language” is a binary variable equal to one if the participant reports that the first language they learned was English. “Age” is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. “Online participant” is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * p<0.10, ** p<0.05, *** p<0.01.

Table B3: Gender differences in ternary beliefs for math task

	(1) First-Order	(2) First-Order	(3) Second-Order, about Men	(4) Second-Order, about Men	(5) Second-Order, about Women	(6) Second-Order, about Women
Man	0.308	0.322	-0.223	-0.366	-0.162	-0.207
	(0.200)	(0.213)	(0.232)	(0.250)	(0.198)	(0.206)
STEM major		-0.095		-0.007		0.166
		(0.223)		(0.266)		(0.218)
Previous experiments		0.001		0.000		0.000
		(0.001)		(0.001)		(0.001)
White		0.214		0.326		0.138
		(0.230)		(0.276)		(0.224)
English first language		-0.017		0.456		0.334
		(0.258)		(0.287)		(0.253)
Age		-0.005		0.011		-0.002
		(0.016)		(0.019)		(0.017)
Mother has graduate degree		0.166		-0.122		-0.334
		(0.516)		(0.594)		(0.612)
Mother has bachelor's degree		0.842*		0.532		-0.379
		(0.504)		(0.555)		(0.581)
Mother has HS or associate's		0.273		1.170**		-0.641
		(0.489)		(0.523)		(0.567)
Father has graduate degree		-0.099		0.387		0.732
		(0.456)		(0.539)		(0.496)
Father has bachelor's degree		-0.243		0.246		0.377
		(0.463)		(0.518)		(0.465)
Father has HS or associate's		0.199		-0.204		0.746*
		(0.434)		(0.432)		(0.433)
Online participant		-0.750***		-0.864**		-0.129
		(0.284)		(0.371)		(0.300)
Observations	354	354	354	354	354	354

Notes: Coefficients and robust standard errors (in parentheses) are from proportional odds (ordered logit) models with ternary dependent variable indicated in the column heading. "Man" is a binary variable equal to one if the participant reports being a man. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "Previous experiments" is a variable we have for online participants only and is equal to the number of experiments the participant completed on Prolific prior to our experiment. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table B4: Gender differences in beliefs that favor men for ultimatum task

	(1)	(2)	(3)	(4)	(5)	(6)
	First-Order, M > W	First-Order, M > W	Second-Order, about Men, M > W	Second-Order, about Men, M > W	Second-Order, about Women, M > W	Second-Order, about Women, M > W
Man	-0.048	-0.077	0.047	0.031	-0.031	-0.078
	(0.052)	(0.051)	(0.053)	(0.054)	(0.052)	(0.052)
STEM major		-0.014		-0.048		0.058
		(0.056)		(0.057)		(0.054)
Previous experiments		0.000		0.000		0.000
		(0.000)		(0.000)		(0.000)
White		0.176***		0.125**		0.161***
		(0.057)		(0.059)		(0.055)
English first language		0.075		0.103		0.099
		(0.069)		(0.068)		(0.063)
Age		-0.007		-0.004		-0.011***
		(0.004)		(0.004)		(0.004)
Mother has graduate degree		-0.058		0.081		-0.423***
		(0.137)		(0.128)		(0.114)
Mother has bachelor's degree		-0.030		0.109		-0.324***
		(0.128)		(0.119)		(0.108)
Mother has HS or associate's		0.045		0.114		-0.165
		(0.123)		(0.114)		(0.104)
Father has graduate degree		0.071		-0.079		0.261**
		(0.113)		(0.106)		(0.102)
Father has bachelor's degree		0.148		-0.133		0.244**
		(0.107)		(0.102)		(0.100)
Father has HS or associate's		-0.050		-0.234**		0.134
		(0.103)		(0.094)		(0.095)
Online participant		-0.129*		-0.010		-0.132*
		(0.074)		(0.074)		(0.072)
Constant	0.648***	0.696***	0.553***	0.526***	0.626***	0.868***
	(0.036)	(0.173)	(0.037)	(0.164)	(0.036)	(0.157)
Observations	354	354	354	354	354	354

Notes: Coefficients and robust standard errors (in parentheses) are from linear regression models with binary dependent variable indicated in the column heading. "Man" is a binary variable equal to one if the participant reports being a man. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "Previous experiments" is a variable we have for online participants only and is equal to the number of experiments the participant completed on Prolific prior to our experiment. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table B5: Gender differences in ternary beliefs for ultimatum task

	(1)	(2)	(3)	(4)	(5)	(6)
	First-Order	First-Order	Second-Order, about Men	Second-Order, about Men	Second-Order, about Women	Second-Order, about Women
Man	-0.161	-0.301	0.124	0.061	-0.205	-0.444*
	(0.215)	(0.223)	(0.208)	(0.220)	(0.212)	(0.232)
STEM major		-0.049		-0.210		0.297
		(0.240)		(0.230)		(0.238)
Previous experiments		0.000		0.001		0.000
		(0.001)		(0.001)		(0.001)
White		0.790***		0.479**		0.697***
		(0.254)		(0.237)		(0.239)
English first language		0.364		0.365		0.422*
		(0.287)		(0.257)		(0.254)
Age		-0.044**		-0.019		-0.051***
		(0.019)		(0.019)		(0.018)
Mother has graduate degree		-0.328		0.595		-1.839***
		(0.610)		(0.589)		(0.588)
Mother has bachelor's degree		-0.201		0.644		-1.498***
		(0.556)		(0.566)		(0.560)
Mother has HS or associate's		0.191		0.587		-0.750
		(0.533)		(0.546)		(0.535)
Father has graduate degree		0.526		-0.407		1.447***
		(0.536)		(0.516)		(0.495)
Father has bachelor's degree		0.809		-0.631		1.216**
		(0.520)		(0.505)		(0.483)
Father has HS or associate's		-0.088		-1.049**		0.837*
		(0.472)		(0.477)		(0.445)
Online participant		-0.634*		-0.197		-0.676**
		(0.325)		(0.295)		(0.319)
Observations	354	354	354	354	354	354

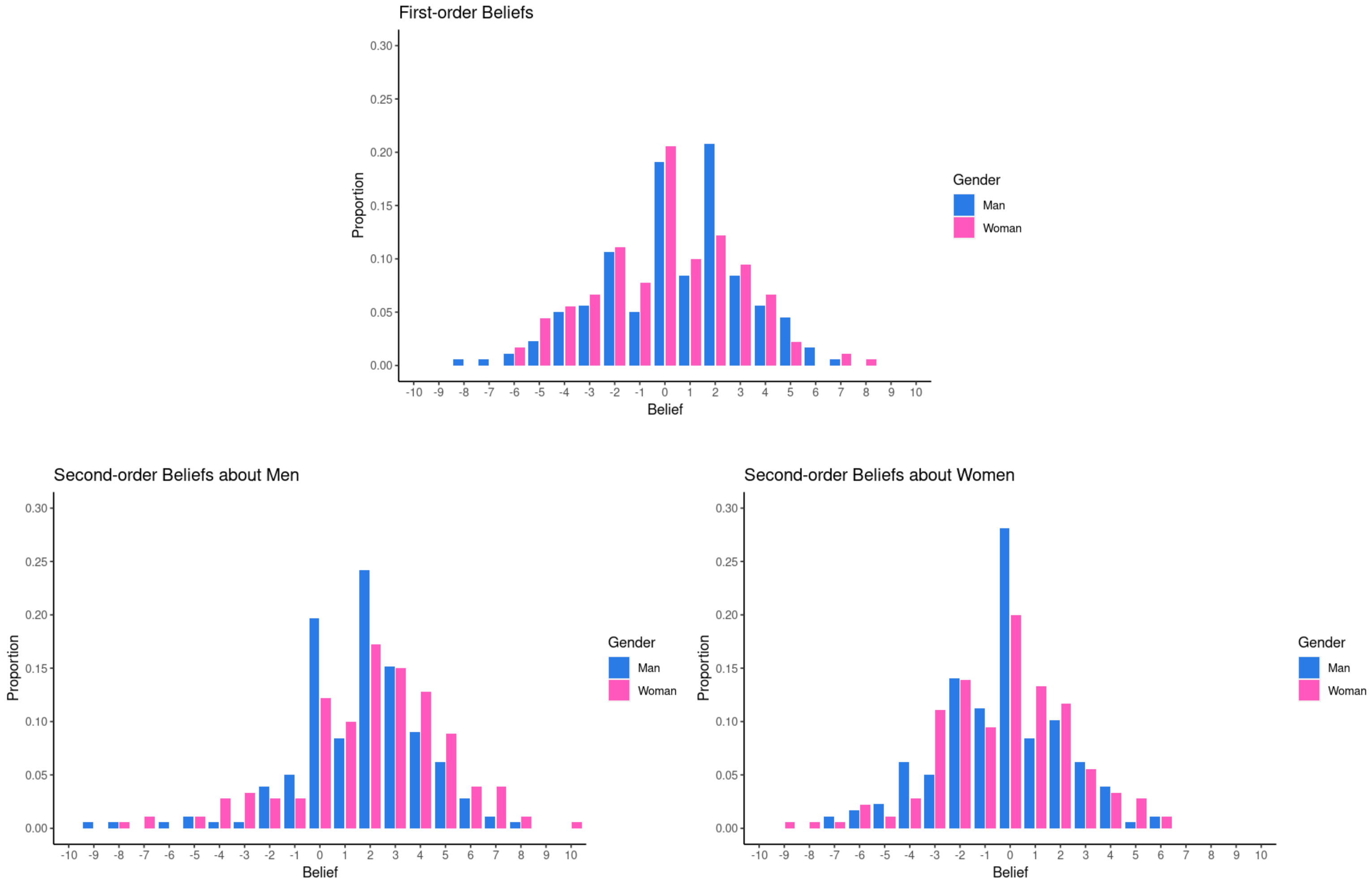
Notes: Coefficients and robust standard errors (in parentheses) are from proportional odds (ordered logit) models with ternary dependent variable indicated in the column heading. "Man" is a binary variable equal to one if the participant reports being a man. "STEM major" is a binary variable equal to one if the participant reporting majoring in a science, technology, engineer, or mathematics field. "Previous experiments" is a variable we have for online participants only and is equal to the number of experiments the participant completed on Prolific prior to our experiment. "White" is a binary variable equal to one if the participant reports identifying as ethnically white or Caucasian. "English first language" is a binary variable equal to one if the participant reports that the first language they learned was English. "Age" is an integer. Mother's (father's) education levels are each a binary variable equal to one if the participant reports their mother (father) has its respective education level. "Online participant" is a binary variable equal to one if the participant is in the online (as opposed to in-person) experiment. * p<0.10, ** p<0.05, *** p<0.01.

Table B6: Proportion of participants with same-gender second-order beliefs matching their own first-order beliefs, by first-order belief

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. Math Task				
W>M	0.525	0.327	0.687	-0.359
	(0.045)	(0.063)	(0.057)	(0.085)
W=M	0.478	0.531	0.432	0.099
	(0.060)	(0.088)	(0.081)	(0.120)
W<M	0.718	0.852	0.560	0.292
	(0.035)	(0.038)	(0.057)	(0.069)
Panel B. Ultimatum Task				
W>M	0.400	0.444	0.357	0.087
	(0.066)	(0.096)	(0.091)	(0.132)
W=M	0.538	0.535	0.543	-0.008
	(0.056)	(0.076)	(0.084)	(0.113)
W<M	0.765	0.733	0.793	-0.060
	(0.029)	(0.043)	(0.038)	(0.057)
Observations	354	175	179	354

Note: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows "W>M", "W=M", and "W<M" report the proportion of participants in each task who believe that the woman outperforms/chooses higher MAO, the woman performs/chooses the same, the woman performs lower/chooses lower MAO compared to the man and whose ternary second-order belief about their own gender is equal to their own ternary first-order belief.

Figure B1: Belief elicitation about the math task

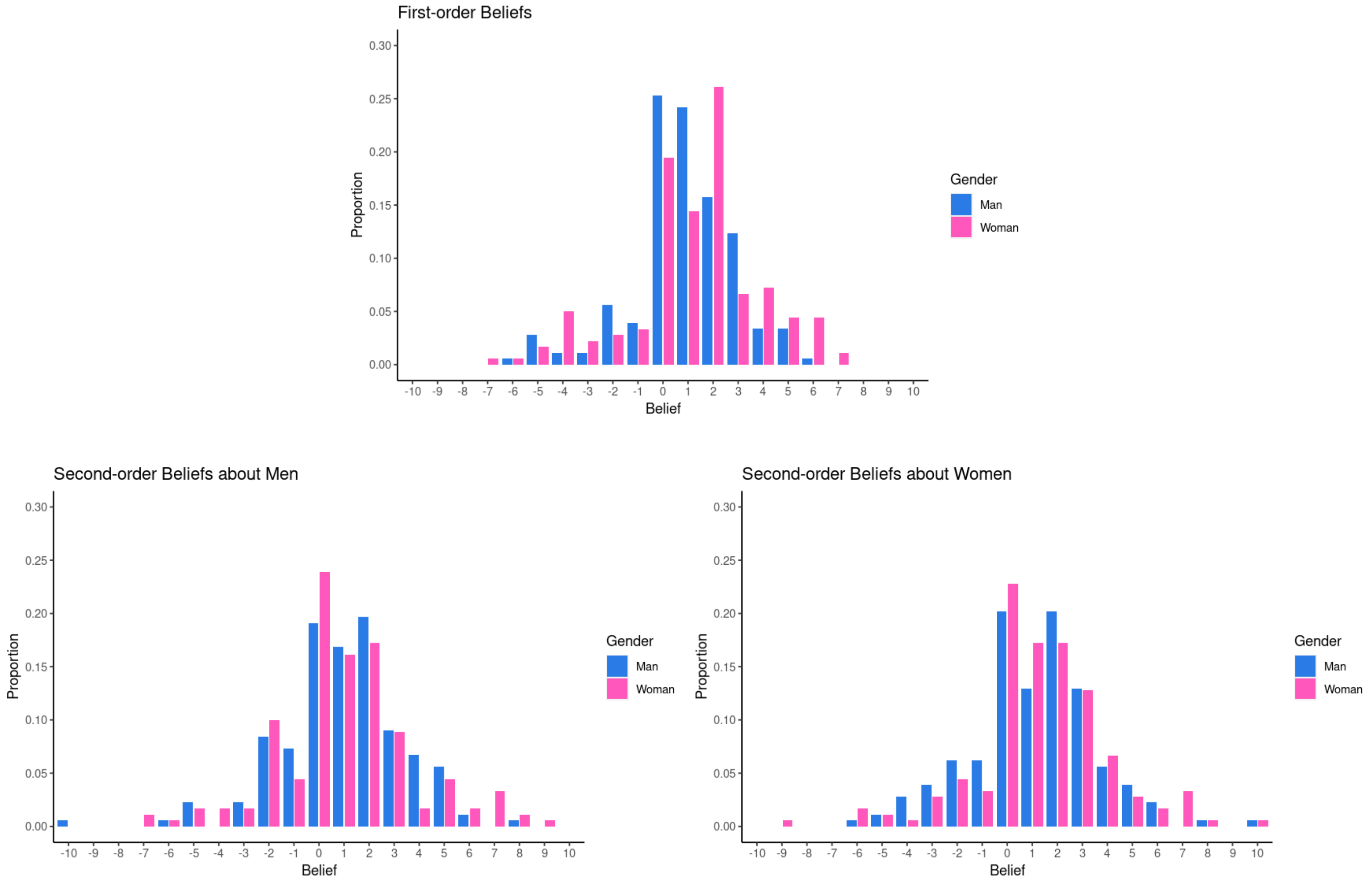

First-order Beliefs

Second-order Beliefs about Men

Second-order Beliefs about Women

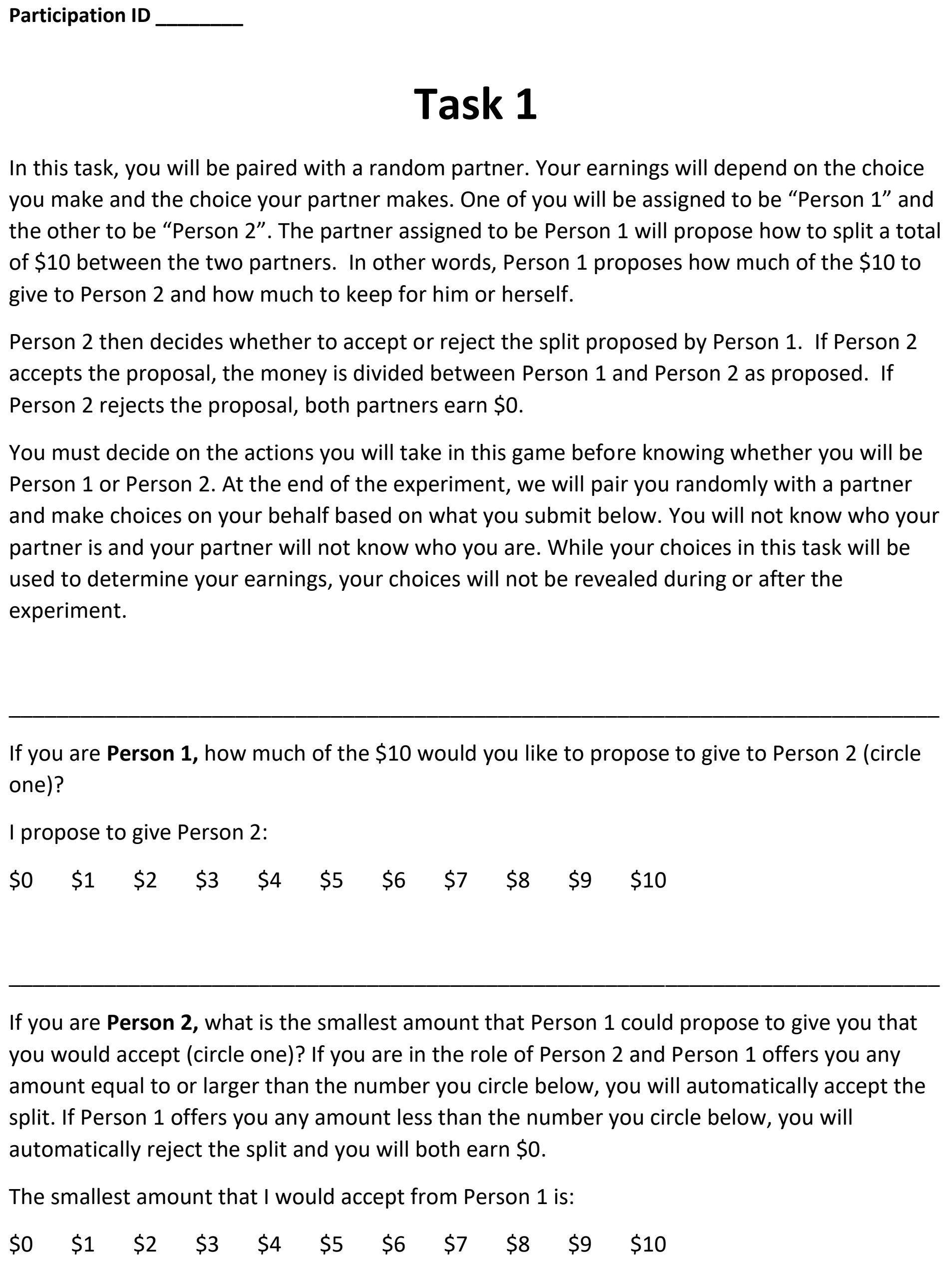
Notes: Each bar represents the proportion of participants who believe (or believe that others believe) that the median difference between a man and a woman on the math task is its respective point. A negative difference means that the woman answered more math summations correctly, while a positive difference means that the man answered more correctly.

Figure B2: Belief elicitations about the ultimatum task



Notes: Each bar represents the proportion of participants who believe (or believe that others believe) that the median difference between a man and a woman on the ultimatum task is its respective point. A negative difference means that the woman chose a higher MAO, while a positive difference means that the man chose a higher MAO.

Online Appendix A: Task Instructions



Participation ID ________

Task 1

In this task, you will be paired with a random partner. Your earnings will depend on the choice you make and the choice your partner makes. One of you will be assigned to be "Person 1" and the other to be "Person 2". The partner assigned to be Person 1 will propose how to split a total of $10 between the two partners. In other words, Person 1 proposes how much of the $10 to give to Person 2 and how much to keep for him or herself.

Person 2 then decides whether to accept or reject the split proposed by Person 1. If Person 2 accepts the proposal, the money is divided between Person 1 and Person 2 as proposed. If Person 2 rejects the proposal, both partners earn $0.

You must decide on the actions you will take in this game before knowing whether you will be Person 1 or Person 2. At the end of the experiment, we will pair you randomly with a partner and make choices on your behalf based on what you submit below. You will not know who your partner is and your partner will not know who you are. While your choices in this task will be used to determine your earnings, your choices will not be revealed during or after the experiment.

If you are **Person 1,** how much of the $10 would you like to propose to give to Person 2 (circle one)?

I propose to give Person 2:

$0 $1 $2 $3 $4 $5 $6 $7 $8 $9 $10

If you are **Person 2,** what is the smallest amount that Person 1 could propose to give you that you would accept (circle one)? If you are in the role of Person 2 and Person 1 offers you any amount equal to or larger than the number you circle below, you will automatically accept the split. If Person 1 offers you any amount less than the number you circle below, you will automatically reject the split and you will both earn $0.

The smallest amount that I would accept from Person 1 is:

$0 $1 $2 $3 $4 $5 $6 $7 $8 $9 $10

Participation ID ________

Task 2

During this task you earn money by correctly summing 2-digit numbers. You will be shown several sets of five two-digit numbers. Each set will be arranged in a row. For example, you could see:

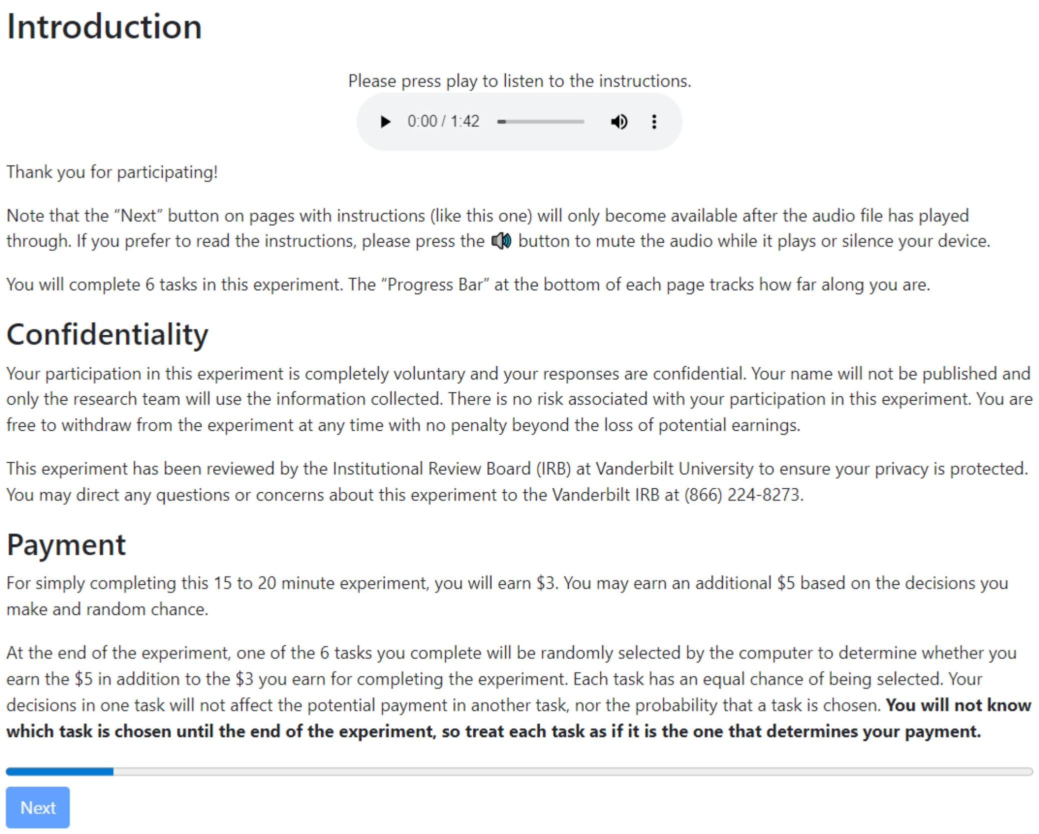
60	71	41	75	81	

For each set, you will write your answer in the empty box on the right. In the above example, the correct answer is 60 + 71 + 41 + 75 + 81 = 328. You would write 328 in the empty box.

For each correct answer, you will earn $0.50. You will not be penalized for incorrect answers. You have 5 minutes to solve as many of the summations as you can. You will be told when time is up, but no time warnings will be issued.

When the experimenter instructs you to do so, please turn to the next page and begin.

Online Appendix B: Experiment Screenshots



Introduction

Please press play to listen to the instructions.

Thank you for participating!

Note that the "Next" button on pages with instructions (like this one) will only become available after the audio file has played through. If you prefer to read the instructions, please press the 🔇 button to mute the audio while it plays or silence your device.

You will complete 6 tasks in this experiment. The "Progress Bar" at the bottom of each page tracks how far along you are.

Confidentiality

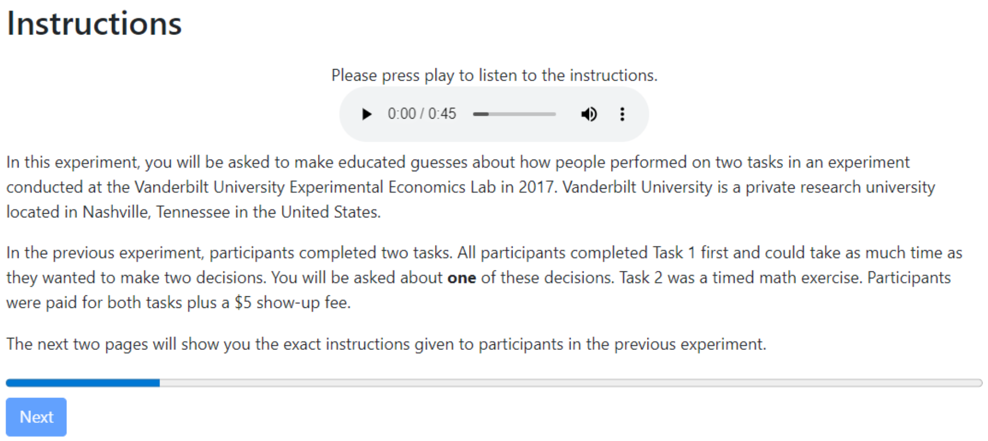
Your participation in this experiment is completely voluntary and your responses are confidential. Your name will not be published and only the research team will use the information collected. There is no risk associated with your participation in this experiment. You are free to withdraw from the experiment at any time with no penalty beyond the loss of potential earnings.

This experiment has been reviewed by the Institutional Review Board (IRB) at Vanderbilt University to ensure your privacy is protected. You may direct any questions or concerns about this experiment to the Vanderbilt IRB at (866) 224-8273.

Payment

For simply completing this 15 to 20 minute experiment, you will earn $3. You may earn an additional $5 based on the decisions you make and random chance.

At the end of the experiment, one of the 6 tasks you complete will be randomly selected by the computer to determine whether you earn the $5 in addition to the $3 you earn for completing the experiment. Each task has an equal chance of being selected. Your decisions in one task will not affect the potential payment in another task, nor the probability that a task is chosen. **You will not know which task is chosen until the end of the experiment, so treat each task as if it is the one that determines your payment.**



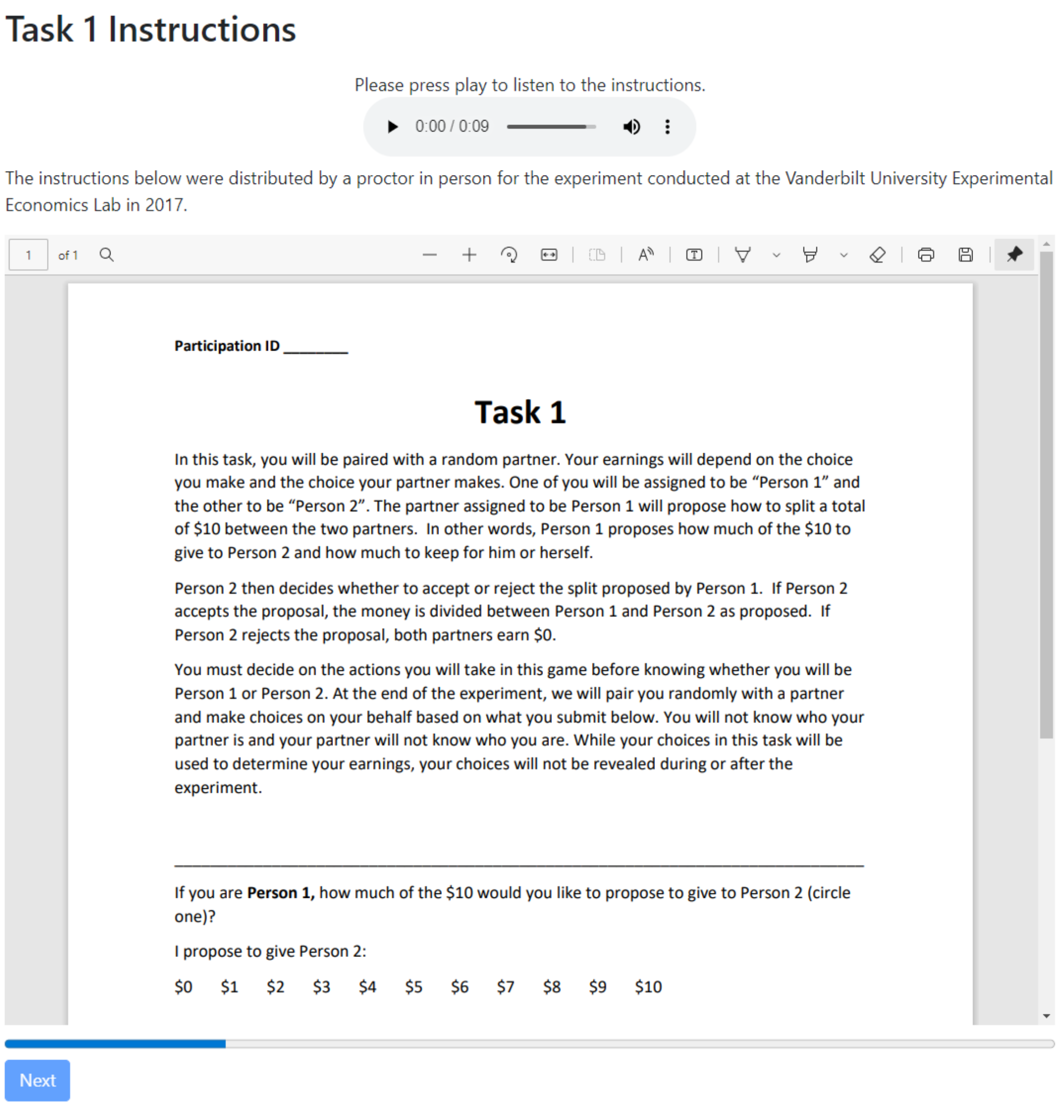
Instructions

Please press play to listen to the instructions.

In this experiment, you will be asked to make educated guesses about how people performed on two tasks in an experiment conducted at the Vanderbilt University Experimental Economics Lab in 2017. Vanderbilt University is a private research university located in Nashville, Tennessee in the United States.

In the previous experiment, participants completed two tasks. All participants completed Task 1 first and could take as much time as they wanted to make two decisions. You will be asked about **one** of these decisions. Task 2 was a timed math exercise. Participants were paid for both tasks plus a $5 show-up fee.

The next two pages will show you the exact instructions given to participants in the previous experiment.



Task 1 Instructions

Please press play to listen to the instructions.

The instructions below were distributed by a proctor in person for the experiment conducted at the Vanderbilt University Experimental Economics Lab in 2017.

Participation ID ________

Task 1

In this task, you will be paired with a random partner. Your earnings will depend on the choice you make and the choice your partner makes. One of you will be assigned to be "Person 1" and the other to be "Person 2". The partner assigned to be Person 1 will propose how to split a total of $10 between the two partners. In other words, Person 1 proposes how much of the $10 to give to Person 2 and how much to keep for him or herself.

Person 2 then decides whether to accept or reject the split proposed by Person 1. If Person 2 accepts the proposal, the money is divided between Person 1 and Person 2 as proposed. If Person 2 rejects the proposal, both partners earn $0.

You must decide on the actions you will take in this game before knowing whether you will be Person 1 or Person 2. At the end of the experiment, we will pair you randomly with a partner and make choices on your behalf based on what you submit below. You will not know who your partner is and your partner will not know who you are. While your choices in this task will be used to determine your earnings, your choices will not be revealed during or after the experiment.

If you are **Person 1,** how much of the $10 would you like to propose to give to Person 2 (circle one)?

I propose to give Person 2:

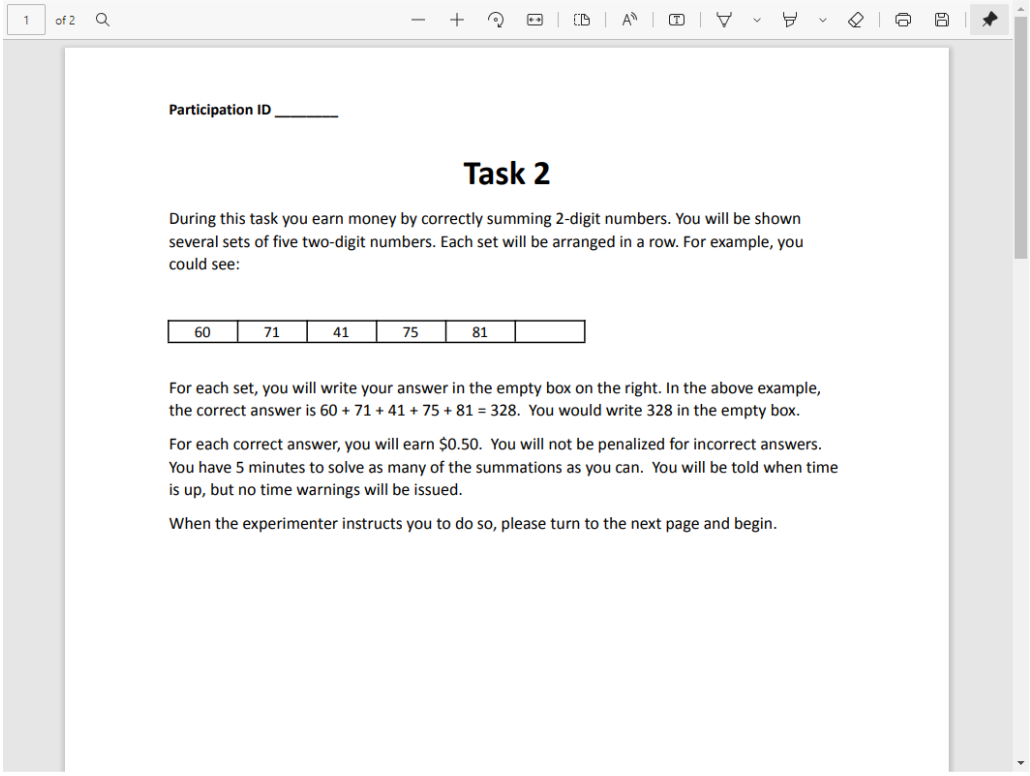
$0 $1 $2 $3 $4 $5 $6 $7 $8 $9 $10

Task 2 Instructions

Please press play to listen to the instructions.

▶ 0:00 / 0:09

The instructions below were distributed by a proctor in person for the experiment conducted at the Vanderbilt University Experimental Economics Lab in 2017.



Participation ID ________

Task 2

During this task you earn money by correctly summing 2-digit numbers. You will be shown several sets of five two-digit numbers. Each set will be arranged in a row. For example, you could see:

60	71	41	75	81	

For each set, you will write your answer in the empty box on the right. In the above example, the correct answer is 60 + 71 + 41 + 75 + 81 = 328. You would write 328 in the empty box.

For each correct answer, you will earn $0.50. You will not be penalized for incorrect answers. You have 5 minutes to solve as many of the summations as you can. You will be told when time is up, but no time warnings will be issued.

When the experimenter instructs you to do so, please turn to the next page and begin.

Next

Payment

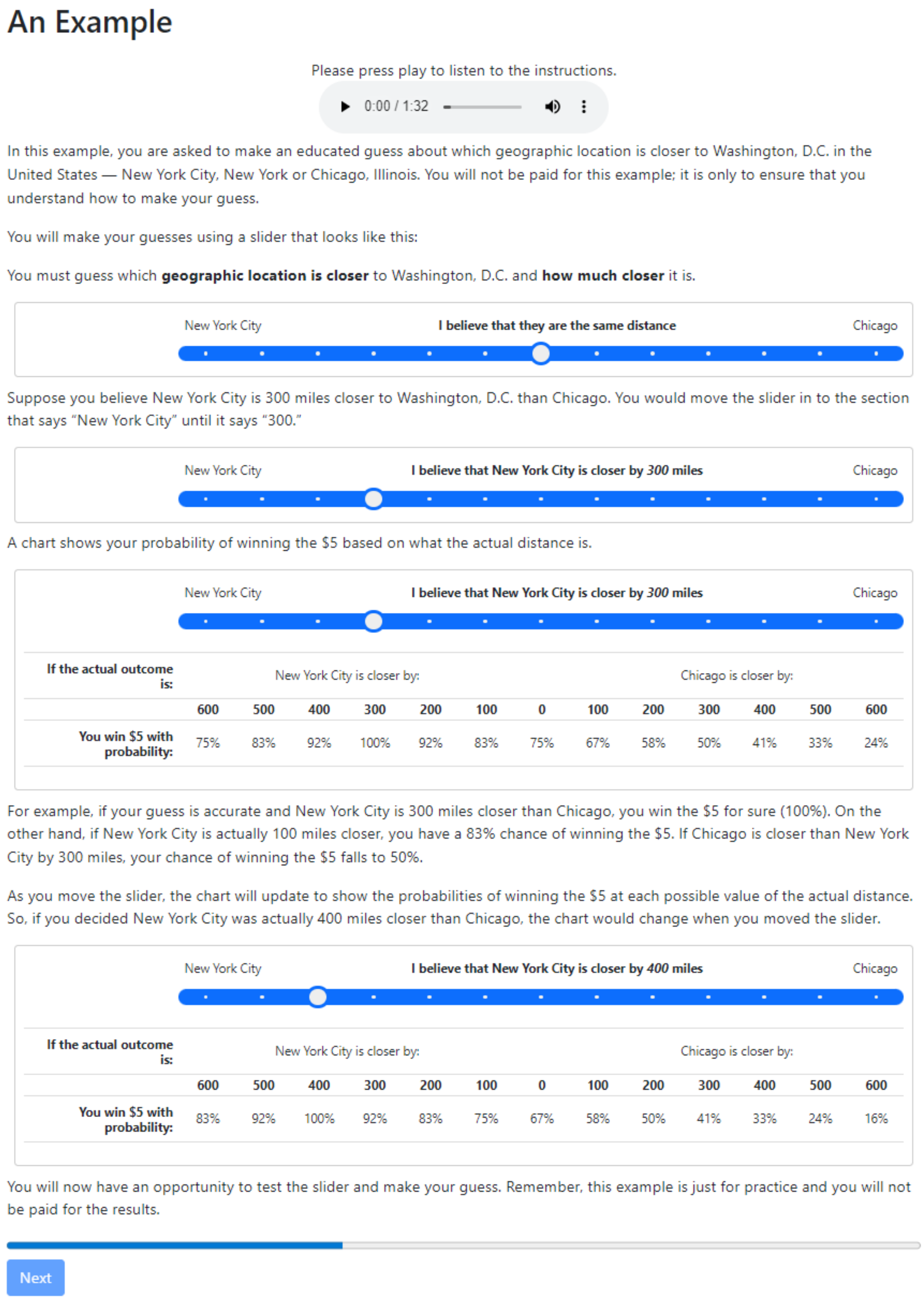
Please press play to listen to the instructions.

▶ 0:00 / 0:31

In this experiment, your payment will be based on a lottery in which you receive either $5 or $0. The likelihood that you receive the larger amount of $5 is determined by how accurate your educated guess is compared to the actual outcome. (If you are interested, the lottery system is carefully designed so that it is mathematically optimal to submit your best guess about the median outcome.) So, **it is in your best interest to submit your true best guess**.

Next, we will go through an example so that you understand the lottery and how to make your guess.

Next



An Example

Please press play to listen to the instructions.

▶ 0:00 / 1:32

In this example, you are asked to make an educated guess about which geographic location is closer to Washington, D.C. in the United States — New York City, New York or Chicago, Illinois. You will not be paid for this example; it is only to ensure that you understand how to make your guess.

You will make your guesses using a slider that looks like this:

You must guess which **geographic location is closer** to Washington, D.C. and **how much closer** it is.

New York City **I believe that they are the same distance** Chicago

Suppose you believe New York City is 300 miles closer to Washington, D.C. than Chicago. You would move the slider in to the section that says "New York City" until it says "300."

New York City **I believe that New York City is closer by *300* miles** Chicago

A chart shows your probability of winning the $5 based on what the actual distance is.

New York City **I believe that New York City is closer by *300* miles** Chicago

If the actual outcome is:			New York City is closer by:							Chicago is closer by:			
	600	**500**	**400**	**300**	**200**	**100**	**0**	**100**	**200**	**300**	**400**	**500**	**600**
You win $5 with probability:	75%	83%	92%	100%	92%	83%	75%	67%	58%	50%	41%	33%	24%

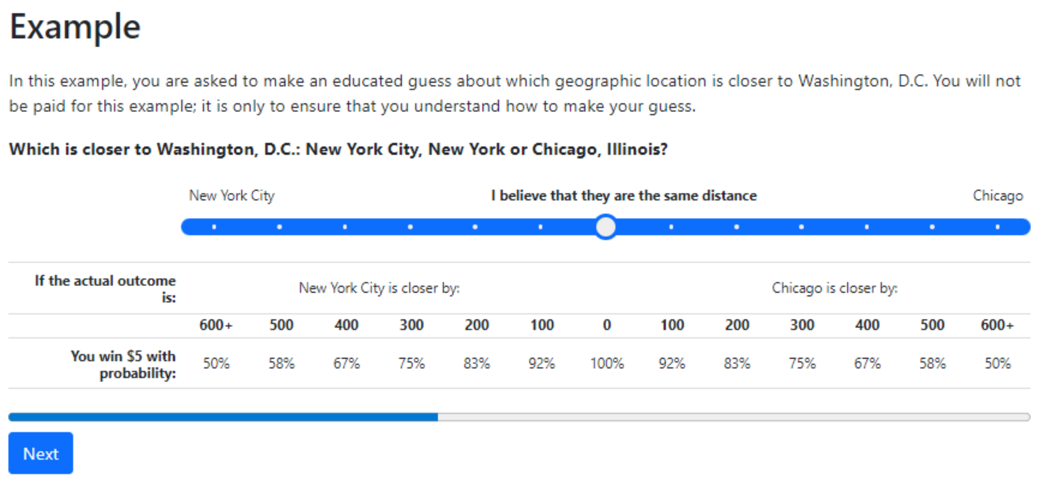
For example, if your guess is accurate and New York City is 300 miles closer than Chicago, you win the $5 for sure (100%). On the other hand, if New York City is actually 100 miles closer, you have a 83% chance of winning the $5. If Chicago is closer than New York City by 300 miles, your chance of winning the $5 falls to 50%.

As you move the slider, the chart will update to show the probabilities of winning the $5 at each possible value of the actual distance. So, if you decided New York City was actually 400 miles closer than Chicago, the chart would change when you moved the slider.

New York City **I believe that New York City is closer by *400* miles** Chicago

If the actual outcome is:			New York City is closer by:							Chicago is closer by:			
	600	**500**	**400**	**300**	**200**	**100**	**0**	**100**	**200**	**300**	**400**	**500**	**600**
You win $5 with probability:	83%	92%	100%	92%	83%	75%	67%	58%	50%	41%	33%	24%	16%

You will now have an opportunity to test the slider and make your guess. Remember, this example is just for practice and you will not be paid for the results.

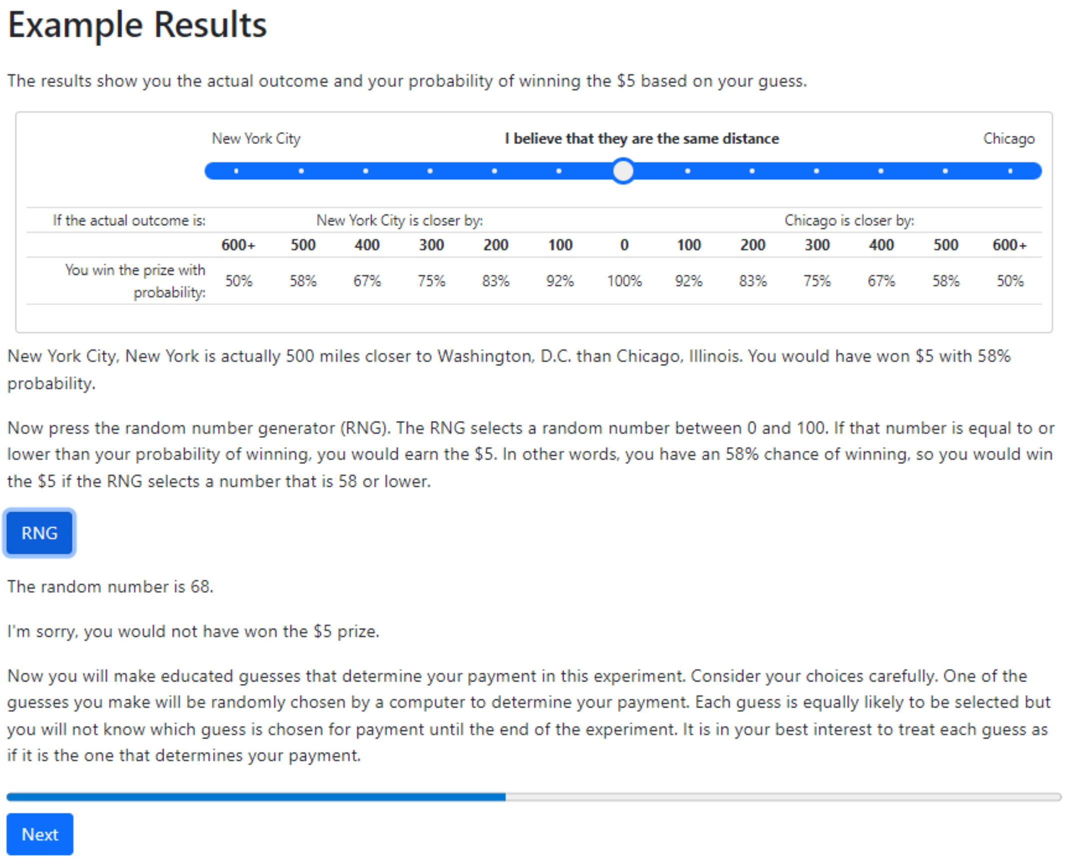


Example

In this example, you are asked to make an educated guess about which geographic location is closer to Washington, D.C. You will not be paid for this example; it is only to ensure that you understand how to make your guess.

Which is closer to Washington, D.C.: New York City, New York or Chicago, Illinois?

If the actual outcome is:	New York City is closer by:							Chicago is closer by:					
	600+	**500**	**400**	**300**	**200**	**100**	**0**	**100**	**200**	**300**	**400**	**500**	**600+**
You win $5 with probability:	50%	58%	67%	75%	83%	92%	100%	92%	83%	75%	67%	58%	50%



Example Results

The results show you the actual outcome and your probability of winning the $5 based on your guess.

If the actual outcome is:	New York City is closer by:							Chicago is closer by:					
	600+	**500**	**400**	**300**	**200**	**100**	**0**	**100**	**200**	**300**	**400**	**500**	**600+**
You win the prize with probability:	50%	58%	67%	75%	83%	92%	100%	92%	83%	75%	67%	58%	50%

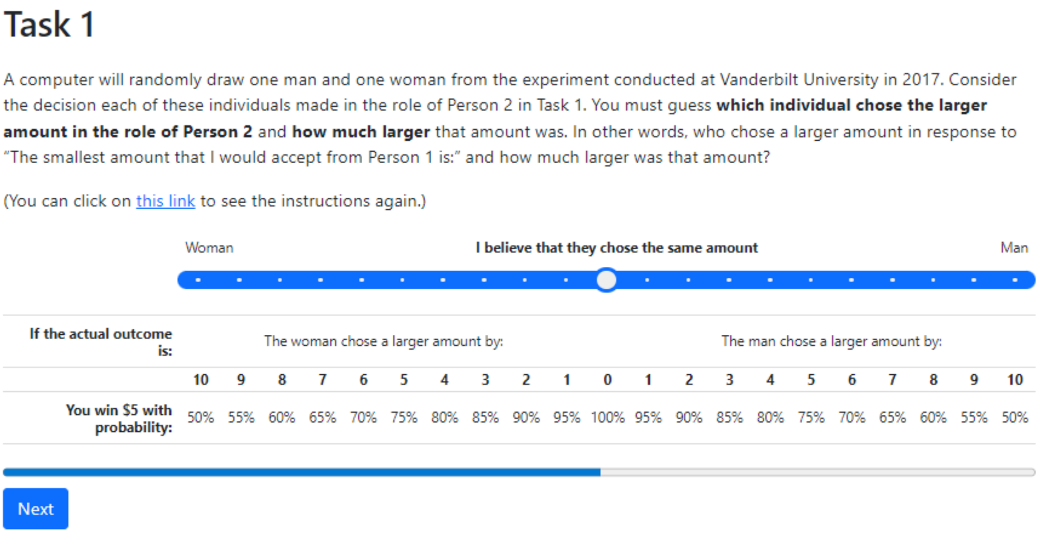
New York City, New York is actually 500 miles closer to Washington, D.C. than Chicago, Illinois. You would have won $5 with 58% probability.

Now press the random number generator (RNG). The RNG selects a random number between 0 and 100. If that number is equal to or lower than your probability of winning, you would earn the $5. In other words, you have an 58% chance of winning, so you would win the $5 if the RNG selects a number that is 58 or lower.

The random number is 68.

I'm sorry, you would not have won the $5 prize.

Now you will make educated guesses that determine your payment in this experiment. Consider your choices carefully. One of the guesses you make will be randomly chosen by a computer to determine your payment. Each guess is equally likely to be selected but you will not know which guess is chosen for payment until the end of the experiment. It is in your best interest to treat each guess as if it is the one that determines your payment.

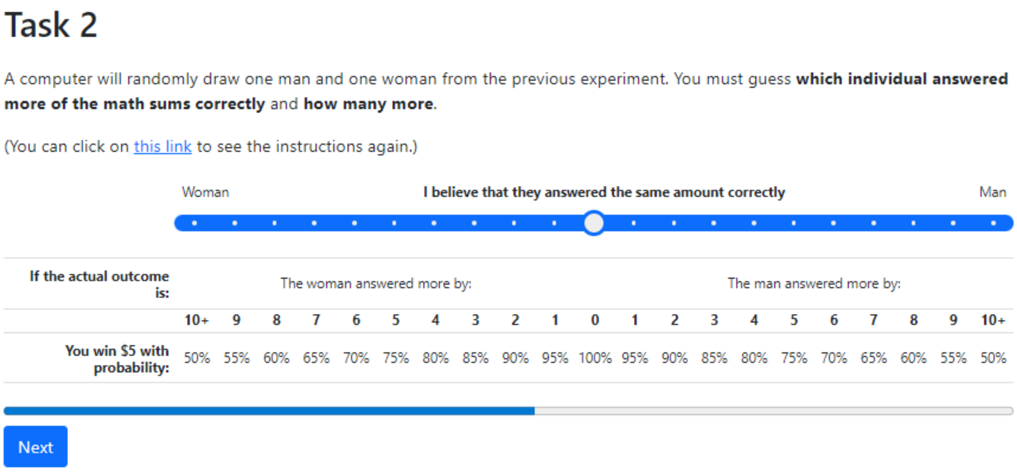


Task 1

A computer will randomly draw one man and one woman from the experiment conducted at Vanderbilt University in 2017. Consider the decision each of these individuals made in the role of Person 2 in Task 1. You must guess **which individual chose the larger amount in the role of Person 2** and **how much larger** that amount was. In other words, who chose a larger amount in response to "The smallest amount that I would accept from Person 1 is:" and how much larger was that amount?

(You can click on this link to see the instructions again.)

If the actual outcome is:	The woman chose a larger amount by:											The man chose a larger amount by:									
	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%

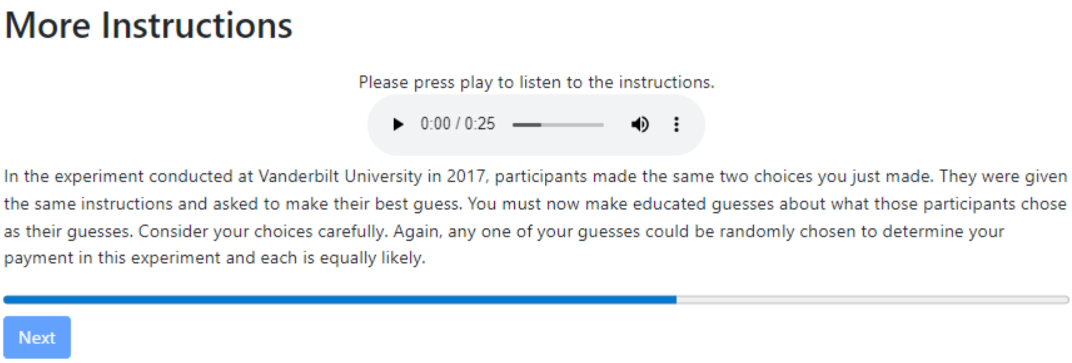


Task 2

A computer will randomly draw one man and one woman from the previous experiment. You must guess **which individual answered more of the math sums correctly** and **how many more**.

(You can click on this link to see the instructions again.)

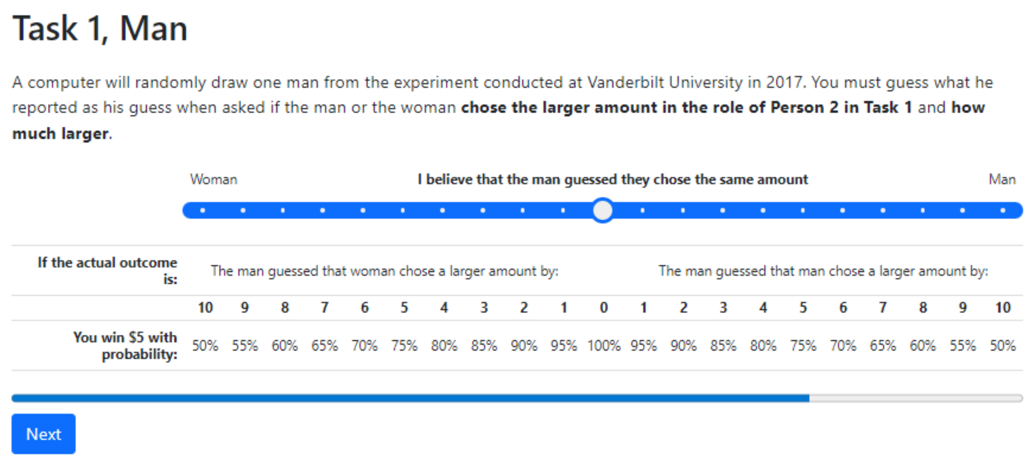
If the actual outcome is:	The woman answered more by:											The man answered more by:									
	10+	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10+
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%



More Instructions

Please press play to listen to the instructions.

In the experiment conducted at Vanderbilt University in 2017, participants made the same two choices you just made. They were given the same instructions and asked to make their best guess. You must now make educated guesses about what those participants chose as their guesses. Consider your choices carefully. Again, any one of your guesses could be randomly chosen to determine your payment in this experiment and each is equally likely.

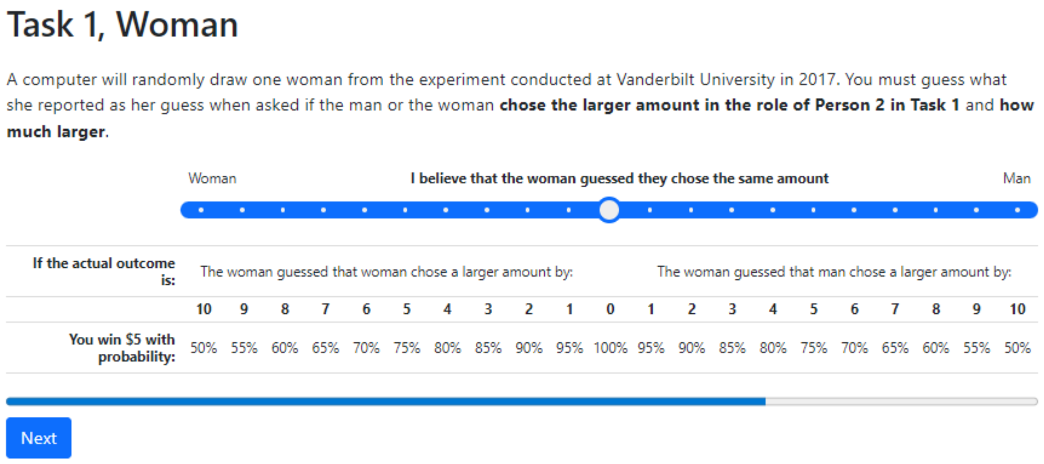


Task 1, Man

A computer will randomly draw one man from the experiment conducted at Vanderbilt University in 2017. You must guess what he reported as his guess when asked if the man or the woman **chose the larger amount in the role of Person 2 in Task 1** and **how much larger**.

Woman | I believe that the man guessed they chose the same amount | Man

If the actual outcome is:	The man guessed that woman chose a larger amount by:											The man guessed that man chose a larger amount by:									
	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%

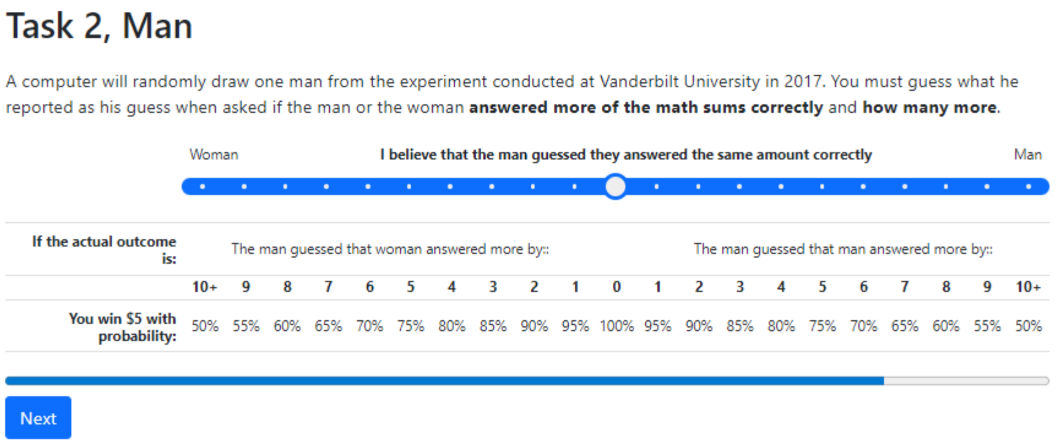


Task 1, Woman

A computer will randomly draw one woman from the experiment conducted at Vanderbilt University in 2017. You must guess what she reported as her guess when asked if the man or the woman **chose the larger amount in the role of Person 2 in Task 1** and **how much larger**.

Woman | I believe that the woman guessed they chose the same amount | Man

If the actual outcome is:	The woman guessed that woman chose a larger amount by:											The woman guessed that man chose a larger amount by:									
	10	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%

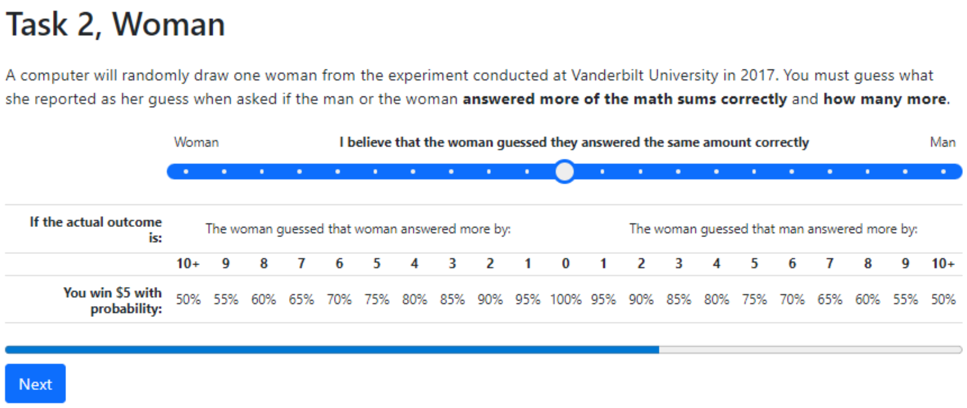


Task 2, Man

A computer will randomly draw one man from the experiment conducted at Vanderbilt University in 2017. You must guess what he reported as his guess when asked if the man or the woman **answered more of the math sums correctly** and **how many more**.

Woman | I believe that the man guessed they answered the same amount correctly | Man

If the actual outcome is:	The man guessed that woman answered more by::											The man guessed that man answered more by::									
	10+	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	10+
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%

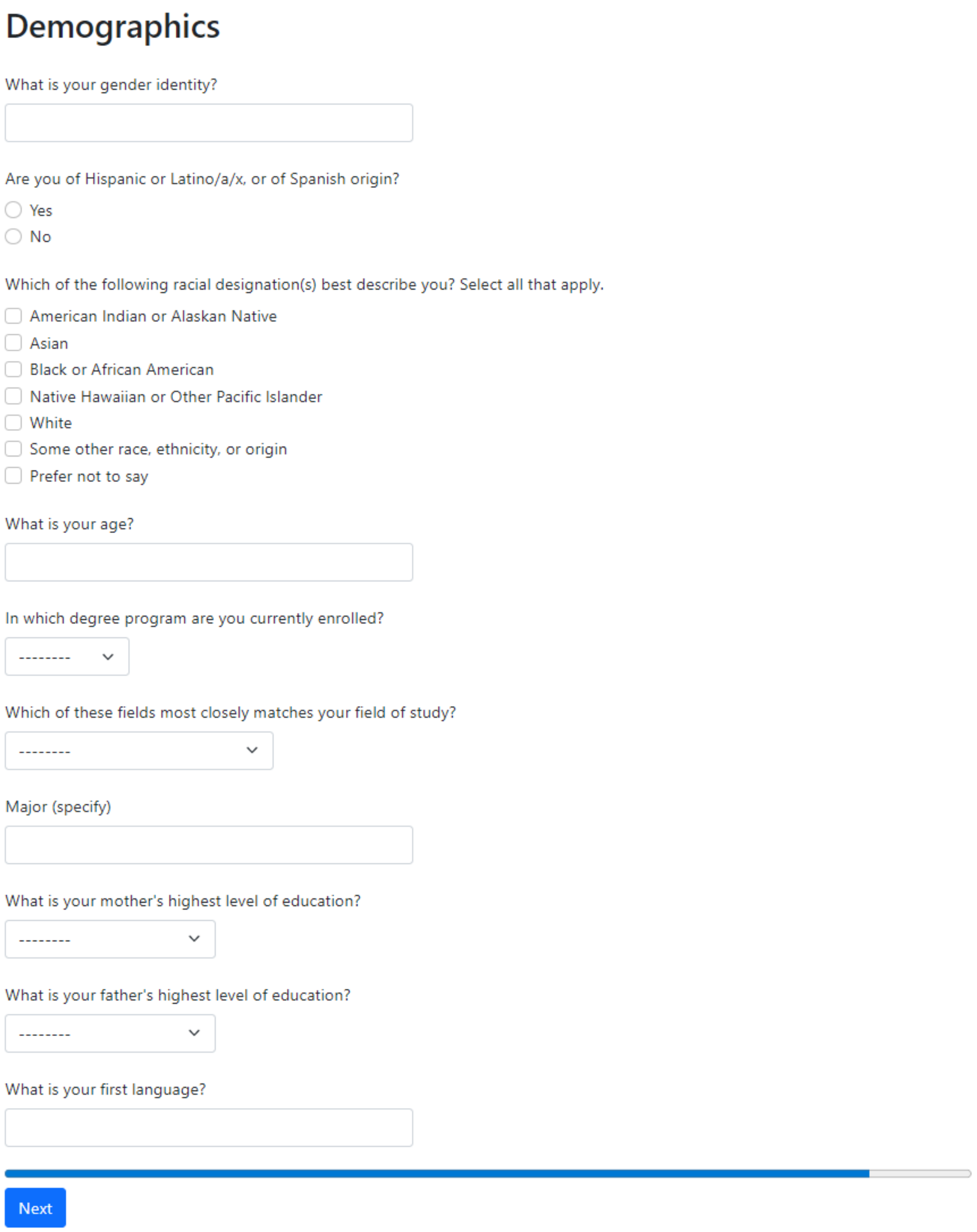


Task 2, Woman

A computer will randomly draw one woman from the experiment conducted at Vanderbilt University in 2017. You must guess what she reported as her guess when asked if the man or the woman **answered more of the math sums correctly** and **how many more**.

Woman
I believe that the woman guessed they answered the same amount correctly
Man

If the actual outcome is:	The woman guessed that woman answered more by:											The woman guessed that man answered more by:									
	10+	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%



Demographics

What is your gender identity?

Are you of Hispanic or Latino/a/x, or of Spanish origin?

○ Yes
○ No

Which of the following racial designation(s) best describe you? Select all that apply.

□ American Indian or Alaskan Native
□ Asian
□ Black or African American
□ Native Hawaiian or Other Pacific Islander
□ White
□ Some other race, ethnicity, or origin
□ Prefer not to say

What is your age?

In which degree program are you currently enrolled?

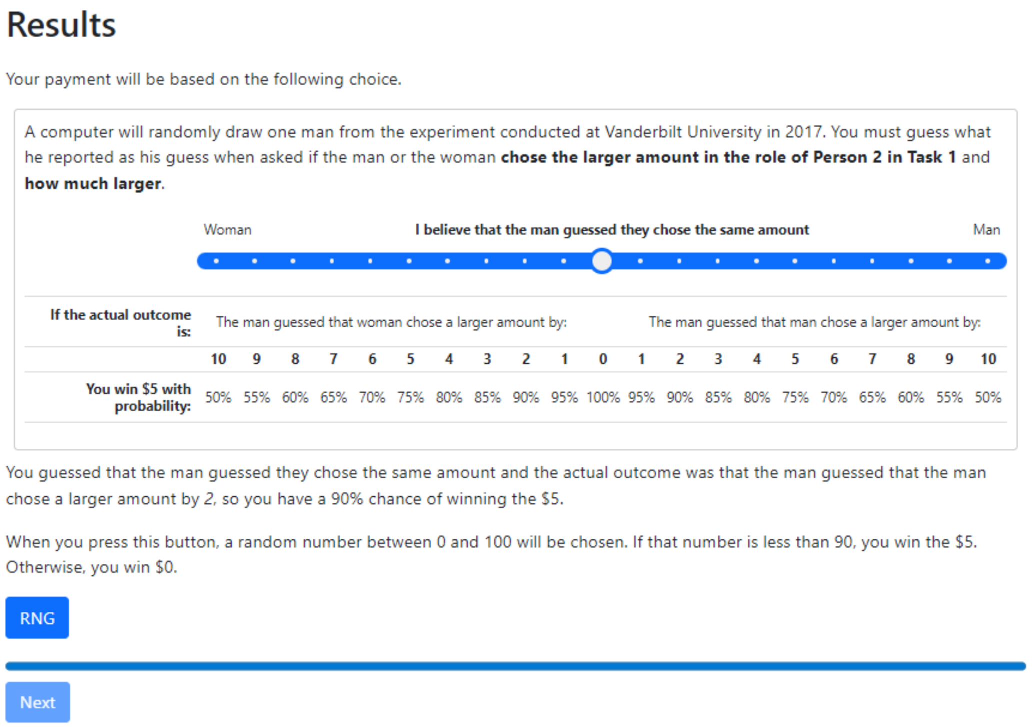
Which of these fields most closely matches your field of study?

Major (specify)

What is your mother's highest level of education?

What is your father's highest level of education?

What is your first language?



Results

Your payment will be based on the following choice.

A computer will randomly draw one man from the experiment conducted at Vanderbilt University in 2017. You must guess what he reported as his guess when asked if the man or the woman **chose the larger amount in the role of Person 2 in Task 1** and **how much larger**.

I believe that the man guessed they chose the same amount

If the actual outcome is:	The man guessed that woman chose a larger amount by:											The man guessed that man chose a larger amount by:									
	10	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
You win $5 with probability:	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%

You guessed that the man guessed they chose the same amount and the actual outcome was that the man guessed that the man chose a larger amount by *2*, so you have a 90% chance of winning the $5.

When you press this button, a random number between 0 and 100 will be chosen. If that number is less than 90, you win the $5. Otherwise, you win $0.

Online Appendix C: Robustness for Attention

In this appendix, we explore the potential effects of inattention on the distributions of reported beliefs. While participants in the in-person experiment sessions were unable to speed through the experiment, online participants may have. After requiring participants to wait on the instruction pages for the duration of the audio playback, we allowed online participants to proceed at their own pace, but recorded the time spent on each page. We set two possible criteria by which to drop potentially inattentive online participants. The first drops anyone who spent less than three seconds on any of the six elicitation tasks. The second drops anyone who did not spend at least 10 seconds on at least one of the tasks. These exclusion criteria reduce the sample size from 354 to 338 and 350 participants, respectively. Appendix Tables C 1 through C 4 report the ternary belief distributions for math and ultimatum tasks under these criteria. Both the levels of these beliefs and their gender differences are nearly unchanged.

Table C 1: Belief elicitation results for math task, dropping participants who ever spend less than 3 seconds on a task

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. First-Order Beliefs				
W>M	0.355	0.329	0.380	-0.051
	(0.026)	(0.036)	(0.037)	(0.052)
W=M	0.163	0.144	0.181	-0.038
	(0.020)	(0.027)	(0.029)	(0.040)
W<M	0.482	0.527	0.439	0.088
	(0.027)	(0.039)	(0.038)	(0.054)
Panel B. Second-Order Beliefs, about Men				
W>M	0.136	0.132	0.140	-0.009
	(0.019)	(0.026)	(0.027)	(0.037)
W=M	0.127	0.156	0.099	0.056
	(0.018)	(0.028)	(0.023)	(0.036)
W<M	0.737	0.713	0.760	-0.048
	(0.024)	(0.035)	(0.033)	(0.048)
Panel C. Second-Order Beliefs, about Women				
W>M	0.438	0.443	0.433	0.010
	(0.027)	(0.038)	(0.038)	(0.054)
W=M	0.210	0.246	0.175	0.070
	(0.022)	(0.033)	(0.029)	(0.044)
W<M	0.352	0.311	0.392	-0.080
	(0.026)	(0.036)	(0.037)	(0.052)
Observations	338	167	171	338

Notes: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows "W>M", "W=M", and "W<M" report the proportion of participants in the math task who believe that the woman scores higher, the woman scores the same, the woman scores lower compared to the man.

Table C 2: Belief elicitation results for math task, dropping participants who never spend at least 10 seconds on a task

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. First-Order Beliefs				
W>M	0.349	0.320	0.376	-0.057
	(0.025)	(0.036)	(0.036)	(0.051)
W=M	0.186	0.169	0.202	-0.034
	(0.021)	(0.029)	(0.030)	(0.041)
W<M	0.466	0.512	0.421	0.090
	(0.027)	(0.038)	(0.037)	(0.053)
Panel B. Second-Order Beliefs, about Men				
W>M	0.137	0.134	0.140	-0.007
	(0.018)	(0.026)	(0.026)	(0.037)
W=M	0.146	0.174	0.118	0.056
	(0.019)	(0.029)	(0.024)	(0.038)
W<M	0.717	0.692	0.742	-0.050
	(0.024)	(0.035)	(0.033)	(0.048)
Panel C. Second-Order Beliefs, about Women				
W>M	0.426	0.430	0.421	0.009
	(0.026)	(0.038)	(0.037)	(0.053)
W=M	0.229	0.262	0.197	0.065
	(0.022)	(0.034)	(0.030)	(0.045)
W<M	0.346	0.308	0.382	-0.074
	(0.025)	(0.035)	(0.036)	(0.051)
Observations	350	172	178	350

Notes: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows "W>M", "W=M", and "W<M" report the proportion of participants in the math task who believe that the woman scores higher, the woman scores the same, the woman scores lower compared to the man.

Table C 3: Belief elicitation results for ultimatum task, dropping participants who ever spend less than 3 seconds on a task

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. First-Order Beliefs				
W>M	0.160	0.162	0.158	0.004
	(0.020)	(0.028)	(0.028)	(0.040)
W=M	0.189	0.210	0.170	0.040
	(0.021)	(0.031)	(0.029)	(0.043)
W<M	0.651	0.629	0.673	-0.044
	(0.026)	(0.037)	(0.036)	(0.052)
Panel B. Second-Order Beliefs, about Men				
W>M	0.219	0.228	0.211	0.017
	(0.022)	(0.032)	(0.031)	(0.045)
W=M	0.180	0.144	0.216	-0.073
	(0.021)	(0.027)	(0.031)	(0.042)
W<M	0.601	0.629	0.573	0.056
	(0.027)	(0.037)	(0.038)	(0.053)
Panel C. Second-Order Beliefs, about Women				
W>M	0.180	0.222	0.140	0.081
	(0.021)	(0.032)	(0.027)	(0.042)
W=M	0.186	0.162	0.211	-0.049
	(0.021)	(0.028)	(0.031)	(0.042)
W<M	0.633	0.617	0.649	-0.032
	(0.026)	(0.038)	(0.036)	(0.052)
Observations	338	167	171	338

Notes: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows "W>M", "W=M", and "W<M" report the proportion of participants in the bargaining task who believe that the woman chooses higher MAO, the woman chooses the same, the woman chooses lower MAO compared to the man.

Table C 4: Belief elicitation results for ultimatum task, dropping participants who never spend at least 10 seconds on a task

	(1) All	(2) Men	(3) Women	(4) Difference
Panel A. First-Order Beliefs				
W>M	0.157	0.157	0.157	-0.000
	(0.019)	(0.028)	(0.027)	(0.039)
W=M	0.211	0.233	0.191	0.042
	(0.022)	(0.032)	(0.029)	(0.044)
W<M	0.631	0.610	0.652	-0.041
	(0.026)	(0.037)	(0.036)	(0.052)
Panel B. Second-Order Beliefs, about Men				
W>M	0.214	0.221	0.208	0.013
	(0.022)	(0.032)	(0.030)	(0.044)
W=M	0.203	0.169	0.236	-0.067
	(0.021)	(0.029)	(0.032)	(0.043)
W<M	0.583	0.610	0.556	0.054
	(0.026)	(0.037)	(0.037)	(0.053)
Panel C. Second-Order Beliefs, about Women				
W>M	0.180	0.215	0.146	0.069
	(0.021)	(0.031)	(0.026)	(0.041)
W=M	0.206	0.180	0.230	-0.050
	(0.022)	(0.029)	(0.032)	(0.043)
W<M	0.614	0.605	0.624	-0.019
	(0.026)	(0.037)	(0.036)	(0.052)
Observations	350	172	178	350

Notes: Gender-specific means and standard errors (in parentheses) are in columns (2) and (3). Column (4) gives the difference in means and its standard error (in parentheses). The rows “W>M”, “W=M”, and “W<M” report the proportion of participants in the bargaining task who believe that the woman chooses higher MAO, the woman chooses the same, the woman chooses lower MAO compared to the man.