Effect of Feedback on Beliefs About Self-Ability *

Ian Chadd † Elif B. Osun‡

September 9, 2026

Abstract

We study the effect of different feedback structures on belief updating in an ego-relevant task using a controlled experiment. Across treatments, subjects receive feedback through a signal with either a noise component, a comparison component, or both. The first two signals are commonly used in the literature, while we develop the latter to systematically analyze the effect of noise and comparison components on belief updating. We find that the signal structure is an important determinant of how subjects update their beliefs. This is driven by men and women exhibiting different biases depending on whether the signal is noisy or comparative. Men underweight bad news when the signal has a noise component and women underweight good news when the signal has a comparison component. These findings have implications for policies aiming to reduce the well-established gender gap in self-confidence through feedback provision.

JEL Codes: J16, C90, D83
Keywords: belief updating, biased beliefs, performance feedback, signal structures, gender differences

*We thank Billur Aksoy, Kai Barron, Alexander Coutts, Christoph Drobner, Keaton Ellis, Emel Filiz-Ozbay, Ginger Jin, Boon Han Koh, Zhenxun Li, Yusufcan Masatlioglu, Zahra Murad, Peter Murrell, Erkut Ozbay, Daniel Reck, John Shea, Neslihan Uler, Daniel Vincent, Joël van der Weele, and Ece Yegane for helpful advice, conversations, and suggestions regarding this project. We also thank participants at the Economic Science Association 2022 European Meeting, the Economic Science Association 2022 North America Meeting, and the University of Maryland Department of Economics Seminars for their helpful comments. We gratefully acknowledge the research support provided by Behavioral and Social Sciences Dean's Research Initiative Award, University of Maryland, College Park.

†Department of Economics, Rensselaer Polytechnic Institute, Email: chaddi@rpi.edu

‡Bates White Economic Consulting, Email: elif.osun@bateswhite.com

1 Introduction

Beliefs about one's own ability shape important life decisions, but there is overwhelming evidence that individuals do not have accurate beliefs about themselves.[1] Distorted beliefs about one's ability can be costly, as they affect economically relevant choices such as which major to declare, which career path to choose, or salary negotiation upon a job offer. One way to correct for misaligned beliefs is to give feedback. However, predicting the effect of feedback on beliefs about ability is not straightforward. The theoretical benchmark for belief updating is Bayes' rule, yet experimental evidence from economics and psychology shows that individuals deviate from Bayes' rule in various ways. Studies focusing on belief updating in ego-relevant domains have not yet reached a consensus on the effect of receiving good versus bad news on how individuals update their beliefs.[2] Understanding the effect of feedback provision is valuable to make well-informed policy recommendations.

In a typical ego-relevant belief updating experiment, subjects complete a task, submit prior beliefs about their relative performance, receive some form of feedback on their performance, and submit posterior beliefs after observing their feedback. There are two signal structures commonly used in the literature for feedback provision: *noisy* and *comparative* signals.[3] Imagine a subject who took a test and is trying to guess whether they are among top or bottom half of performers among a group of individuals who completed the same task. A *noisy* signal reveals whether the subject is among top or bottom half with some accuracy rate, sometimes erroneously revealing the incorrect state. Such a *noisy* signal structure in the field might look like an employee receiving a performance evaluation stating that they are "among the top performers" in their office, where noise is introduced through subjective evaluation, incomplete or noisy performance metrics, among other channels. A *comparative* signal truthfully reveals whether the subject performed better or worse than a randomly chosen opponent among those who completed the same task.[4] The field analogue of this *comparative* signal might be a student who can compare her exam grade to that of a peer, in the absence of complete information about the overall grade distribution. It is not clear whether the differences between the signal structures themselves affect updating behavior, yet the two signals are commonly used in ego-relevant

[1] For example, 88% of U.S. drivers rate themselves safer than the median driver (Svenson, 1981) and only 3.8% of subjects in Niederle and Vesterlund (2007) guess they have the worst performance in a group of 4 people.

[2] We discuss studies documenting belief updating deviations from Bayes' rule in ego-relevant contexts in more detail in Section 2. For a survey of deviations from Bayesian updating in a broader context, see Benjamin (2019).

[3] For example, Buser et al. (2018), Coutts (2019), Schwardmann and Van der Weele (2019), Barron (2021), Möbius et al. (2022), and Drobner and Goerg (2024) use a *noisy* signal, while Eil and Rao (2011), Zimmermann (2020), Coffman et al. (2021), and Drobner (2022) use a *comparative* signal to provide feedback.

[4] To be clear, even the comparative signal structure involves noise in the sense that the opponent to which a subject is compared is chosen randomly. We use the terms *noisy* and *comparative* to distinguish between the two feedback structures based on what is salient about them, recognizing that both involve some component of noise and comparison.

belief updating experiments without acknowledgement of this possibility. In this paper, we systematically analyze the effect of these two commonly used signal structures on belief updating.

We design an experiment to compare belief updating under a *noisy* signal to that under a *comparative* signal structure. In the first part of the experiment, subjects complete an ego-relevant IQ task. We place the subjects in a group of other individuals who completed the same task and rank them based on their performance. In the second part of the experiment, subjects submit their beliefs on their relative performance twice: once before (prior beliefs) and once after they receive some feedback (posterior beliefs). We use a behavioral model based on Grether (1980), which uses weighted log-likelihood ratios of prior beliefs, good news, and bad news to construct posterior beliefs. Estimating the weight on each component allows us to detect deviations from the weights under the Bayesian benchmark.

We vary the signal structure used to generate feedback and compare belief updating behavior across treatments. A direct comparison between updating behavior under *noisy* and *comparative* signals raises two issues. First, the two signals do not necessarily have the same informational content: the informativeness of the *noisy* signal is determined by the accuracy rate of the signal and is the same for all subjects, but the informativeness of the *comparative* signal varies by subject, as it depends on the subject's prior belief distribution over ranks. Second, the *noisy* signal has a noise component but lacks a comparison component, while the reverse is true for the *comparative* signal. Hence, there is a two-dimensional change across the two signals. To address these two issues, we design a novel signal structure that includes both noisy and comparative elements. In this signal structure, the signal subjects receive resembles that of the noisy signal: they are told whether they are in the top half of the performance distribution in a probabilistic manner. However, like in the comparative signal structure, the subject's performance is compared to that of a randomly chosen opponent, which determines the accuracy rate of the signal. The manner in which this accuracy rate is constructed is such that the informational content of the overall signal is isomorphic to that in the treatment with a *noisy* signal.[5] Implementing a signal that has both components allows us to detect the effect of noise and comparison in a controlled way by changing one component at a time.

Gender may affect how individuals update beliefs about their own ability. A large body of experimental literature documents robust gender differences in self-confidence, with men displaying more overconfidence than women (Barber and Odean, 2001; Niederle and Vesterlund, 2007).[6] This gender difference

[5]The behavioral model that we use incorporates the informativeness of the signals, hence a comparison between *noisy* and *comparative* signals is still possible despite the differences in information contents across treatments. By generating an additional signal that has a comparison component but is informationally isomorphic to the *noisy* signal, we are able to rule out information differences across treatments to be the driving mechanism of updating differences across treatments.

[6]Men also exhibit motivated reasoning with respect to beliefs about relative performance, while women do not (Thaler, 2021).

in self-confidence may contribute to the well-established gender gap in labor market outcomes through human capital choices.[7] This gap remains persistent, with women earning 83 cents on the dollar relative to men (Shrider et al., 2021) even though women make up more than half (50.7%) of the college-educated labor force in the United States (Fry, 2022). Feedback provision can be a valuable intervention to shrink the gender gap in labor market outcomes, yet what type of feedback is the most appropriate for this purpose is an open question. Evidence suggests that men and women react differently to feedback. Women are documented to dislike competition (Niederle and Vesterlund, 2007), so they might react to signals that have a salient competitive element differently than men. In a study of responsiveness to exposure to peer excellence in education, women are i) more likely than men to rate an essay they wrote as "worse" than a peer's and ii) less likely than men to persist in response to this negative comparison when exposed to high-quality peer work (Rogers and Feller, 2016).[8] Women are also more sensitive to receiving lower grades than men when it comes to persistence in economics courses at the undergraduate level (Rask and Tiefenthaler, 2008). For academic economists, female assistant professors react to a rejection during peer review by lowering the perceived likelihood of subsequently publishing in a leading journal more strongly than male assistant professors (Shastry and Shurchkov, 2024). Taken together, these results suggest that both salient noise and comparison may affect how men and women update differently, justifying a systematic test using a controlled experiment.

This work on the effects of feedback structure builds on experimental work focused on gender differences in belief updating more generally. Coffman et al. (2019) explore how feedback affects gender differences in self-assessments and find that men and women exhibit different updating patterns upon receiving feedback, depending on the gender-congruency of the task. They document that men update their beliefs more optimistically than women if the task is male-typed (and vice versa if the task is female-typed). Since subjects have higher self-confidence for their performance on tasks that are in their gender's domain to begin with, receiving feedback in this setup actually fuels persistence in the gender gap in self-confidence. The focus of Coffman et al. (2019) is the gender-congruency of the task and not the type of the signal used to provide feedback. Shastry et al. (2020) show that men tend to attribute bad news to luck while women attribute it to ability. In this paper, we examine whether signals with noise and comparison components affect belief updating behavior differently for men and women, as this could help in designing policies to

[7] For example, Cortés et al. (2021) find that gender differences in overconfidence lead to differences in job search behavior of men and women college students.

[8] Though this difference is statistically insignificant because the study was underpowered for subgroup analysis. Note that this result was not reported in Rogers and Feller (2016); we conducted this analysis using their publicly available data.

reduce the gender gap in self-confidence.[9]

We find that using different signal structures affects belief updating behavior. Although isomorphic in their informational content, receiving a signal with only a noise component leads to different deviations from the theoretical benchmark compared to receiving a signal with both noise and comparison components. The difference is driven by men and women exhibiting different updating behavior depending on whether the signal has a noise or comparison component. We examine updating behavior of men and women separately under three treatments. We find that women never underweight bad news and men never underweight good news relative to the Bayesian benchmark. In contrast, how women update under good news and how men update under bad news is sensitive to signal type. Men underweight bad news in both treatments in which the signal has a noise component, whereas women underweight good news in both treatments in which the signal has a comparison component. These findings imply that for policies aiming to shrink the well documented gender gap in self-confidence, providing feedback with a noise component is not ideal if bad news is more prevalent, whereas providing feedback with a comparison component is not ideal if good news is more common. We conduct an ex-post analysis on gender differences in posterior beliefs and find suggestive evidence in line with these implications.

This paper contributes to the literature in several ways. This is the first study to systematically analyze the effect of different feedback structures on belief updating in a unified framework. We generate a novel signal structure that allows comparison between the effect of noise and comparative components of feedback in a controlled manner.[10] This is also the first paper documenting gender differences in how men and women perceive news under different signal structures. Our findings suggest that policies aiming to reduce the gender gap in self-confidence by providing feedback on performance should carefully take the feedback structure and the performance of the target population into account; otherwise, providing feedback might be ineffective.[11]

The remainder of this paper is organized as follows. Section 2 discusses the related literature. Section 3 introduces the experimental design, treatments, and experimental protocol. Section 4 explains the methodology for measuring biases in belief updating and presents the results. Section 5 concludes.

[9] In Section 2, we discuss other papers that document gender differences in belief updating behavior.

[10] Coutts et al. (2020) independently developed a similar feedback structure to examine self-serving bias when updating beliefs under multiple sources of uncertainty.

[11] In fact, performance feedback may sometimes lead to worse outcomes. For example, Azmat et al. (2019) find that providing relative performance feedback decreases students' educational performance in a higher education setting field experiment.

2 Related Literature

The main assumption of the neoclassical approach to belief formation is that upon receiving new information, individuals revise their beliefs according to Bayes' rule. Early experiments in the psychology literature documenting deviations from Bayes' rule using hypothetical belief updating questions include Edwards (1968) and Kahneman and Tversky (1972). These studies provide exogenous priors and compare the subjects' posterior beliefs to the Bayesian benchmark for the given signals about the underlying state. In contexts such as beliefs about self-ability, the prior beliefs are endogeneous and heterogeneous across subjects, so comparing posterior beliefs to the Bayesian benchmark is not sufficient to determine the source of updating deviations. Grether (1980) introduced a model of belief updating that allows one to detect deviations from the Bayesian weights on prior beliefs and signals separately through parameter estimation. A number of recent studies on belief updating, including this paper, use Grether's model to detect updating deviations from Bayes' rule (e.g. Möbius et al., 2022; Barron, 2021).

In the context of belief updating when information is ego-relevant, such as information about one's ability, the theoretical literature proposes several models to explain deviations from Bayes' rule. Landier (2000) proposes a model in which beliefs have a hedonic component through anticipation utility. Köszegi (2006)'s subjects derive ego utility from positive views about their ability to do well in a skill-sensitive task. Mayraz (2011) provides an axiomatic model in which beliefs are affected by desires. More recently, Möbius et al. (2022) build a model of optimally-biased Bayesian updating. The common prediction of all these models is that good news is weighted more than bad news.[12]

Experimental studies focusing on belief updating in ego-relevant domains lack consensus on the weights assigned to good versus bad news when updating beliefs. Eil and Rao (2011) are among the first to document belief updating deviations for good versus bad news. They study updating in response to news about beauty and intelligence, and find that subjects give more weight to good compared to bad news. Möbius et al. (2022), Drobner (2022), and Drobner and Goerg (2024) also find that positive information is weighted more heavily than negative when updating beliefs in an IQ-related quiz. In contrast, Ertac (2011) examines updating in response to news about performance on tasks requiring ability and effort and finds that individuals incorporate bad news more into their beliefs than good news. Coutts (2019) finds that bad news receives more weight compared to good news when updating beliefs in ego-relevant, financially-relevant, and neu-

[12] Confirmation bias is one additional mechanism proposed to explain deviations from Bayes' rule. Rabin and Schrag (1999) build a model of confirmation bias, in which individuals give more weight to information that conforms with their prior beliefs. However, the experimental studies (including this paper) did not find direct evidence for confirmation affecting belief updating in ego-relevant domains. For example, Eil and Rao (2011) find that valence is the underlying cause of confirmatory bias and that confirmation alone has no effect. Möbius et al. (2022) examine and find no evidence of confirmation bias.

tral domains. Grossman and Owens (2012) document that subjects have overconfident beliefs about their performance on an intelligence-based task, but their belief updating follows the Bayesian benchmark upon receiving both good and bad news. Barron (2021) uses a financially-relevant task that is not ego-relevant but with payoffs such that subjects prefer one state over the other, and also finds updating in line with Bayes' rule. Buser et al. (2018), Schwardmann and Van der Weele (2019), and Zimmermann (2020) find that subjects do not give enough weight to their signals when updating their beliefs relative to Bayes' rule, but give equal weight to good and bad news.[13] All the studies mentioned here use a single feedback structure in their experimental design, and none examine the effect of the feedback structure on belief updating.

A few papers propose mechanisms that can contribute to the lack of consensus in belief updating behavior across experimental studies. Drobner (2022) shows that expectations about resolution of uncertainty affect belief updating behavior. Using an IQ test and exogeneously manipulating subjects' expectations about the resolution of uncertainty, he finds that those who are informed that their true rank will not be revealed at the end of the experiment update their beliefs optimistically, while those who are informed that they will learn their true rank at the end of the experiment update their beliefs neutrally. Drobner and Goerg (2024) manipulate the ego-relevance of a similar IQ test by providing evidence for and against the importance of IQ tests across subjects. They find that positive information is weighted more heavily than negative only in the high-ego treatment. Coffman et al. (2019) examine subjects' beliefs about their performance on tasks that vary in their gender-congruency and document that gender stereotypes influence belief updating: subjects give more weight to good news over bad when the signal arrives in a gender-congruent domain. Coutts (2019) uses a feedback structure that noisily informs subjects whether they were among the top 15% of performers and finds that subjects give more weight to bad news compared to good news when updating their beliefs. Even though his experiment is not designed to test the effect of the informational content of signals on belief updating, he considers the use of negatively skewed signals as an ex-post explanation of bad news receiving more weight than good news. In this paper, we consider the type of signal structure used to give feedback as another mechanism that can affect belief updating behavior.

In addition to Coffman et al. (2019), discussed in the introduction, several papers document gender differences in belief updating. Ertac (2011) finds that women update their beliefs more pessimistically, by giving less weight to good news compared to men. The gender difference in belief updating arises only in the verbal GRE task, which was perceived as more difficult by the subjects, but not in the easy algebraic addition task. Möbius et al. (2022) and Coutts (2019) find that women update their beliefs more conservatively than

[13] Zimmermann (2020) also examines belief updating in the long run. The results reported here are the findings immediately after feedback. In the long run, he finds that the effect of receiving good news persists, but the effect of bad news fade over time.

men both for good and bad news, but are not significantly more asymmetric. In Coutts et al. (2020), men are significantly more responsive to good news relative to bad news when they receive feedback about their own performance, while women do not update their beliefs asymmetrically. Coffman et al. (2021) examine the effect of feedback on beliefs in a dynamic setting and show that while both men and women underweight both type of signals relative to the Bayesian benchmark, the effect of bad news on men's beliefs fades more over time compared to women's beliefs, leading to persistent gender differences in self-confidence in the long run. Similar to other papers studying belief updating, these studies use a single feedback structure in their experimental design and are not designed to test the gender differences in belief updating across signal structures.

Signals with noise and comparative components are frequently used in the literature.[14] This paper is concerned with investigating the effect of different feedback structures on belief updating and their differential effect by gender. One cannot address this question by sorting the existing literature on the type of the signal used and making a meta-analysis, as various other aspects of the experimental design differ across studies, including type of the task (e.g. SAT questions, logic questions, raven's matrices, ASVAB questions, summation task, beauty task) and the performance measure used for belief elicitations (e.g. the likelihood of being among top or bottom performers, the likelihood of being among a pre-determined percentile, expected rank, absolute score). Hence, there is need for a controlled experiment to investigate the effect of different signal structures on belief updating in a unified framework.

[14] Grossman and Owens (2012), Buser et al. (2018), Coutts (2019), Schwardmann and Van der Weele (2019), Barron (2021), Möbius et al. (2022), and Drobner and Goerg (2024) use a noisy signal. Ertac (2011) uses a variation of the comparative signal in which the comparison is not against a single opponent, but against a group of opponents. Eil and Rao (2011), Zimmermann (2020), Coffman et al. (2021), Drobner (2022) use a comparative signal.

3 Experimental Design

We designed the experiment using the experimental software oTree (Chen et al., 2016) and conducted it online on Prolific during April and May 2022. We recruited 901 subjects from the U.S. subject pool.[15] No subject participated in the experiment more than once. Median completion time was about 10 minutes and median payment was about $13 per hour excluding the completion fee.[16]

The experiment consisted of four parts, detailed below, and an exit questionnaire. Figure 1 summarizes the timeline of the experiment. In the first part of the experiment, subjects completed an IQ task. Upon completing the test, subjects were informed that they were randomly placed in a group of 9 other participants who previously solved the same test.[17] Then, subjects submitted their beliefs on their relative performance among this group, both before and after they received feedback. The experiment had a between-subject design, with each treatment using a different signal structure for feedback provision.

Figure 1: Timeline of the experiment

3.1 Test Stage

Subjects had four minutes to answer as many questions as possible. The test consisted of questions typically used to measure IQ, an ego-relevant belief domain. Questions were standard logic questions similar to those used in Möbius et al. (2022) and Cognitive Reflection Test questions (Frederick, 2005), such as:

1. Which one of the five choices makes the best comparison? LIVED to DEVIL as 6323 is to: a) 2336 b) 6232 c) 3236 d) 3326 e) 6332

2. Assume that these two statements are true: All brown-haired men have bad tempers. Larry is a brown-haired man. The statement "Larry has a bad temper" is: a) True b) False c) Unable to determine

[15] From this initial pool, we drop the data of 9 subjects whose reply to the survey question about their gender is inconsistent with their demographic data on Prolific, one subject who revoked their consent after completing the study, and one subject who timed out and as a result could not be paid.

[16] The completion fee was $1.1 for about one third of the participants and was increased to $1.4 for the remaining two third after Prolific increased the minimum hourly participant reward from $6.5 to $8 on April 21, 2022. The total payment including the bonus payments was already above the updated minimum required hourly payment, however the platform makes the minimum payment calculation at the time of announcing the study, and does not take the bonus payments into account. Rather than changing the duration of the experiment to keep the completion fee the same, We increased the completion fee and kept the duration of the experiment the same after the price change.

[17] The 9 other participants for each subject were randomly chosen from a group of Prolific participants who previously completed the same IQ test before data collection for the main experiment began.

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? a) 7 days b) 13 days c) 24 days d) 47 days e) 48 days

Earnings for the quiz were \$0.20 per question answered correctly, so the subjects had a monetary incentive to perform as well as possible. At the test stage, subjects knew that there was going to be another part of the experiment, but they did not know the content of the following part. Hence, subjects did not know that they would submit their beliefs about their relative performance compared to a group of other participants when solving the quiz. This was to avoid any incentive to perform poorly in order to guarantee having correct beliefs about relative performance later. Subjects did not learn their earnings until the end of the experiment, so they could not make any inferences about their performance from their quiz earnings.

3.2 Prior Belief Elicitation Stage

Once the test stage was over, subjects were informed that their performance would be compared to 9 other randomly chosen participants who previously completed the same quiz. To examine belief updating, We focus on subjects' beliefs about whether they were among the top half or bottom half of performers in their group, which is a subjective probability over a binary type space, as updating a single number is more intuitive compared to updating a distribution. However, we also measure subjects' belief distribution over ranks, to calculate the likelihood ratio of the signals when the feedback structure is comparative.[18] For these purposes, subjects were asked the following questions:

Question 1. How do you estimate the likelihood (in percent) of being in each rank when your performance is compared to the other 9 members of your group?

Question 2. What do you think is the likelihood (in percent) that you rank among the top and bottom halves of the performers in the group? In other words, in the group of 10, what do you think is the likelihood that your rank is 1, 2, 3, 4, or 5 (you are among the top half performers) and what do you think is the likelihood that your rank is 6, 7, 8, 9, or 10 (you are among the bottom half performers)?

We asked the above two questions on the same screen to eliminate concerns about anchoring. The experimental interface was split in two parts. The left side of the screen consisted of the first question and allowed subjects to submit their beliefs for the likelihood of being in each rank. The right side of the screen consisted of the second question, and the probabilities of being among the top and bottom half of

[18] The signal types and calculation of likelihood ratios are explained in more detail in Subsection 3.3

performers were calculated in real time as subjects modified their answer to the first question. Once subjects submitted their beliefs, they were asked to confirm that the likelihood of being among the top and bottom half performers reflected their true beliefs in a separate screen, to further increase the salience of the second question. Subjects could go back to the previous screen to edit their answer if they wished (see Appendix B for screenshots).

To eliminate hedging motives, either prior or posterior beliefs were randomly chosen for payment. We incentivized prior beliefs using the quadratic scoring rule (Selten, 1998) with the following formula:

$$100 - 50 \times \sum_{i=1}^{10} (\mathbb{1}[rank = i] - \frac{p_i}{100})^2$$

where $\mathbb{1}[rank = i]$ is an indicator variable that takes the value 1 if subject's rank was equal to i and 0 otherwise, and p_i is their estimate for being in rank $i \in 1, 2, ..., 10$.

Note that subjects were incentivized for their estimates on the likelihood of each rank and not separately for their likelihood of being among the top or bottom half, as incentive compatibility in one question leads to incentive compatibility in the other. An alternative would be to incentivize both questions on prior beliefs separately and randomly choose one to implement. We avoided this in order to minimize the complexity of information we provided to subjects on the screen.

Even though incentive compatibility of the quadratic scoring rule requires assuming risk neutrality, there are several reasons to suggest that this is not an obstacle for interpeting the results of this paper. First, possible earnings from each belief elicitation question ranged from \$0 to \$1, stakes over which one would expect risk-neutrality. Secondly, similar to Eil and Rao (2011) and other papers using the quadratic scoring rule (e.g. Zimmermann, 2020; Barron, 2021), We explicitly told the subjects that they would maximize their expected earnings if they report their true beliefs. Thirdly, Danz et al. (2020) show that truthful likelihood reporting is maximized when subjects are not provided with the exact formula for the payoff calculation. Following this argument, the main experimental screen did not include the explicit formula for payoff calculation, but only included the sentence "Your expected payoff will be the highest if you report your true beliefs." The interested subjects could click on a link to access the exact formula. Lastly, the main results that we focus on in this paper compare belief updating behavior across signal structures. Any tendency to hedge beliefs due to risk preferences in one treatment would likely be the same in other treatments, having no effect on the relative bias across treatments.

3.3 Feedback Stage

The signal structure used to provide feedback varied by treatment and had either a noise component, a comparison component, or both. We call these treatments *Noisy*, *Comparative*, and *NoisyComparative*, respectively. In all of the treatments, subjects received instructions about how their signal would be determined and needed to answer a comprehension question correctly before receiving feedback.

3.3.1 Noisy Treatment

Feedback in the *Noisy* Treatment consisted of a signal with an accuracy rate of 7/9: if a subject was among the top half of performers of their group, they would receive a signal stating that they were among the top half performers (good news) with probability 7/9, and a message stating that they were among the bottom half performers (bad news) with probability 2/9. If a subject was among the bottom half of performers of their group, they would receive bad news with probability 7/9 and good news with probability 2/9. In this treatment, the meaning of the signal has a "noise" component in the sense that it is incorrect with some probability. It does not have a "comparison" component, since the signal is not determined through the subject being compared to another individual. This signal structure is commonly used in the belief updating literature (e.g. Buser et al. (2018), Coutts (2019), Barron (2021), and Möbius et al. (2022)).

Table 1: Signals in *Noisy* Treatment

Performance	**Signal Received**
Top half	"Top half" with 7/9 chance
	"Bottom half" with 2/9 chance
Bottom half	"Bottom half" with 7/9 chance
	"Top half" with 2/9 chance

3.3.2 Comparative Treatment

The signals in the *Comparative* Treatment informed subjects whether their performance was better (good news) or worse (bad news) than a randomly chosen participant in their group. Hence, the signal has a comparison component. There is no noise component in the meaning of the signal, as it always conveys correct information. This is another signal structure commonly used in the belief updating literature (e.g. Eil and Rao (2011), Zimmermann (2020), Coffman et al. (2021), and Drobner (2022)).

Table 2: Signals in *Comparative* Treatment

Performance	**Signal Received**
Better than randomly chosen participant	"Better than the other participant"
Worse than randomly chosen participant	"Worse than the other participant"

Comparison Between Noisy and Comparative Treatments: There are two issues with directly comparing belief updating behavior between the *Noisy* and *Comparative* treatments. First, the informativeness of the signals under the two treatments are not the same. The likelihood ratios of receiving good and bad news in the *Noisy* Treatment are homogeneous across subjects and are determined by the accuracy rate of the signal:

$$LR_G^N = \frac{Pr(s = G \mid t = H)}{Pr(s = G \mid t = L)} = \frac{7}{2} \quad , \quad LR_B^N = \frac{Pr(s = B \mid t = H)}{Pr(s = B \mid t = L)} = \frac{2}{7} \tag{1}$$

where LR_G^N and LR_B^N are the likelihood ratios of receiving a good and a bad signal in the *Noisy* Treatment, s is the signal received (where G and B denote good and bad news), and t is the performance type (where H and L denote being among the top half and bottom half of performers).

The likelihood ratios of receiving good and bad news in *Comparative* Treatment, in contrast, are determined by the subject's prior beliefs over ranks, and hence vary by subject. Denote p_i as the prior probability given to being in each rank $i = \{1, 2, ..., 10\}$ (where 1 is the best performance and 10 is the worst performance). The probability of randomly choosing a participant with worse performance (i.e. receiving a good signal) given rank i in a group of 9 others is $(10 - i)/9$, and the probability of randomly choosing a participant with better performance (i.e. receiving a bad signal) is $(i - 1)/9$. Then,

$$Pr(s = G|t = H) = \frac{Pr((s = G) \cap (t = H))}{Pr(t = H)} = \frac{p_1 \times \frac{9}{9} + p_2 \times \frac{8}{9} + p_3 \times \frac{7}{9} + p_4 \times \frac{6}{9} + p_5 \times \frac{5}{9}}{Pr(t = H)}$$

$$Pr(s = G|t = L) = \frac{Pr((s = G) \cap (t = L))}{Pr(t = L)} = \frac{p_6 \times \frac{4}{9} + p_7 \times \frac{3}{9} + p_8 \times \frac{2}{9} + p_9 \times \frac{1}{9} + p_{10} \times \frac{0}{9}}{Pr(t = L)}$$

$$Pr(s = B|t = H) = \frac{Pr((s = B) \cap (t = H))}{Pr(t = H)} = \frac{p_1 \times \frac{0}{9} + p_2 \times \frac{1}{9} + p_3 \times \frac{2}{9} + p_4 \times \frac{3}{9} + p_5 \times \frac{4}{9}}{Pr(t = H)}$$

$$Pr(s = B|t = L) = \frac{Pr((s = B) \cap (t = L))}{Pr(t = L)} = \frac{p_6 \times \frac{5}{9} + p_7 \times \frac{6}{9} + p_8 \times \frac{7}{9} + p_9 \times \frac{8}{9} + p_{10} \times \frac{9}{9}}{Pr(t = L)}$$

Using the above equations, the likelihood ratios of receiving good and bad news in the *Comparative* Treatment are:

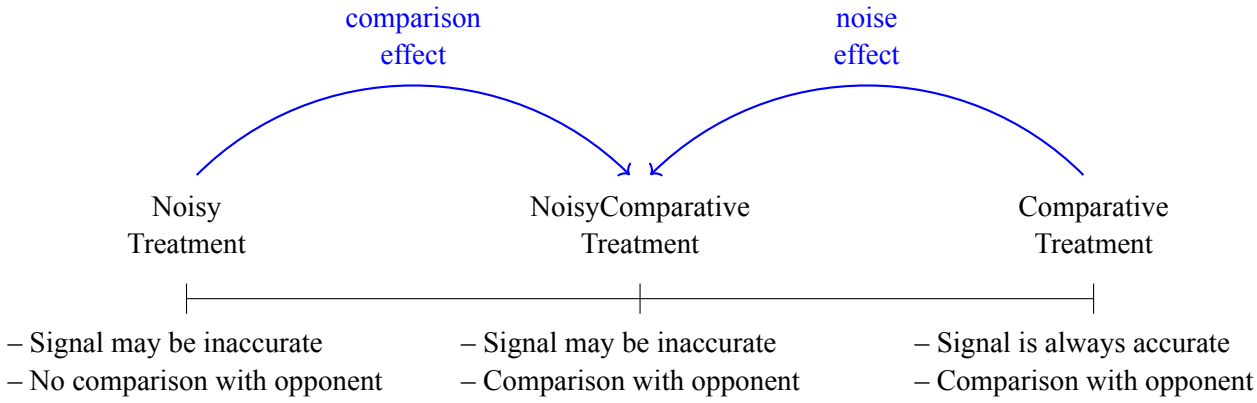
$$LR_G^C = \frac{Pr(s=G|t=H)}{Pr(s=G|t=L)} = \frac{p_1 \times \frac{9}{9} + p_2 \times \frac{8}{9} + p_3 \times \frac{7}{9} + p_4 \times \frac{6}{9} + p_5 \times \frac{5}{9}}{p_6 \times \frac{4}{9} + p_7 \times \frac{3}{9} + p_8 \times \frac{2}{9} + p_9 \times \frac{1}{9} + p_{10} \times \frac{0}{9}} \times \frac{Pr(t=L)}{Pr(t=H)} \quad (2)$$

$$LR_B^C = \frac{Pr(s=B|t=H)}{Pr(s=B|t=L)} = \frac{p_1 \times \frac{0}{9} + p_2 \times \frac{1}{9} + p_3 \times \frac{2}{9} + p_4 \times \frac{3}{9} + p_5 \times \frac{4}{9}}{p_6 \times \frac{5}{9} + p_7 \times \frac{6}{9} + p_8 \times \frac{7}{9} + p_9 \times \frac{8}{9} + p_{10} \times \frac{9}{9}} \times \frac{Pr(t=L)}{Pr(t=H)} \quad (3)$$

Since it is possible that $LR_G^N \neq LR_G^C$ and $LR_B^N \neq LR_B^C$, as can be seen by comparing equations (1)-(3), the informativeness of the signals are not necessarily the same under the *Noisy* and *Comparative* Treatments. We design a novel signal structure that determines the accuracy rate of a noisy signal by comparing the subject to a randomly chosen opponent in such a way that the informational content of the signal is isomorphic to that in the *Noisy* Treatment, as we explain in further detail in the following subsection. The behavioral model that we use incorporates the informativeness of the signals when examining belief updating behavior, so a comparison between *Noisy* and *Comparative* Treatments is still possible, yet having a signal structure with a comparison component that is also informationally isomorphic to the signal in the *Noisy* Treatment allows us to rule out information differences across treatments being the driving mechanism of updating differences across treatments.

The second issue with directly comparing belief updating behavior between the *Noisy* and *Comparative* Treatments is that there are two changes across treatments: the signal in the *Noisy* Treatment has a noise component but lacks a comparison component, while the the reverse is true for the signal in the *Comparative* Treatment. As illustrated in Figure 2, the *NoisyComparative* Treatment acts as a bridge between the two treatments, allowing us to consider the effect of one change at a time.

Figure 2: Summary of treatments



3.3.3 NoisyComparative Treatment

In the *NoisyComparative* Treatment, the accuracy of the signal is determined by comparing the subject's performance type to a randomly chosen participant's. If the subject and the randomly chosen opponent are in different halves of the distribution (i.e. if one is among the top half while the other is among the bottom half of performers), then the signal correctly reveals whether the subject is among the top or bottom. If the subject and the randomly chosen opponent are in the same half of the distribution (i.e. if both are among the top half or both are among the bottom half of performers), then the signal has a 50% chance of revealing the correct type and 50% chance of revealing the incorrect type.

Table 3: Signals in *NoisyComparative* Treatment

Subject Performance	**Opponent Performance**	**Signal Received**
Top half	Bottom half	"Top half"
Bottom half	Top half	"Bottom half"
Top half	Top half	"Top half" with 50% chance "Bottom half" with 50% chance
Bottom half	Bottom half	"Top half" with 50% chance "Bottom half" with 50% chance

To see that the informational content of the signals in the *NoisyComparative* Treatment is equivalent to those in the *Noisy* treatment, note that:

$$\begin{aligned} Pr(s = G \mid t_i = H) &= Pr(t_{-j} = L \mid t_j = H) \times 1 + Pr(t_{-j} = H \mid t_j = H) \times 1/2 \\ &= 5/9 \times 1 + 4/9 \times 1/2 = 7/9 \end{aligned}$$

where s is the signal received (where G and B denote good and bad news), t is the performance type (where H and L denote being among the top half and the bottom half of performers), j is the index for the subject, and $-j$ is the index for the randomly chosen opponent. Similarly,

$$Pr(s = G \mid t_j = L) = Pr(t_{-j} = H \mid t_j = L) \times 0 + Pr(t_{-j} = L \mid t_j = L) \times 1/2 = 2/9$$

$$Pr(s = B \mid t_j = H) = Pr(t_{-j} = L \mid t_j = H) \times 0 + Pr(t_{-j} = H \mid t_j = H) \times 1/2 = 2/9$$

$$Pr(s = B \mid t_j = L) = Pr(t_{-j} = L \mid t_j = L) \times 1/2 + Pr(t_{-j} = H \mid t_j = L) \times 1 = 7/9$$

Hence, the likelihood ratios of receiving good and bad news in the *NoisyComparative* Treatment are equivalent to those calculated in Equation (1):

$$LR_G^{NC} = \frac{Pr(s = G \mid t = H)}{Pr(s = G \mid t = L)} = \frac{7}{2} \quad , \quad LR_B^{NC} = \frac{Pr(s = B \mid t = H)}{Pr(s = B \mid t = L)} = \frac{2}{7} \tag{4}$$

3.4 Posterior Belief Elicitation Stage

We elicited subjects' posterior beliefs about the likelihood of being among the top and bottom half of performers of their group after they observe their signal. We do not elicit beliefs over ranks in this stage, as we only need beliefs over being among the top half or the bottom half of performers to examine subjects' updating behavior.

As discussed above, either prior or posterior beliefs were randomly chosen for payment to prevent hedging motives. We incentivized posterior beliefs using the quadratic scoring rule (Selten, 1998) with the following formula:

$$100 - 50 \times \sum_{k \in \{top, bottom\}} (\mathbb{1}[half = k] - \frac{p_k}{100})^2$$

where $\mathbb{1}[half = k]$ is an indicator variable that takes the value 1 if the subject was among the k half of performers in the group and 0 otherwise, and p_k is the subject's posterior likelihood of being among half $k \in \{top, bottom\}$.

4 Results

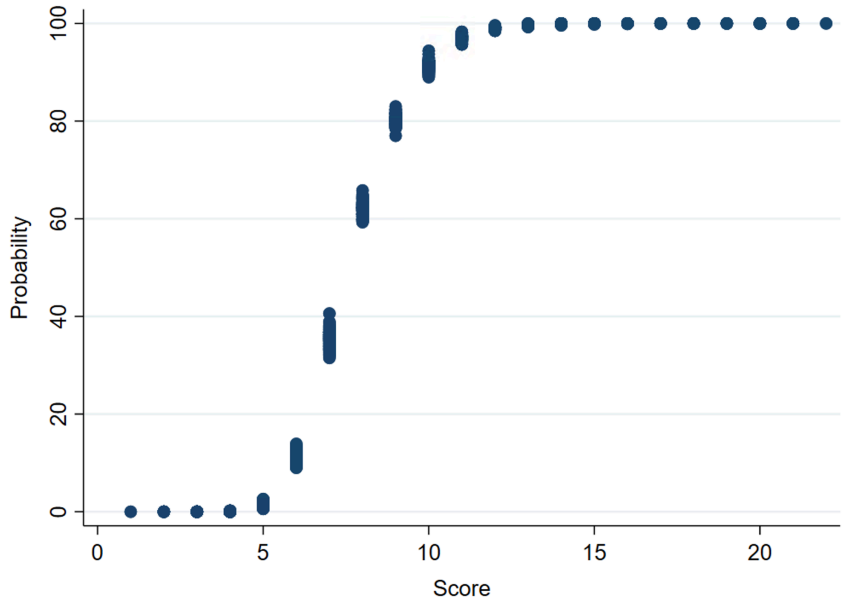
From the main data, we exclude some observations to minimize noise stemming from lack of comprehension or not paying attention to instructions. Similar to previous studies (e.g. Möbius et al., 2022, Barron, 2021, Coutts, 2019), we exclude subjects who reported posterior beliefs that were updated in the opposite direction compared to the Bayesian prediction (i.e., an upward shift in the belief of being among top performers after a bad signal or a downward shift in the belief of being among top performers after a good signal). These observations correspond to 9.9% of the subjects, which is in line with findings from previous studies. Secondly, given the online nature of the experiment, subjects who did not read the instructions require caution. We exclude subjects who spent less than 10 seconds both on the screen with instructions about the signal structure and on the comprehension question screen (in which instructions about the signal structure were

also accessible), resulting in the exclusion of 2.4% of the remaining subjects.[19] The proportion of subjects meeting these exclusion criteria is similar across feedback structures (Pearson $\chi^2 = 0.786$, $p > 0.10$ for updating incorrectly and Pearson $\chi^2 = 2.09$, $p > 0.10$ for time spent on instructions and the comprehension question). We take this as evidence that comprehension was similar across feedback structures.

4.1 Overview of Prior Beliefs and Confidence

As a preliminary analysis, we examine subjects' prior beliefs relative to their actual performance. As a belief accuracy benchmark, we generate each subject's bootstrapped probability of being among the top half of performers given their score. We run a simulation with 1,000 repetitions in which we randomly match each subject with 9 other participants and generate an indicator variable representing whether the subject was among the top half of performers. The bootstrapped probability of subject k being among the top half is $\sum_{r=1}^{1000} d_{r,k}/1000$, where $d_{r,k}$ is an indicator variable that takes the value 1 if subject k was among the top half performers of their group in the r^{th} replication of the simulation, and 0 otherwise. Figure 3 depicts the bootstrapped probability for each score. The bootstrapped probability is a cleaner benchmark for subjects' belief accuracy compared to simply using a dummy variable indicating whether they were among the top half of performers in their experimental group, as two subjects with the same score can have different realized outcomes due to luck.

Figure 3: Bootstrapped probability of being among the top half of performers given score



[19] We choose the 10 seconds cutoff in an ad-hoc manner, aiming for a lower bar on how fast the signal structure summary page can be read with comprehension. The main results are qualitatively similar if no subject is excluded based on time spent on instructions.

The mean absolute error in prior beliefs (calculated by taking the absolute value of the difference between the subject's prior belief and the bootstrapped probability of their being among the top half of performers) is equal to 35.5 points and is significantly different than 0 ($p < 0.001$) using a Wilcoxon signed-rank test.[20] In line with previous findings, we find that subjects do not have accurate beliefs about their relative performance on average.

We also find a significant gender difference in self-confidence, even though men and women perform similarly on the IQ test. On average, men and women answer 8.46 and 8.25 questions correctly, respectively. The difference is not statistically significant ($p = 0.427$). Figure 4 illustrates men and women's prior beliefs for each score, illustrating that women have lower priors for each possible score. Table 4 provides further evidence that women have significantly lower self-confidence, based on OLS regressions relating prior beliefs to gender and actual performance ($p < 0.001$).

[20] Unless otherwise stated, all p-values to compare distributions are obtained using the Mann Whitney U-test, while all p-values to compare measures to benchmarks are obtained using the Wilcoxon signed-rank test. Reported p-values come from the corresponding two-sided test.

Figure 4: Prior beliefs of being among the top half of performers by actual score

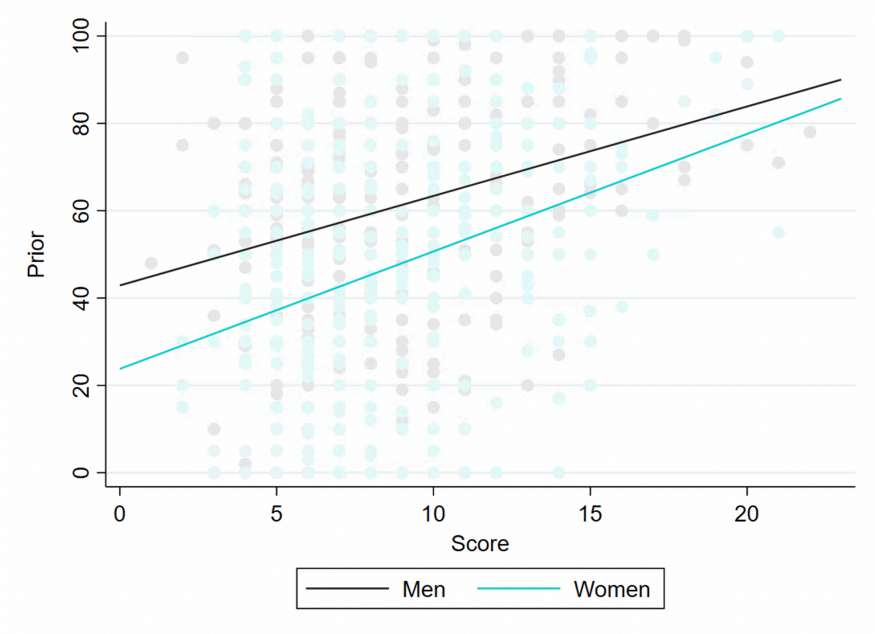


Table 4: OLS regressions relating prior beliefs to gender and performance

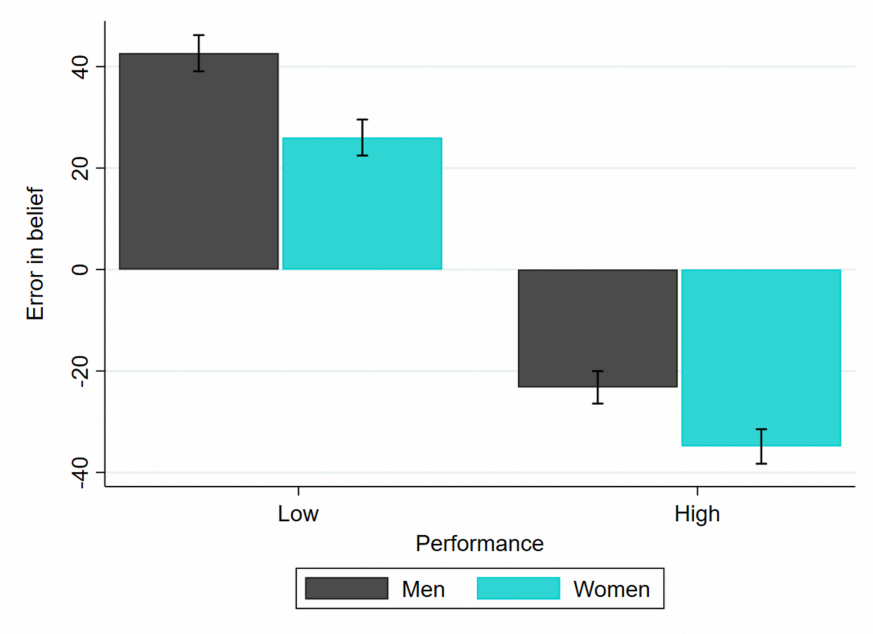
Prior	(1)	(2)	(3)
Female	-14.2***	-13.8***	-19.1***
	(2.020)	(1.923)	(4.737)
Score		2.4***	2.0***
		(0.259)	(0.357)
Female*Score			0.641
			(0.519)
Constant	60.2***	40.3***	42.9***
	(1.411)	(2.569)	(3.301)
N	783	783	783

Notes: Prior is the prior belief of being among the top half of performers. *Female* is a dummy variable equal to 1 if gender is female and 0 if male. *Score* is the number of questions answered correctly. *Female*Score* is the interaction of gender and score. Standard errors are in parentheses, * p<0.1, ** p<0.05, *** p<0.01.

Finally, we find significant gender differences in over and underconfidence. The mean error in prior beliefs, calculated as the difference between the prior belief and the bootstrapped probability of being among the top half of performers, is 9.1 points for men and -2.8 points for women. Note that positive values correspond to overconfidence and negative values correspond to underconfidence. The difference across genders is significant ($p < 0.001$). Since low performers are more likely to get bad news and high performers

are more likely to get good news, we also examine the gender difference in self-confidence by performance level. We classify low performers as those with less than a 50% probability of being among the top half and high performers as those with more than a 50% probability of being among the top half of performers. As Figure 5 illustrates, both men and women with low performance are overconfident, but men are more overconfident than women (the mean errors in prior beliefs are 42.6 vs 26.0, $p < 0.001$). Both men and women with high performance are underconfident, but women are more underconfident than men (the mean errors in prior beliefs are -23.2 vs -34.9, $p < 0.001$).

Figure 5: Mean error in prior beliefs by gender and performance



4.2 Belief Updating Behavior

We investigate belief updating behavior using the model originated by Grether (1980), which maintains the general Bayesian structure but allows for different weights on the prior, good news, or bad news compared to the Bayesian benchmark.[21] Consider the following likelihood ratio:

$$\frac{Pr(H|S)}{Pr(L|S)} = \left(\frac{Pr(H)}{Pr(L)}\right)^{\delta} \times \left(\frac{Pr(S|H)}{Pr(S|L)}\right)^{\beta} \tag{5}$$

where H *(High)* and L *(Low)* correspond to being among top and bottom half of performers among the reference group, $Pr(H|S)$ and $Pr(L|S)$ are posterior beliefs given signal $S \in \{Good, Bad\}$, and $Pr(H)$ and

[21] This model is also used by Möbius et al. (2022), Coutts (2019), Coffman et al. (2019), and Holt and Smith (2009).

$Pr(L)$ are prior beliefs of being among top half and bottom half of performers.[22] In the standard Bayesian model, $\delta = \beta = 1$. Adding indicator variables to distinguish between good and bad news, Equation 5 becomes:

$$\frac{Pr(H|S)}{Pr(L|S)} = \left(\frac{Pr(H)}{Pr(L)}\right)^{\delta} \times \left(\frac{Pr(S=G|H)}{Pr(S=G|L)}\right)^{\beta_G \times \mathbb{1}[S=G]} \times \left(\frac{Pr(S=B|H)}{Pr(S=B|L)}\right)^{\beta_B \times \mathbb{1}[S=B]} \tag{6}$$

where G and B denote receiving a signal with good news and bad news, respectively. Finally, log-linearizing Equation 6 allows us to test for behavioral biases on priors and signals using an OLS regression:

$$ln\left(posterior\right) = \delta \times ln\left(prior\right) + \beta_G \times \mathbb{1}[S=G] \times ln\left(LR_G\right) + \beta_B \times \mathbb{1}[S=B] \times ln\left(LR_B\right) \tag{7}$$

where $ln\left(posterior\right) = ln\left(Pr(H|S)/Pr(L|S)\right)$ is the posterior log-likelihood ratio for being among the top half given signal $S \in \{Good, Bad\}$. A positive value indicates allocating higher probability to being among the top half and a negative value indicates allocating higher probability to being among the bottom half. δ is the weight given to prior log-likelihood ratio for being among the top half, $ln\left(prior\right) = ln\left(Pr(H)/Pr(L)\right)$. LR_G and LR_B are the likelihood ratios of observing good and bad news, respectively.[23] $\mathbb{1}[S=G]$ and $\mathbb{1}[S=B]$ are indicator variables that are equal to 1 for the corresponding signal and 0 otherwise. β_G and β_B measure the responsiveness of the posterior to receiving good and bad news, respectively. Borrowing from the nice summary provided by Benjamin (2019) and Barron (2021) on interpreting the values of the δ, β_G, and β_B coefficients, Table 5 presents various belief updating biases documented in the literature.

[22] In all notation, we use type (H)igh and (L)ow to represent being among top and bottom halves instead of (T)op and (B)ottom. This is to avoid confusion with the notation of (G)ood and (B)ad signals.

[23] We show the calculations of LR_G and LR_B for each treatment in Subsection 3.3. See Equations (1), (2), (3), and (4).

Table 5: Interpretation of OLS Coefficients

Coefficient	Interpretation
$\delta = \beta_G = \beta_B = 1$	Bayesian updating
$\delta < 1$	Base-rate neglect
$\delta > 1$	Base-rate overuse
$\beta_G < 1$ or $\beta_B < 1$	Conservatism
$\beta_G > 1$ or $\beta_B > 1$	Overinference
$\beta_G \neq \beta_B$	Asymmetry

This Bayesian framework is silent with regard to prior and posterior beliefs at the boundary (i.e. beliefs equal to 0% or 100% for which the log-likelihood ratio is undefined). Following Charness and Dave (2017), Holt and Smith (2009) and Grether (1992), we truncate the data so that beliefs about being among the top and bottom performers lie in the interval [1%, 99%]. We replace posterior beliefs equal to 0% with 1% and those equal to 100% with 99%. For prior beliefs over ranks, we replace probabilities of 0% with 0.2% and subtract the total added probability from all non-zero probability ranks, weighted by the prior in the corresponding rank.[24]

$$p_i^* = \begin{cases} 0.2 & \text{if } p_i = 0 \\ p_i - \frac{p_i \times 0.2 \times n_0}{100} & \text{if } p_i \neq 0 \end{cases}$$

where p_i is the prior belief on rank $i \in \{1, 2, ..., 10\}$ before truncation, p_i^* is the same belief after truncation, and n_0 is the number of ranks with 0 prior belief. Truncated prior beliefs over being among the top half and bottom half of performers are the sum of the truncated prior beliefs over relevant ranks ($\sum_{i=1}^{5} p_i^*$ for top, $\sum_{i=6}^{10} p_i^*$ for bottom).[25]

4.2.1 Belief Updating Compared to the Bayesian Benchmark

Bayesian updating is consistent with all of the coefficients of Equation (7) being equal to 1. Behavioral models, such as ego utility or confirmation bias, predict alternative estimates of these coefficients, yet there is no existing model that predicts differential updating behavior across treatments based on the signal structure.

[24] The truncation of beliefs over being among the top half and bottom half of performers prevents the $ln(posterior)$ and $ln(prior)$ terms (in all treatments) and the truncation of beliefs over ranks prevents the LR_G and LR_B terms (in the *Comparative* Treatment) from blowing up in Equation (7).

[25] Our findings are robust to using alternative truncation methods.

Using the analysis described above, we compare updating behavior to the Bayesian benchmark across treatments. Table 6 reports the coefficients from estimating the OLS regression in Equation (7). The upper part of the table reports coefficients and their corresponding standard errors. A coefficient significantly different than 0 (as indicated by stars) indicates that prior beliefs and receiving good or bad news significantly affect posterior beliefs. As expected, all coefficients δ, β_G, and β_B are significantly different than 0 ($p < 0.001$). The bottom half of Table 6 compares estimated coefficients to the Bayesian benchmark. Any coefficient different than 1 is a deviation from Bayes' rule.

Table 6: Belief updating across treatments

	N	NC	C
Regressor	(1)	(2)	(3)
δ	0.666***	0.731***	0.718***
	(0.030)	(0.037)	(0.028)
β_G	0.937***	0.770***	0.725***
	(0.075)	(0.081)	(0.076)
β_B	0.878***	0.703***	0.815***
	(0.077)	(0.089)	(0.094)
p-values for H_0 :			
$\delta = 1$	0.000	0.000	0.000
$\beta_G = 1$	0.377	0.008	0.000
$\beta_B = 1$	0.126	0.001	0.094
$\beta_G = \beta_B$	0.580	0.570	0.499
N	261	255	267
R^2	0.752	0.718	0.745

Notes: Columns (1)-(3) report results from the OLS regression on *Noisy* Treatment, *NoisyComparative* Treatment, and *Comparative* Treatment, respectively. The first half of the table reports coefficient values and their associated standard errors below in parentheses with * p<0.1, ** p<0.05, *** p<0.01. The second half of the table reports p-values from Chow-tests on equality of coefficients to 1 or to each other.

Subjects exhibit base-rate neglect in all treatments ($p < 0.001$ for $H_0 : \delta = 1$) and do not update asymmetrically in any of the treatments ($p > 0.1$ for $H_0 : \beta_G = \beta_B$). The existence of conservatism varies by signal structure. There is no evidence of conservatism in the *Noisy* Treatment ($p = 0.377$ for $H_0 : \beta_G = 1$, $p = 0.126$ for $H_0 : \beta_B = 1$), while there is conservative updating of both good and bad news in the *NoisyComparative* Treatment ($p = 0.008$ for $H_0 : \beta_G = 1$, $p = 0.001$ for $H_0 : \beta_B = 1$), and conservative updating of only good news in the *Comparative* Treatment ($p < 0.001$ for $H_0 : \beta_G = 1$, $p = 0.094$ for $H_0 : \beta_B = 1$) at the 5% level. Even though the *Noisy* and *NoisyComparative* treatments are informationally isomorphic, adding a comparison component to the noisy signal results in different updating behavior across treatments.

Result 1 *Updating behavior is sensitive to the signal structure, even when the informational content of the two signals is equivalent. Subjects do not update conservatively in the Noisy Treatment but exhibit conservatism in the NoisyComparative and Comparative Treatments.*

4.2.2 Belief Updating Across Genders

Next, we examine whether there is a gender difference in how noisy and comparative signals are processed. Table 7 reports the results from estimating the OLS regressions based on Equation (7) separately for each gender and signal structure. The upper part of the table reports coefficients and their corresponding standard errors. Again, all coefficients are significantly different than 0, verifying that prior beliefs, good news, and bad news significantly affect posterior belief formation for both genders in all treatments. The bottom half of Table 7 reports coefficients compared to the Bayesian benchmark for men and women in each treatment. Any coefficient different than 1 is a deviation from Bayes' rule.

Table 7: Belief updating across treatments by gender

	N		NC		C	
	Men	Women	Men	Women	Men	Women
Regressor	(1)	(2)	(3)	(4)	(5)	(6)
δ	0.715***	0.589***	0.714***	0.725***	0.718***	0.703***
	(0.044)	(0.040)	(0.063)	(0.047)	(0.044)	(0.038)
β_G	0.919***	0.921***	0.814***	0.737***	0.778***	0.654***
	(0.104)	(0.106)	(0.128)	(0.104)	(0.116)	(0.102)
β_B	0.673***	1.107***	0.629***	0.775***	0.758***	0.908***
	(0.118)	(0.100)	(0.141)	(0.120)	(0.136)	(0.132)
p-values for H_0 :						
$\delta = 1$	0.000	0.000	0.004	0.000	0.000	0.000
$\beta_G = 1$	0.456	0.403	0.203	0.023	0.063	0.000
$\beta_B = 1$	0.004	0.320	0.004	0.053	0.119	0.544
$\beta_G = \beta_B$	0.139	0.194	0.366	0.803	0.922	0.172
N	133	128	133	122	135	132
R^2	0.745	0.783	0.663	0.777	0.729	0.771

Notes: Columns (1)-(2), (3)-(4), and (5)-(6) report results from the OLS regression using the data of men and women separately from *Noisy* Treatment, *NoisyComparative* Treatment, and *Comparative* Treatment, respectively. The first half of the table reports coefficient values and their associated standard errors below in parentheses with * p<0.1, ** p<0.05, *** p<0.01. The second half of the table reports p-values from Chow-tests on equality of coefficients to 1 or to each other.

We find that how women update their beliefs upon receiving good news and how men update their beliefs upon receiving bad news is sensitive to signal type. Relative to the Bayesian benchmark, men underweight bad news in both treatments in which the signal has a noise component ($p = 0.004, 0.004, 0.119$ for $H_0 : \beta_B = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively), whereas women underweight good news in both treatments in which the signal has a comparison component ($p = 0.403$, 0.023, 0.000 for $H_0 : \beta_G = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments). Furthermore, men do not significantly underweight good news in any treatment at the 5% level, ($p = 0.456, 0.203, 0.063$ for $H_0 : \beta_G = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments), while women do not underweight bad news in any treatment ($p = 0.320$, 0.053, 0.544 for $H_0 : \beta_B = 1$ in *Noisy*, *NoisyComparative*, and *Comparative* Treatments).

Result 2 *Noise and comparison components in a signal have a differential effect on belief updating by gender. Relative to the Bayesian benchmark, men underweight bad news if the signal has a noise component, whereas women underweight good news if the signal has a comparison component. Regardless of the signal structure, women do not underweight bad news and men do not underweight good news.*

4.2.3 Gender Differences in Posterior Beliefs

While the Grether (1980) framework presented above is the workhorse method of estimating departures from the Bayesian baseline, its application to gender differences in updating behavior is tricky. Because men and women differ systematically in their prior beliefs, comparing raw updating behavior across gender can conflate bias with baseline differences in expectations. In this section, we instead estimate posterior beliefs conditional on priors to isolate gender differences in responsiveness to feedback and to document which feedback mechanisms are effective at closing the gap in posterior beliefs. We examine the gender difference in posterior beliefs across treatments separately for recipients of good and bad news, running the following OLS regression of posterior beliefs on gender, priors, test score, and other individual characteristics:

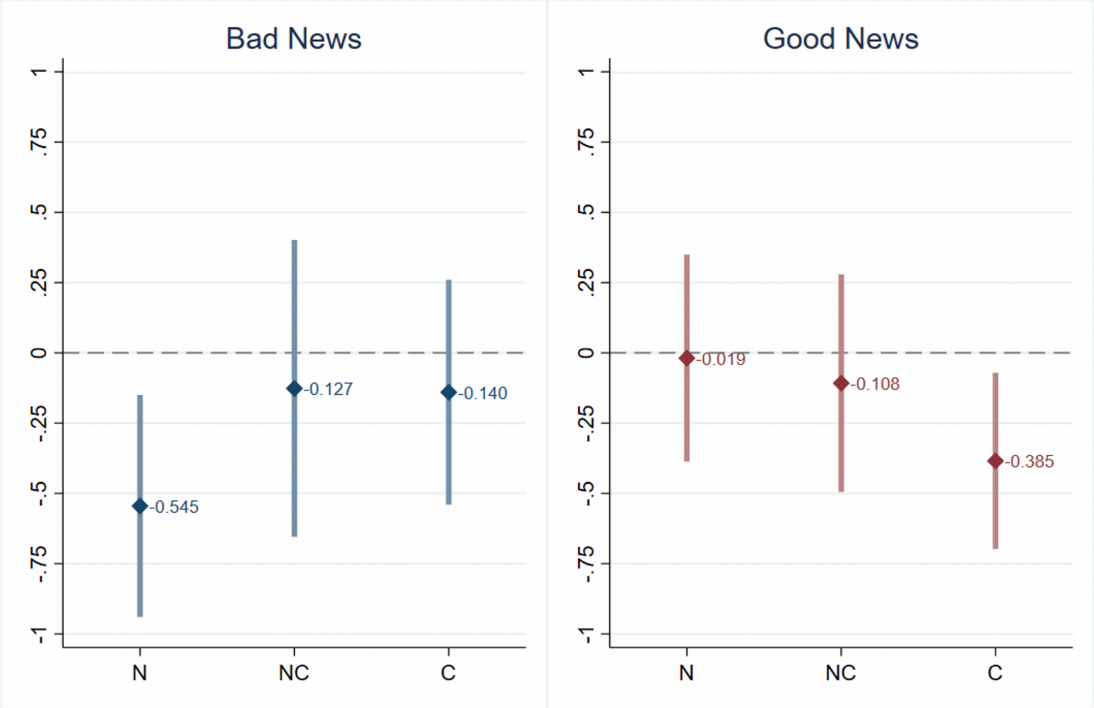
$$posterior_j = \beta_0 + \beta_F \times female_j + \beta_P \times prior_j + \beta_S \times score_j + \gamma \times C_j + \epsilon_j \tag{8}$$

where $posterior$ is the posterior log-likelihood ratio for being among the top half of performers, $female$ is a dummy variable equal to 1 if the gender is female and 0 if male, $prior$ is the prior log-likelihood ratio for being among the top half, $score$ is the number of correct answers in the IQ test, C is a vector of individual

characteristics including age, education, and income, and j denotes the subject index.[26]

Figure 6 plots the β_F coefficient in Equation (8), which captures the gender difference in posterior beliefs after feedback provision that is induced by feedback. In line with the conjectures given above, the gender gap in posterior beliefs for those who receive bad news is only significant in the *Noisy* Treatment, while the gender gap in posterior beliefs for those who receive good news is only significant in the *Comparative* Treatment. Women who receive bad news in the *Noisy* Treatment and women who receive good news in the *Comparative* Treatment have significantly lower posterior beliefs compared to men after controlling for score and prior beliefs (with $p\text{-}values$ 0.007 and 0.017, respectively). The complete list of coefficients and their corresponding standard errors are depicted in Table A.3.

Figure 6: Gender Gap in Posterior Beliefs Across Treatments



Notes: N, NC, and C correspond to *Noisy* Treatment, *NoisyComparative* Treatment, and *Comparative* Treatment, respectively. The figure illustrates β_F coefficient of Equation (8) with 95% confidence intervals.

The findings indicate that receiving bad news in the *Noisy* Treatment and good news in the *Comparative* Treatment result in a gender gap even after controlling for prior beliefs. Given that women have lower prior beliefs on being among the top half of performers compared to men, the documented gender

[26]The number of men and women who receive good news is balanced across treatments. Testing the equivalence of percent of men and women who receive good news with a test of proportions yields p-values 0.157, 0.987, and 0.572 for *Noisy*, *Noisy-Comparative*, and *Comparative* Treatments, respectively. See Table A.2 for the breakdown of number of subjects.

differences on posterior beliefs can be seen as a lower bound on the adverse effects of receiving noisy bad news or comparative good news. If prior beliefs are dropped from the set of controls, the gender differences in the *Noisy* Treatment under bad news and in the *Comparative* Treatment under good news become even more pronounced. Furthermore, the gender gap in posteriors for both types of news in the *NoisyComparative* Treatment becomes significant at the 5% level, also in line with predictions based on the belief updating patterns documented in Subsection 4.2.2. However, when the data is broken down by treatment, gender, and signal type, we are left with smaller sample sizes and there are some imbalances in prior beliefs across subgroups. Hence, we focus on the regressions conditioning on prior beliefs in the main body of the paper. The regression results not conditioning on priors can be found in Appendix Table A.4.

5 Conclusion

People are not great at forming accurate beliefs about their abilities, which leads to sub-optimal economically-relevant decisions. Giving performance feedback is one way to correct for misaligned beliefs, but there is no consensus on how individuals update their beliefs. In this paper, we show that the structure of feedback is an important factor affecting belief updating biases. The results of our controlled experiment show that the weights subjects give to good and bad news vary by whether the signal has a noise component or a comparison component. In previous work examining belief updating biases, noisy and comparative signals have been used in the absence of a comprehensive understanding of the behavioral responses to each.

A gender breakdown of misaligned beliefs shows that men have higher self-confidence than women with similar abilities, a result commonly found in previous studies. Furthermore, men and women react differently to signals with a noise or a comparison component. We find evidence that, relative to the Bayesian benchmark, men underweight bad news when receiving noisy signals, while women underweight good news when receiving comparative signals. Understanding how feedback mechanisms affect gender differences in belief updating can help us design more efficient policies to shrink the gender gap in self-confidence through feedback provision. For example, in settings where an average person is underconfident (i.e., more likely to receive good news), a policy maker may choose to provide feedback with a salient noisy component to avoid gender asymmetries in updating, as evidenced by our results that show that men and women react to good news symmetrically in the noisy treatment. In contrast, in settings where most people are overconfident, a policy maker may choose to provide feedback with a salient comparative component, where our results show that men and women update symmetrically in response to bad news. We caveat this discussion by noting that many feedback structures in the field include both a noisy and comparative component, so that more

research is necessary to determine optimal policy responses for field settings.

References

G. Azmat, M. Bagues, A. Cabrales, and N. Iriberri. What you don't know… can't hurt you? a natural field experiment on relative performance feedback in higher education. *Management Science*, 65(8):3714–3736, 2019.

B. M. Barber and T. Odean. Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics*, 116(1):261–292, 2001.

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

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

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

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

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

K. Coffman, M. Collis, and L. Kulkarni. *Stereotypes and belief updating*. Harvard Business School, 2019.

K. B. Coffman, P. U. Araya, and B. Zafar. A (dynamic) investigation of stereotypes, belief-updating, and behavior. 2021.

P. Cortés, J. Pan, L. Pilossoph, and B. Zafar. Gender differences in job search and the earnings gap: Evidence from business majors. 2021.

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

A. Coutts, L. Gerhards, and Z. Murad. What to blame? self-serving attribution bias with multi-dimensional uncertainty. 2020.

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

C. Drobner. Motivated beliefs and anticipation of uncertainty resolution. *American Economic Review: Insights*, 4(1):89–105, 2022.

C. Drobner and S. J. Goerg. Motivated belief updating and rationalization of information. *forthcoming, Management Science*, 2024.

W. Edwards. Conservatism in human information processing. *Formal representation of human judgment*,

1968.

D. Eil and J. M. Rao. The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38, 2011.

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

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

R. Fry. Women now outnumber men in the u.s. college-educated labor force. 2022.

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

D. M. Grether. Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1):31–57, 1992.

Z. Grossman and D. Owens. An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510–524, 2012.

C. A. Holt and A. M. Smith. An update on bayesian updating. *Journal of Economic Behavior & Organization*, 69(2):125–134, 2009.

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

B. Köszegi. Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4 (4):673–707, 2006.

A. Landier. Wishful thinking: A model of optimal reality denial. *Massachusetts Institute*, 2000.

G. Mayraz. Priors and desires: A model of payoff-dependent beliefs. *CEP Discussion Paper*, 1047, 2011.

M. M. Möbius, M. Niederle, P. Niehaus, and T. S. Rosenblat. Managing self-confidence: Theory and experimental evidence. *Management Science*, 2022.

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

M. Rabin and J. L. Schrag. First impressions matter: A model of confirmatory bias. *The quarterly journal of economics*, 114(1):37–82, 1999.

K. Rask and J. Tiefenthaler. The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review*, 27(6):676–687, 2008.

T. Rogers and A. Feller. Discouraged by peer excellence: Exposure to exemplary peer performance causes

quitting. *Psychological science*, 27(3):365–374, 2016.

P. Schwardmann and J. Van der Weele. Deception and self-deception. *Nature human behaviour*, 3(10): 1055–1061, 2019.

R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–61, 1998.

G. K. Shastry and O. Shurchkov. Reject or revise: Gender differences in persistence and publishing in economics. *Economic Inquiry*, 2024.

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

E. A. Shrider, M. Kollar, F. Chen, J. Semega, et al. Income and poverty in the united states: 2020. *US Census Bureau, Current Population Reports*, (P60-273), 2021.

O. Svenson. Are we all less risky and more skillful than our fellow drivers? *Acta psychologica*, 47(2): 143–148, 1981.

M. Thaler. Gender differences in motivated reasoning. *Journal of Economic Behavior & Organization*, 191: 501–518, 2021.

F. Zimmermann. The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–61, 2020.

Appendix A Additional Tables and Figures

Table A.1: Demographics Breakdown Across Treatments

	Treatment			p-values		
	N	NC	C	N-NC	N-C	NC-C
Gender						
Male	50.5%	50.3%	50.2%	0.97	0.93	0.97
Female	49.5%	49.7%	49.8%	0.97	0.93	0.97
Age	37.21	35.70	35.27	0.31	0.13	0.59
Education						
High School or Less	13.90	11.49	7.02	0.38	0.01	0.06
Associate Degree	9.15	11.15	10.37	0.42	0.62	0.76
Some College	23.39	23.65	27.76	0.94	0.22	0.25
Bachelor's Degree	37.29	37.16	39.80	0.97	0.53	0.51
Post Graduate Degree	16.27	16.55	15.05	0.93	0.68	0.62
Income						
Less than $20K	10.51	13.51	12.04	0.26	0.56	0.59
Between $20K and $30K	11.86	11.15	9.70	0.79	0.39	0.56
Between $30K and $50K	18.31	19.59	16.72	0.69	0.61	0.36
Between $50K and $70K	19.66	18.92	21.40	0.82	0.60	0.45
Between $70K and $150K	29.49	24.32	26.42	0.16	0.40	0.56
More than $150K	10.17	12.50	13.71	0.37	0.18	0.66
N	295	296	299			

Notes: The columns N, NC, and C correspond to *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively. The last three columns compare values across the associated treatment pairs and report p-values obtained by a test of proportions (for ratios) or by a Mann Whitney U-test (for the continuous variable age).

Table A.2: Number of Subjects in Each Treatment Broken Down by Gender and Signal Received

	Bad News			Good News		
	N	NC	C	N	NC	C
$Female$	68	54	59	60	68	73
$Male$	59	59	65	74	74	70

Notes: There is no significant difference in percent of males and females who receive good news in any treatment. p-values obtained by a test of proportions are 0.157, 0.987, and 0.572 for *Noisy*, *NoisyComparative*, and *Comparative* Treatments, respectively.

Table A.3: OLS Regressions Relating Posterior Beliefs on Gender and Individual Characteristics

	Bad News			Good News		
Coefficient	N	NC	C	N	NC	C
$Female$	-0.545***	-0.127	-0.140	-0.019	-0.108	-0.385**
	(0.199)	(0.266)	(0.202)	(0.186)	(0.196)	(0.159)
$Prior$	0.658***	0.760***	0.740***	0.629***	0.672***	0.602***
	(0.051)	(0.067)	(0.048)	(0.043)	(0.055)	(0.037)
$Score$	0.078**	0.004	0.062	-0.029	0.009	0.039
	(0.030)	(0.044)	(0.044)	(0.025)	(0.028)	(0.025)
$Education$	-0.117	0.121	-0.231	0.124	-0.072	-0.109
	(0.211)	(0.246)	(0.204)	(0.205)	(0.230)	(0.158)
$Income$	0.128	-0.078	0.089	-0.036	0.239	0.149
	(0.202)	(0.260)	(0.207)	(0.203)	(0.217)	(0.155)
Age	-0.007	-0.008	-0.002	-0.001	-0.016*	0.005
	(0.006)	(0.010)	(0.008)	(0.008)	(0.009)	(0.007)
$Constant$	-1.103***	-0.564	-0.983**	1.434***	1.431***	0.421
	(0.399)	(0.554)	(0.468)	(0.439)	(0.432)	(0.367)
N	127	113	124	134	142	143
R^2	0.704	0.617	0.712	0.676	0.626	0.730

Notes: Dependent variable is the is the posterior log-likelihood ratio for being among top-half performers, $Female$ is a dummy variable equal to 1 if the gender is female and 0 if male, $Prior$ is the prior log-likelihood ratio for being among top-half performers, $Score$ is the number of correct answers in the IQ test, $Education$ is a dummy variable equal to 1 if education is Bachelors' Degree or higher and 0 otherwise. $Income$ is a dummy variable equal to 1 if annual income is higher than $\$50k$ and 0 otherwise. Standard errors are reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table A.4: OLS Regressions Relating Posterior Beliefs on Gender and Individual Characteristics Without Controlling for Prior Beliefs

	Bad News			Good News		
Coefficient	N	NC	C	N	NC	C
$Female$	-1.174***	-1.241***	-0.656*	-0.370	-0.634**	-0.990***
	(0.297)	(0.368)	(0.344)	(0.299)	(0.276)	(0.262)
$Score$	0.211***	0.047	0.277***	0.102***	0.122***	0.158***
	(0.044)	(0.065)	(0.072)	(0.038)	(0.038)	(0.040)
$Education$	0.402	0.161	-0.140	0.680**	0.784**	0.267
	(0.317)	(0.366)	(0.353)	(0.327)	(0.317)	(0.265)
$Income$	-0.192	0.474	0.204	0.322	0.282	0.095
	(0.308)	(0.380)	(0.356)	(0.327)	(0.314)	(0.264)
Age	-0.009	-0.005	-0.002	0.006	-0.032**	0.007
	(0.010)	(0.014)	(0.013)	(0.013)	(0.013)	(0.012)
$Constant$	-1.721***	-0.668	-2.109***	-0.489	0.974	-0.674
	(0.609)	(0.824)	(0.797)	(0.679)	(0.622)	(0.614)
N	127	113	124	134	142	143
R^2	0.296	0.146	0.137	0.141	0.212	0.213

Notes: Dependent variable is the is the posterior log-likelihood ratio for being among top-half performers, $Female$ is a dummy variable equal to 1 if the gender is female and 0 if male, $Score$ is the number of correct answers in the IQ test, $Education$ is a dummy variable equal to 1 if education is Bachelors' Degree or higher and 0 otherwise. $Income$ is a dummy variable equal to 1 if annual income is higher than $\$50k$ and 0 otherwise. Standard errors are reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Appendix B Instructions

B.1 Welcome Page

Introduction

Welcome and thank you for participating in this study. This study consists of 2 paying sections and an exit questionnaire.

Final Earnings

The payoffs in this experiment are in terms of points with a conversion rate 1 USD = 100 points. You will be paid the sum of your earnings in two sections.

In order to receive the participation payment and the additional rewards, you have to answer every question. You will receive payment within the next 48 hours of completing the study.

Next

B.2 Part I, Introduction

Section I

In the next screen, you will be asked to solve several multiple choice questions. You will have 4 minutes to solve as many questions as possible. Your payoff in this section will be 20 points for each correctly answered question.

Next

B.3 Part II, Introduction

Section II

You completed the problem-solving part of the study. There were other Prolific participants who previously solved the exact same questions you answered in Section I. We randomly selected 9 of these participants. Together with these randomly selected participants, you now form a group of 10 participants.

We constructed a ranking of this group based on performance in the multiple choice questions in Section I. The group member that scored highest obtained rank 1. The group member with the second highest score obtained rank 2, etc... The group member with the worst performance obtained rank 10. If there was a tie between group members, the computer randomly decided who is ranked higher with equal chances.

Next, we would like to ask you about how you think you did on the questions you answered in Section I compared to others in your group. We will ask you 2 questions. In all questions, your expected earnings will be highest when you state your true beliefs. The computer will randomly select one of the 2 questions, and this question will be relevant for your earnings for Section II.

Next

B.4 Part II, Prior Belief Elicitation

Question 1 of 2

We are interested in how you think your performance (number of questions you answered correctly) in Section I compares to 9 other Prolific participants in your group. Specifically, we are interested in the following questions:

- How do you estimate the likelihood (in percent) of being in each rank when your performance is compared to the other 9 members of your group?
- What do you think is the likelihood (in percent) that you rank among the top and bottom halves of the performers in the group? In other words, in the group of 10, what do you think is the likelihood that your rank is 1, 2, 3, 4, or 5 (you are among the top half performers) and what do you think is the likelihood that your rank is 6, 7, 8, 9, or 10 (you are among the bottom half performers)?

To answer these questions, please fill out the table below. **Note that your answer to question on the right (likelihood of being among top/bottom performers) automatically updates based on your answers to question on the left (likelihood of being in each rank), so please make sure that you are content with your answers to both questions when submitting your answer.**

If this is the question that is selected for payment, you can earn up to 100 points from this question. Your expected payoff will be the highest if you report your true beliefs. (If you are interested in reading more about how the payoffs are calculated, click here .)

What do you think is the likelihood (in percent) that you rank in each of the following positions when your performance is compared to other members of the group?

Rank 1 (better than all other 9 participants):	10	percent
Rank 2 (better than 8, worse than 1 other participant):	20	percent
Rank 3 (better than 7, worse than 2 other participants):	20	percent
Rank 4 (better than 6, worse than 3 other participants):	15	percent
Rank 5 (better than 5, worse than 4 other participants):	10	percent
Rank 6 (better than 4, worse than 5 other participants):	10	percent
Rank 7 (better than 3, worse than 6 other participants):	10	percent
Rank 8 (better than 2, worse than 7 other participants):	5	percent
Rank 9 (better than 1, worse than 8 other participants):	0	percent
Rank 10 (worse than all other 9 participants):	0	percent

You can only enter whole numbers. The lowest possible number is 0 (percent). The highest possible number is 100 (percent). The sum of estimates should add up to 100.

What do you think is the likelihood (in percent) that you rank among the top and bottom halves of the performers in the group?

Likelihood of being among top 5 performers:	**75**	**percent**
Likelihood of being among bottom 5 performers:	**25**	**percent**

The sum of estimates should add up to 100. Currently, this sum is **100**.

Submit

B.4.1 Pop-up box upon clicking on ”click here”

If this is the question that is selected for payment, you can earn up to 100 points from this question. Your expected payoff will be the highest if you report your true beliefs. (If you are interested in reading more about how the payoffs are calculated, click here .)

You will be paid according to the following formula:

$$100 - 50 \times \sum_{i=1}^{10} (\mathbb{1}\{\text{rank=i}\} - \frac{p_i}{100})^2$$

where $\mathbb{1}\{\text{rank=i}\}$ is an indicator variable that takes the value 1 if your rank was equal to i and 0 otherwise, and p_i is your estimate for being in rank i (for each i in $\{1, 2, \ldots, 10\}$).

While this payoff formula may look complicated, what it means for you is simple: you get paid the most on average when you honestly report your best guesses of the probabilities for each rank (and so, your best guesses of the probabilities for being among top/bottom half performers of your group).

B.5 Part II, Prior Belief Confirmation

Please take a moment to verify the information below. Based on your answers to Question 1, the likelihood that you rank among the top and bottom halves of the performers in the group are:

Likelihood of being among top 5 performers:	**75**	**percent**
Likelihood of being among bottom 5 performers:	**25**	**percent**

Does this reflect your belief on being among the top and bottom halves of the performers in the group?

No, I want to edit my answer | Yes, confirm my answer

B.6 Part II, Signal Instructions

B.6.1 Noisy Signal

We would now like to provide you with some feedback on your performance. The computer is going to show you a signal about your relative performance in the group. The signal can be either "You are among the top half performers of your group", or "You are among the bottom half performers of your group".

We call this message a "signal" because it will not always show your true performance.

Here is how the signal is determined:
There is a 7/9 chance that the signal the computer shows you is your true relative performance. Imagine there are 9 balls, numbered 1-9 in a bag. The computer will draw one of those 9 balls out of the bag.

- If the computer draws a ball with a number between **1-7**, the computer will show you a signal that is your true performance.
  *This means that if you were among the **top half** performers of your group you will see "You are among the **top half** performers of your group" and if you were among the **bottom half** performers of your group you will see "You are among the **bottom half** performers of your group"*

- If the computer draws a ball with a number **8 or 9**, the computer will show you a signal that is the opposite of your true performance.
  *This means that if you were among the **top half** performers of your group you will see "You are among the **bottom half** performers of your group" and if you were among the **bottom half** performers of your group you will see "You are among the **top half** performers of your group"*

Next

Summary of how the signal is determined:
Each row shows what signals you may see for the corresponding event.

Your true performance	Signal you receive
Top half	"Top half" with 7/9 chance, "Bottom half" with 2/9 chance
Bottom half	"Bottom half" with 7/9 chance, "Top half" with 2/9 chance

Previous Next

B.6.2 Comparative Signal

We would now like to provide you with some feedback on your performance. The computer is going to show you a signal about your relative performance in the group.

Here is how the signal is determined:
The computer will randomly choose another participant among the other 9 members of your group. Each member has equal chance to be picked.

- If you performed better than the randomly chosen participant, you will see "You performed better than the randomly chosen participant from your group"
- If you performed worse than the randomly chosen participant, you will see "You performed worse than the randomly chosen participant from your group"

Next

Summary of how the signal is determined:
Each row shows what signals you may see for the corresponding event.

Randomly selected participant's rank	Your rank	Signal you receive
Between 1 and 10	Better than randomly selected participant's rank	Your performance is better than the other participant's
	Worse than randomly selected participant's rank	Your performance is worse than the other participant's

Previous Next

B.6.3 NoisyComparative Signal

We would now like to provide you with some feedback on your performance. The computer is going to show you a signal about your relative performance in the group. The signal can be either "You are among the top half performers of your group", or "You are among the bottom half performers of your group".

We call this message a "signal" because it will not always show your true performance.

Here is how the signal is determined:
The computer will randomly choose another participant among the other 9 members of your group. Each member has equal chance to be picked.

- Compared to the half that you are in in your group (top or bottom half), if the chosen participant is in the other half, the computer will show you a signal that is your true performance. *This means that:*
  *If you were among the **top half** performers of your group while the randomly chosen participant was among the **bottom half** performers, you will see "You are among the **top half** performers of your group".*
  *If you were among the **bottom half** performers of your group while the randomly chosen participant was among the **top half** performers, you will see "You are among the **bottom half** performers of your group"*

- If the chosen participant is in the same half that you are in in your group (top or bottom half), the computer will flip a coin to decide whether to send you a signal stating you are in the top or bottom half, which may or may not be your true performance. *This means that if both you and the randomly chosen participant were among the **top half** performers of your group or if both you and the randomly chosen participant were among the **bottom half** performers of your group, you are equally likely to see "You are among the **top half** performers of your group" or "You are among the **bottom half** performers of your group".*

Next

Summary of how the signal is determined:
Each row shows what signals you may see for the corresponding event.

Your true performance	Randomly selected participant's performance	Signal you receive
Top half	Bottom half	Top half
Bottom half	Top half	Bottom half
Top half	Top half	"Top half" with 50% chance, "Bottom half" with 50% chance
Bottom half	Bottom half	"Top half" with 50% chance, "Bottom half" with 50% chance

Previous Next

B.7 Part II, Comprehension Question

B.7.1 Noisy Treatment

Understanding question: which of the following statements is true about the signal you will see? (Click here to remember how the signal is determined.)

- ○ Your signal will always show your true relative performance.
- ○ The computer will tell you whether or not your signal is your true performance.
- ○ Your signal will be your true performance with 7/9 chance and it will be opposite of your true performance with 2/9 chance.

Submit

B.7.2 Comparative Treatment

Understanding question: which of the following statements is true about the signal you will see? (Click here to remember how the signal is determined.)

- ○ Your signal will tell you whether you are in top half or bottom half performers of your group.
- ○ Your signal will compare your rank with one of the 9 other members of your group and will tell you whether you performed better or worse than that participant.

Submit

B.7.3 NoisyComparative Treatment

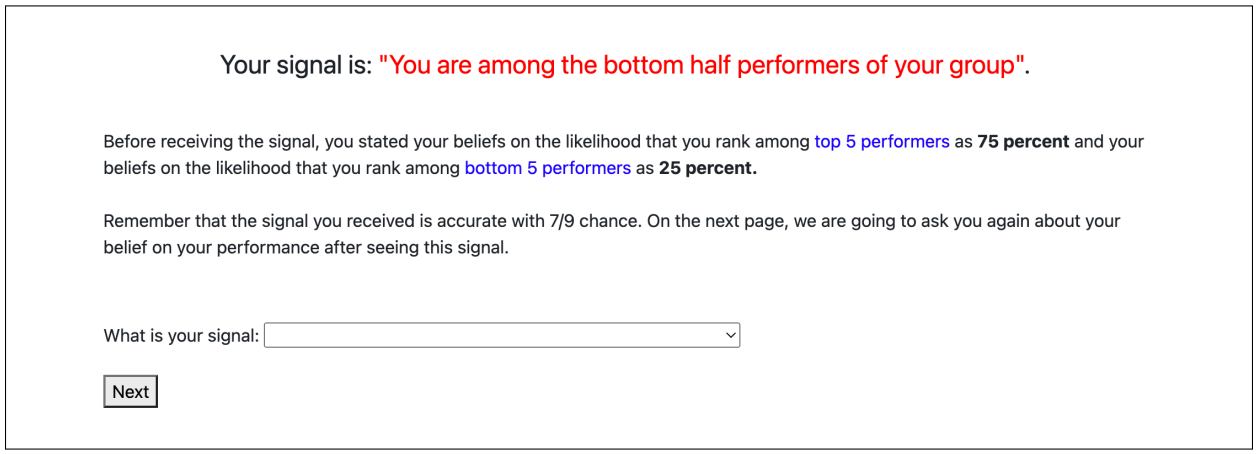
Understanding question: which of the following statements is true about the signal you will see? (Click here to remember how the signal is determined.)

- ○ Your signal will always show your true relative performance.
- ○ The computer will tell you whether or not your signal is your true performance.
- ○ When you and the randomly chosen participant are in different halves of the group ("top & bottom" or "bottom & top"), your signal will be your true performance. When you and the randomly chosen participant are in the same half of the group ("top & top" or "bottom & bottom"), there is a 50% chance that your signal will be your true performance and 50% chance that your signal will be opposite of your true performance.

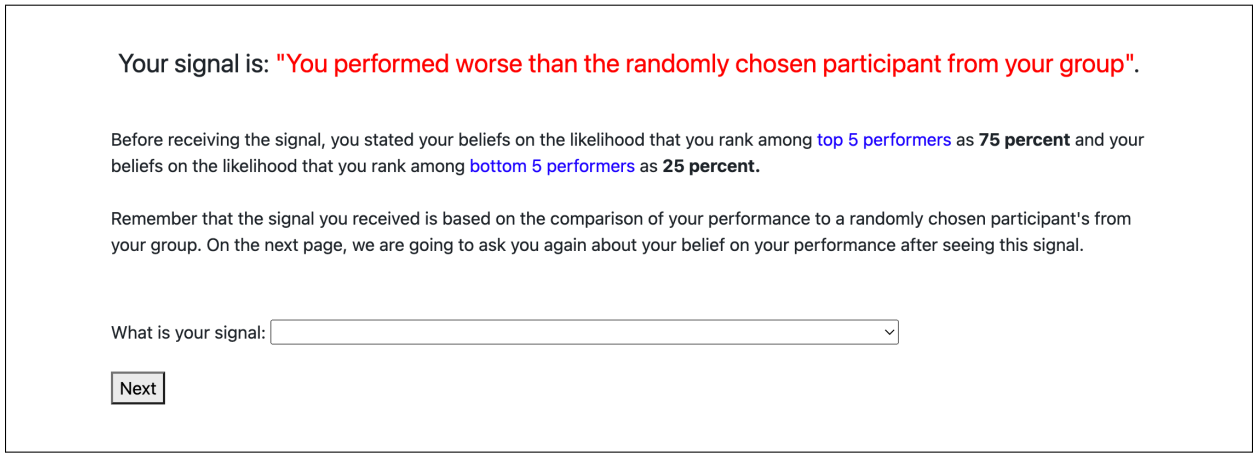
Submit

B.8 Part II, Feedback if Bad News

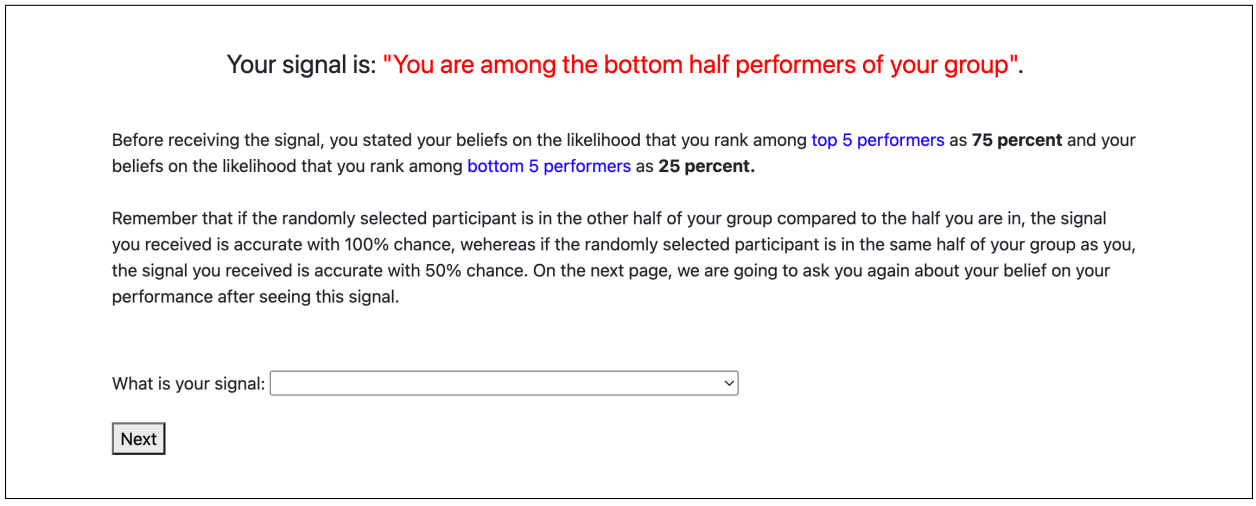
B.8.1 Noisy Treatment



B.8.2 Comparative Treatment

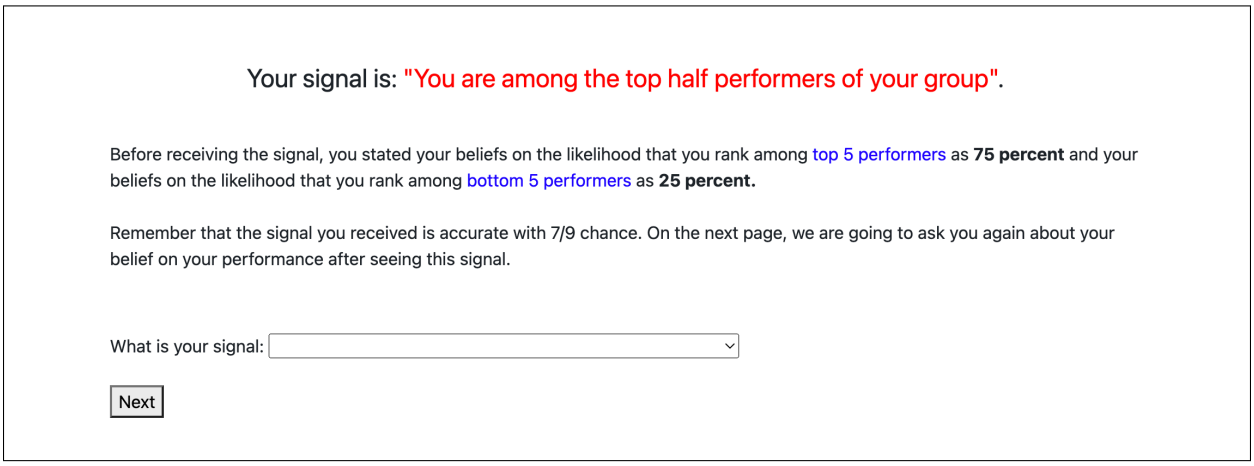


B.8.3 NoisyComparative Treatment

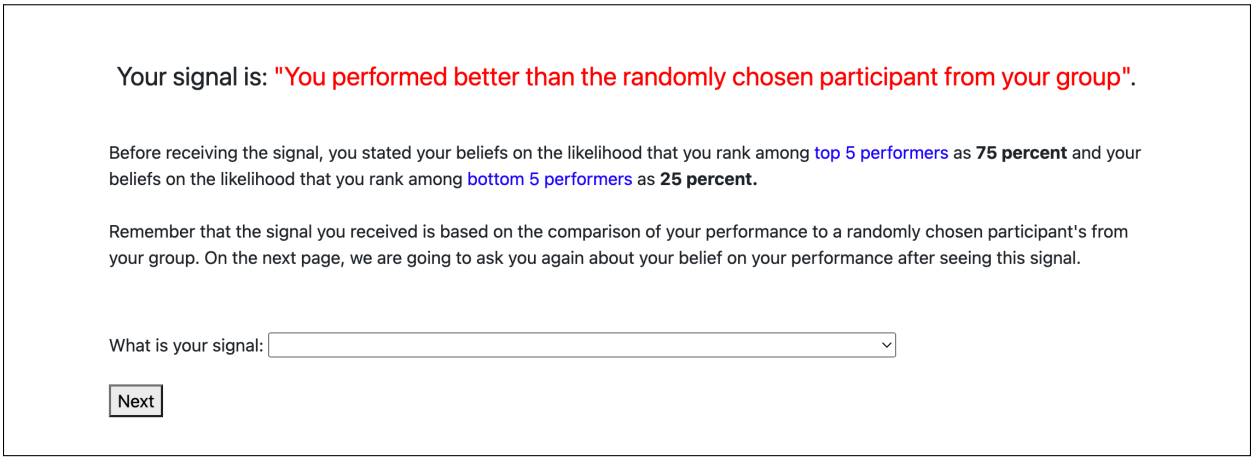


B.9 Part II, Feedback if Good News

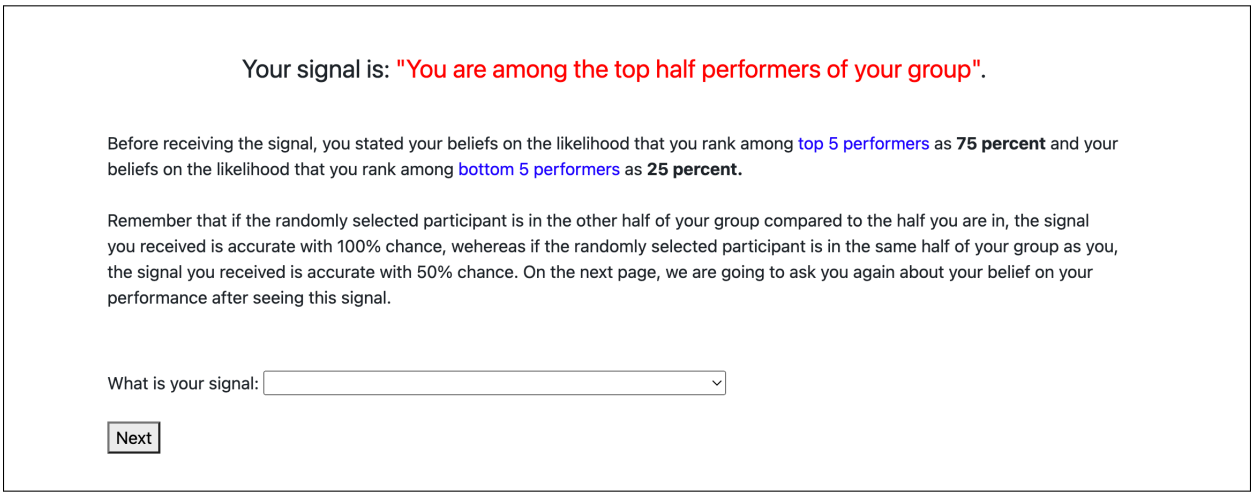
B.9.1 Noisy Treatment



B.9.2 Comparative Treatment



B.9.3 NoisyComparative Treatment



B.10 Part II, Posterior Beliefs

Question 2 of 2

Now that you received some feedback on your performance, what do you think is the likelihood (in percent) that you rank among the top half/bottom half performers in the group?

Likelihood of being among top 5 performers: [ ] percent
Likelihood of being among bottom 5 performers: [ ] percent

You can only enter whole numbers. The lowest possible number is 0 (percent). The highest possible number is 100 (percent). The sum of two estimates should add up to 100.

If this is the question that is selected for payment, you can earn up to 100 points from this question. Your expected payoff will be the highest if you report your true beliefs. (If you are interested in reading more about how the payoffs are calculated, click here .)

Submit

B.10.1 Pop-up box upon clicking on "click here"

If this is the question that is selected for payment, you can earn up to 100 points from this question. Your expected payoff will be the highest if you report your true beliefs. (If you are interested in reading more about how the payoffs are calculated, click here .)

You will be paid according to the following formula:

$$100 - 50 \times \sum_{k \in \{top, bottom\}} (\mathbb{1}\{\text{half}=\text{k}\} - \frac{p_k}{100})^2$$

where $\mathbb{1}\{\text{half}=\text{k}\}$ is an indicator variable that takes the value 1 if you you ranked among the k half performers in the group and 0 otherwise, and p_k is your estimate for being among k half performers for k in $\{top, bottom\}$.

While this payoff formula may look complicated, what it means for you is simple: you get paid the most on average when you honestly report your best guess of the probability for being among top/bottom half performers of your group.