Discrimination Without Reason:
Biases in Statistical Discrimination

Ignacio Esponda Ryan Oprea Sevgi Yuksel

January, 2022

Abstract

We report experimental evidence that people have difficulty effectively engaging in statistical discrimination, leading to lower accuracy gains from discriminating than a rational model would predict. As a result, discrimination can be significantly reduced without lowering accuracy, simply by improving the way people use information. We show that this inefficiency stems from subjects putting excess weight on their subjective judgements while simultaneously applying crude contrast-driven group-level biases. A series of treatment interventions give us insight into the psychological drivers of these errors and guidance on policies likely to be effective at removing them.

Esponda: Economics Department, University of California, Santa Barbara, Santa Barbara, CA, 93106, iesponda@ucsb.edu; Oprea: Economics Department, University of California, Santa Barbara, Santa Barbara, CA, 93106, roprea@gmail.com; Yuksel: Economics Department, University of California, Santa Barbara, Santa Barbara, CA, 93106, sevgi.yuksel@ucsb.edu.

1 Introduction

Models of statistical discrimination illustrate how information about group identity (e.g., race, gender, age, nationality) can be rationally used to form more accurate inferences when there is imperfect information about an individual (Phelps 1972, Arrow 1973). Consider, as an example, a manager evaluating the quality of a job candidate. The candidate is from one of two groups (advantaged vs. disadvantaged) which differ, on average, in candidate quality. Before forming an assessment, the manager studies the candidate individually (looking at the CV, conducting an interview, etc.), but such information is often limited. In such a case, the manager can use information on group identity to improve the accuracy of his assessment. But, as a consequence, candidates from the advantaged group systematically receive higher assessments than similarly qualified counterparts from the disadvantaged group.

In this sense, models of statistical discrimination descibe a theoretical tradeoff between *accuracy* and *discrimination*. On the one hand, inferences systematically differ regarding individuals from different groups who are otherwise identical when information on group identity is used. This discrimination is often viewed as negative, leading to unfair treatment of group members and perpetuation of historical inequities. On the other hand, using group identity increases the accuracy of such inferences. The bad news these models deliver is that (to the degree people negotiate this tradeoff in the optimal, Bayesian way described in such models) there is little room to improve on either one of these margins without making things worse on the other.[1]

Yet, there are behavioral reasons to wonder whether the accuracy-discrimination tradeoff is as "tight" as statistical discrimination models suggest. After all, statistical discrimination is a type of statistical reasoning and behavioral economists have collected substantial evidence over the past few decades that people are prone to serious mistakes when making such inferences. To the degree such errors carry over to statistical discrimination, there may be behavioral "free lunches" in negotiating the accuracy-discrimination tradeoff that don't arise in rational statistical discrimination, producing scope for cognitively-motivated policy interventions.

To study the accuracy-discrimination tradeoff through a "behavioral lens", we conduct a laboratory experiment designed to study how subjects combine individual and group level information to

[1]This discussion assumes the differences between groups (and the characteristics of the information environment) to be exogenous to the inference problem. When this is not true, rational statistical discrimination can generate further inefficiencies by distorting incentives to invest in activities (such as education) that (i) impact the underlying differences in group distributions; (ii) change informativeness of signals about individuals. See Lundberg & Startz (1983) for further discussion.

form inferences. Importantly, we deliberately designed this experiment to be completely abstract, cleanly removing other, non-inferential sources of discrimination (such as animus or taste-based discrimination). By studying a setting in which subjects assess the properties of fictitious members of artificial groups, we can identify purely perceptual and cognitive errors in statistical discrimination that would be impossible to isolate using framed experiments or naturally occurring data in which confounding affective or preference-based deviations from the model can also arise.

In our main, Baseline treatment subjects are asked to estimate the "value" (a number between 1 and 100) of an unspecified attribute of a fictitious member of one of two "groups" ("green" or "orange") that differ only in their mean value (40 or 60). Prior to making this assessment, the subject is (i) told which group the fictitious person is a member of (green or orange), and (ii) shown a number of dots equal to the fictitious person's true value for a split second. The short exposure to the dots means that subjects cannot perfectly observe the true value and therefore receive only a noisy, subjective signal.

This design reproduces what we take to be the crucial set of ingredients of many real-world settings in which the accuracy-discrimination tradeoff is of greatest public concern. The agent has to combine statistical information on group differences with imperfect subjective information about the individual, aggregating two distinct types of evidence into an assessment. By giving the subject a perceptual problem as a signal (rather than, e.g., a noisy piece of additional statistical information), we make it possible for a number of realistic errors to emerge that are plausibly important in relevant real-world settings, such as (i) policing, (ii) evaluation of intellectual arguments, (iii) hiring and promotion, (iv) evaluation of political candidates, (v) school admissions etc. in which prior statistical knowledge often must be combined with subjective evaluations to form assessments. This design choice thus allow us to reproduce what we take to be a key feature of this problem in the world: the possibility that statistical information interferes with perception, and that perception interferes with the formation of statistical posteriors.

Other details of the design are motivated by a structural model of statistical discrimination, built to identify behavioral determinants of decision making in this setting. The model points to three determinants of the accuracy-discrimination tradeoff that we designed the experiment to estimate. First, the model uses subjects' accuracy to estimate the precision of the subjective signal, allowing us to benchmark how a perfectly Bayesian decision would trade off accuracy and discrimination. Next, the model allows us to estimate the weight subjects put on group identity relative to the subjective signal, allowing us to look for evidence of classical errors in statistical reasoning like overprecision (i.e., excessive weight on the subjective signal). Finally, the model allows us to

estimate group-driven biases in assessment of the sort we might expect from perceptual distortions or irrational stereotyping.

We find that an accuracy-discrimination tradeoff exists, but is much weaker than predicted by theory. Subjects make significantly more discriminatory assessments when given group information (the "Baseline" treatment) than when not (the "NoGroup" treatment), as predicted by the standard model. But the accuracy gain of this additional discrimination is marginal at best, and it falls significantly short of what the Bayesian benchmark predicts. In particular, even when taking measured noise in perception as given, our subjects could dramatically improve the accuracy of their assessments without discriminating any more than they do. Alternatively, they could discriminate far less without making less accurate assessments.

Structural estimates show that this weak empirical tradeoff between accuracy and discrimination derives from two errors. First, subjects suffer from *overprecision*, meaning that they rely too much on their impressions from the perceptual signal and place less than optimal weight on group-level information. Second, subjects display *group bias*, meaning that they make biased use of group-level information, leading them to exaggerate group differences in a non-Bayesian way. Both mistakes reduce accuracy, but have opposite implications for discrimination. The net result is irrational statistical discrimination, and a potential avenue for costlessly reducing discrimination. Importantly, we find that these errors and the inefficiencies they give rise to are highly resistant to learning, persisting even after dozens of rounds of play with feedback on the mistakes subjects make.

Diagnostic treatments significantly narrow down the source of this irrationality and provide important clues as to how to remove it. In our "SignalFirst" treatment, we study the possibility that this inefficiency is driven in part by perceptual distortions, caused by foreknowledge of group identity. The treatment is identical to Baseline, but we show subjects the perceptual signal *before* revealing the group instead of after. Doing this causes a significant decrease in discrimination *at no cost to accuracy*. Structural estimates show that this works by reducing group bias, suggesting that this bias operates partly by distorting perception of the signal.

More dramatically, our "OneGroup" treatment shows that this group bias (and the inaccuracy it produces) occurs only when subjects have to assess members of two different groups. In this treatment, instead of alternating between assessing fictional people from different groups, we specialize subjects to evaluate only members of one group or the other (high-mean or low-mean) throughout the experiment. Removing scope for group contrasts causes subjects to make significantly more accurate assessments while simultaneously dramatically reducing discrimination. Structural esti-

mates show that this works by increasing the precision of the subjective signal and by eliminating the group bias altogether, suggesting that this bias is driven by irrational stereotypes that arise when subjects contrast one group with another.

These results are valuable because they identify purely cognitive and perceptual failures of rational statistical discrimination, and suggest a set of policy instruments to reduce them. Once again, by running the experiment in an abstract environment, we avoid confounds that might easily arise using naturally occurring data. In particular, had we run the experiment with frames of race, gender or age or had we estimated our statistical model in a real labor market, we might have inferred that the group bias we observe arose as consequence of other classical sources of discrimination such as taste-based discrimination. While we have no doubt animus and taste-based discrimination is important in the field, there is a value to isolating an additional force that has a purely cognitive and perceptual nature. The value lies especially in the very different policy remedies they open up for reducing discrimination with little cost to accuracy. We discuss some of these policy upsides in our concluding discussion.

Our paper contributes to an extensive literature studying discrimination, its potential causes, and policies intended to counteract it. We refer the reader to Charles & Guryan (2011), Bertrand & Duflo (2017), and Neumark (2018) for recent reviews. A key goal of this literature has been to identify the degree to which observed discrimination in different settings can be categorized as taste-based (Becker 1957) vs. belief-based or statistical in nature (Phelps 1972, Arrow 1973). Existing work has provided evidence for both channels. By using an abstract setting consisting of fictitious people from an "orange" or "green" group, we eliminate the scope for taste-based discrimination. By doing so, we can sidestep the issue of identification of these distinct channels and focus exclusively on statistical discrimination.

More specifically, our research builds on recent work in behavioral and experimental economics documenting the multiple ways in which behavior or beliefs diverge from the Bayesian benchmark in discrimination problems. One important point made by this literature is that what is often categorized as statistical discrimination might be irrational in the sense that it might be driven by *incorrect beliefs* about group differences. For example, Fershtman & Gneezy (2001) use trust and dictator games in the laboratory to document systematic mistrust in Israel Jewish society toward men of Eastern origin, due to mistaken ethnic stereotypes. Mobius & Rosenblat (2006) show that subjects wrongly believe that attractive people are more productive and that such belief differences translate into a wage beauty premium. Arnold, Dobbie & Yang (2018) conclude that racial bias in bail decisions is possibly driven by judges exaggerating the relative danger of releasing

black defendants. More recently, Bohren, Haggag, Imas & Pope (2019) outline the implication of incorrect beliefs for identifying the source of discrimination. They also document discrimination against Americans partly based on wrong stereotypes. There is also evidence that providing statistical information decreases or eliminates statistical discrimination driven by inaccurate beliefs (e.g., Reuben, Sapienza & Zingales (2014), Bohren, Haggag, Imas & Pope (2019)). While inaccurate prior beliefs are an important source of discrimination in some settings, we focus on other behavioral factors in how statistical assessments are made that could also play an important role in discriminatory decisions and have received little attention in the literature.

Our findings on biased inference also relate to a literature in psychology on confirmation bias (see Nickerson (1998) and Klayman (1995) for reviews). In our design, the signal providing individual-level information is based on a perception task and, thus, is purposefully subjective and ambiguous. Prior research suggests that, in such settings, people are prone to interpret evidence in a way that favors (or is consistent) with their initial beliefs. Namely, there is a tendency to 'see what one is looking for'.[2] The comparison between our Baseline and SignalFirst treatments provide evidence for this channel. We find inferences to be less biased in the latter treatment in which subjects' perception of the signal is untainted by knowledge of group identity.

Since our focus is on discrimination—namely, how people from different groups are differently evaluated—our results on biased inference also connect to recent work by Bordalo, Coffman, Gennaioli & Shleifer (2016) which models beliefs distortions based on stereotype formation. Stereotypes contain a "kernel of truth" as they are rooted in true differences between groups, but overweight the prevalence of 'representative' types in each population: for example, because there are more older people in Florida than California, one might incorrectly infer most Floridians are old and most Californians are young. Bordalo, Coffman, Gennaioli & Shleifer (2016) also provide experimental evidence in support of their model. In one of their experiments, subjects observe a screenshot with men and women with different shirt colors for a few seconds, and are later asked to recall what they observed. They show that recall is distorted in the direction of representative types. Our finding of a bias in the perception of the signal in the Baseline treatment but not in the OneGroup treatment is consistent with their finding and suggests that the act of contrasting different groups is partially

[2]Kelley (1950), Darley & Gross (1983), andLord, Ross & Lepper (1979) are prominent early examples in psychology providing evidence on how people's perceptions can be distorted by what they expect to see. More recent work on this issue include Enke (2020), Charness, Oprea & Yuksel (2021), and Oprea & Yuksel (2021)). This also connects to a literature that studies belief persistence in the presence of feedback. As Nickerson (1998) concludes "People often form an opinion early in the process and then evaluate subsequently acquired information in a way that is partial to that opinion".

responsible for the biased inference we observe in our experiment.

Our results on how subjects overweight individual-level subjective signals (relative to group-level statistical information) relate to the literature in psychology and economics on overprecision. Moore & Healy (2008) argue that overprecision is a type of overconfidence that is characterized by excessive certainty regarding the accuracy of one's beliefs and provide experimental evidence of this phenomenon. Overprecision has also been documented in the psychology literature as well as in the US Cellular and life insurance markets.[3,4] In a particularly relevant example from the discrimination literature, Mengel & Campos Mercade (2021) show in an experiment that subjects underweight objective statistical information relative to subjective information. Specifically, subjects are first asked to subjectively assess candidates and report a prior based on their characteristics (such as gender). This can be seen as analogous to the subjective signal in our environment. Subsequently, subjects are given objective, statistical signals about the candidates. The authors find that subjects update too little relative to their subjective prior. Both their finding and ours suggest that subjects put excessive weight on subjective information relative to objective, statistical information. While both in our experiment and the experiment by Mengel & Campos Mercade (2021) subjects behave as if the accuracy of their subjective signal were higher than what it is, in their experiment this results in conservatism (because the subjective signal provides group-level information) and in our experiment it results in base-rate neglect (because the subjective signal provides individual-level information).

Overprecision in our environment manifests itself as a form of base-rate neglect, since subjects behave as if they place excessive weight on the subjective signal and, therefore, too little weight on the prior, which here takes the form of statistical information about the two groups. Base-rate neglect is a bias that goes back to Kahneman & Tversky (1972) (see Benjamin, Bodoh-Creed & Rabin (2019) and Esponda, Vespa & Yuksel (2019) for recent work). We contribute to this literature by highlighting the implications of base-rate neglect in a design that features key characteristics of the settings where we worry about discrimination, specifically settings where people from different groups are contrasted. On this issue, the comparison between our Baseline

[3]In the psychology literature, Soll & Klayman (2004) find that 90% confidence intervals contain the correct answer less than 50% of the time. In the US Cellular market, Grubb (2009) and Grubb & Osborne (2015) argue that consumers in the US Cellular market underestimate the noise in their forecasts about future demand for calls, leading them to systematically choose the wrong calling plans and incur significant overcharges. In the market for life insurance, Gottlieb & Smetters (2021) argue that overoptimism about future liquidity shocks leads customers to underweight the risk of allowing policies to lapse.

[4]Mobius, Niederle, Niehaus & Rosenblat (2021) provide evidence of under-responding to statistical information in a context where subjects form beliefs about their own performance, and suggest this bias could be due to ego utility.

and OneGroup treatments is particularly informative. The latter, by focusing on a single group, resembles more closely the classic experiments on belief updating. While there is base-rate neglect in both treatments, we find subjects put more weight on group-level (relative to individual-level) information in the Baseline; this suggests that awareness of the contrast of different groups could have an impact on how much weight is put on group-level information in statistical inferences.[5]

While previous work either focuses on contrast effects or on the (over)weighting of subjective signals, our paper introduces an empirical model that includes both of these behavioral phenomena and quantifies their relative importance in explaining deviations from the Bayesian benchmark.

Finally, while we focus on environments where signals are exogenous, several experimental discrimination papers consider settings where the signal observed by the subject is endogenous. In one of the experiments considered by Mengel & Campos Mercade (2021), the signal is the applicant's education, and in equilibrium higher ability candidates are more likely to get educated. In Reuben, Sapienza & Zingales (2014), the signal is a decision to report on performance, and women are more likely than men to under-report performance. Both of these papers show evidence of naïveté, in the sense that subjects tend to under-infer from the signal. This bias is consistent with an experimental literature that shows that people have trouble making inferences from others' actions.[6]

The remainder of the paper is organized as follows. In Section 2 we describe our experimental design. In Section 3 we describe the theoretical setting and outline our estimation approach. In Section 4 we report our main experimental findings. Finally, in Section 5 we discuss implications of our results and suggest directions for future work.

2 Experimental Design

In our experiment subjects make assessments by combining (i) statistical information about group differences with (ii) imperfect subjective information about individuals. Our goal is to examine whether subjects combine these pieces of information in an efficient, Bayesian manner as predicted by models of statistical discrimination, and study implications of such behavior on the accuracy-

[5]This finding is consistent with earlier results from psychology. Koehler (1996) in the review of this literature concludes that subjects respond more to base-rates if they are varied within-subject (or alternatively if there is no variation in signal characteristics within).

[6]Eyster & Rabin (2005), Weizsäcker (2010), Araujo, Wang & Wilson (2021), Esponda & Vespa (2014), Barron, Huck & Jehiel (2019), Ngangoué & Weizsäcker (2021).

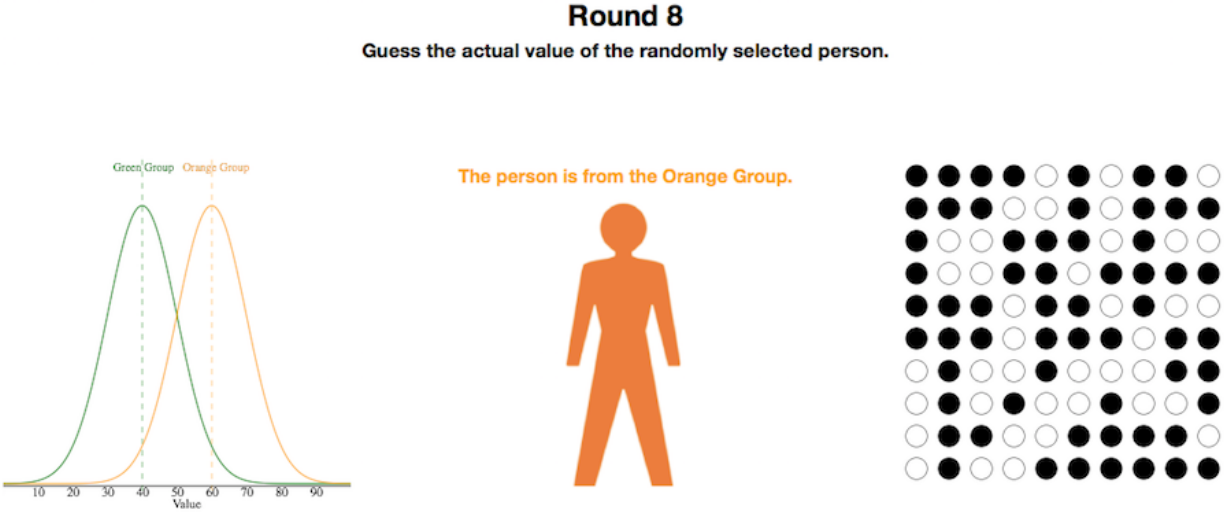


Figure 1: Screenshot from the Inference Task as Employed in Baseline. *Notes: The square grid showing the actual value of the person (number of black dots) disappears after 0.25 seconds.*

discrimination tradeoff. We designed the experiment around the model and empirical framework described in Section 3. In Section 2.1, we describe the inference task used in the experiment as employed in our Baseline treatment. In Section 2.2, we describe how this task is used in our sessions and how it varies across treatments. In Section 2.3, we explain how this design allows us to achieve the key goals of our paper. Finally, in Section 2.4, we describe details on the implementation of the experiment.

2.1 The Inference Task

The experiment consists of a series of 75 inference tasks. In each task, the computer first randomly selects one of two distributions (or "groups") that are approximately normally distributed and differ only in their mean.[7] The "high-mean group" (called the "Orange group" in the experiment) has a mean score of 60, the "low-mean group" (the "Green group") has a mean score of 40 and both have a standard deviation of 10. The computer then randomly draws the "value" of a fictitious "person" from the distribution of the selected group. The subject's task is to assess or guess the value of the person the computer has selected by typing a number between 0 and 100, and they are paid based on the accuracy of this assessment (as described in Section 2.4).

In our Baseline treatment, subjects are tasked with combining two types of information to make

[7] We discretize the distribution using integers between 0 and 100.

their assessment. Figure 1 shows a screenshot of the task.[8]

First, the subject is reminded of the two distributions on the left side of the screen. Then, the subject is given "Group Information," shown in the center of the screen. In the example, the subject has been told the person is from the Orange group.

Next, the subject is shown a perceptually noisy "Signal," on the right side of her screen. The signal is a grid of 100 dots (white and black) which flashes on the screen for 0.25 seconds. The number of black dots in this grid is the actual value of the person, but the subject does not have time to exactly count. Therefore, the signal is, in practice, noisy.

After seeing the grid flash on her screen, the subject is given a text box to input her guess. She is, afterwards, immediately shown the true value for the task and then clicks a button to move on to the next task, which will feature a new draw and possibly a different group.

2.2 Session and Treatment Design

The experiment employs a between-subjects design consisting of four treatments. One is the Baseline treatment, described above. The other three are variations on the same design, consisting of 75 independent inference tasks, the same set of distributions, the same grid signals displayed for the same amount of time (0.25 seconds) and the same incentives.

One of the treatments allows us to verify the comparative statics of statistical discrimination models:

> **NoGroup:** Like Baseline, but subjects make their assessments without ever receiving Group Information. Subjects observe the distributions shown on the left side of Figure 1 but are **never** provided the Group Information (the middle panel). Instead, they click a button to observe the signal (the right panel) for 0.25 seconds and enter their assessment.

The other two treatments provide the subject with the same information as the Baseline treatment (i.e., both group and signal information). However, they provide the information in two distinct ways and provide insight into how subjects perceive signals and integrate the two pieces of information we provide them:

> **SignalFirst:** Like Baseline, but subjects observe the signal **before** knowing which group the

[8] Please see instructions for our Baseline treatment in Appendix H on how we implemented these distributions in the laboratory and trained subjects in terms of their properties.

person belongs to rather than after. Subjects first click a button to observe the signal for 0.25 seconds. Three seconds later, Group Information appears on their screen and remains there for the remainder of the task.[9]

OneGroup: Like Baseline, but instead of randomly observing members of the low-mean and high-mean groups over the course the 75 tasks, subjects are assigned to always observe members of one group (high-mean or low-mean) over all 75 tasks. Subjects are only shown one of the two distributions on the left side of Figure 1 and are only informed about that group in the instructions.

2.3 Understanding the Design

We designed the experiment to serve several goals. We deliberately consider fictitious members of artificial groups in order to focus on purely perceptual and cognitive errors in statistical discrimination. Moreover, of all possible inference problems, we focus on one that is relevant to issues of discrimination, in which people make an assessment by combining (i) statistical information about group charactersitics with (ii) direct subjective information about the individual. In order to achieve this, we made two design choices.

First, we designed the experiment so that subjects have objective information about group differences. As discussed in the introduction, existing literature already highlights how incorrect beliefs about group differences can result in biased predictions. By deliberately focusing on abstract groups (green vs. orange) and by training subjects carefully about the statistical properties of these group distributions, our goal is to study inferences in a benchmark setting where beliefs about group characteristics are controlled.

Second, we designed the experiment so that information about the individual (the "signal") is *subjective* but fairly *costless* to acquire. Our aim is study how the perception of the signal is affected by knowledge of group identity. Our conjecture was that a visual signal that cannot be internalized fully (given the time it is shown on screen) can create sufficient uncertainty for us to study this. At the same time, we wanted to simplify the problem (e.g., abstracting from strategic information acquisition) by minimizing costs associated with seeing the signal. In our design, while the signal is difficult to see perfectly, the costs associated with assessing it are limited to the quarter second a subject spends looking at the screen.[10]

[9] In the Baseline, there was a similar three second delay after Group information was revealed before subjects could click to see the signal.

[10] By choosing the same variance for the green and orange groups, we also make sure that, even if subjects could

Some additional design choices were driven by the empirical and theoretical framework presented in the next section. This framework allows us to decompose the behavior of a subject into three components: a bias term, a relative weight on the subjective signal vs. group level information, and the precision of the signal. As we explain in the next section, one assumption in this framework is the (testable) requirement that the expected predicted value conditional on the true value be linear in the true value. The choice of a normally distributed prior distribution (as opposed to, say, a binary state of the world) is motivated by this requirement. Moreover, in order to be able to identify the behavioral parameters of the model at the individual level and to assess the extent to which experience affects inferences, we have each subject perform the prediction task a total of 75 times.

Finally, we designed the SignalFirst, NoGroup and OneGroup treatments such that their contrast to the Baseline would be informative about how the accuracy-discrimination tradeoff should be evaluated through a behavioral lens. Furthermore each of these treatments suggest a different policy approach to how we could design institutions in light of this tradeoff.

- The contrast between the Baseline and NoGroup treatments tests whether the theoretical tradeoff between accuracy and discrimination arises empirically. In particular, we test the extent to which accuracy decreases in the NoGroup treatment when discrimination is institutionally eliminated.

- The contrast between the Baseline and SignalFirst treatments allows us to study the degree to which behavioral biases in statistical inference are driven by biased perception of the subjective signal. The contrast allows us to study whether discrimination can be reduced by causing individual-level information to be revealed before group identity.

- The contrast between the Baseline and OneGroup treatments provide insights on whether behavioral biases in statistical inference are driven by contrast effects (possibly due to reasoning based on stereotypes), allowing us to explore whether specialization (to different groups) in assessments could decrease discrimination.

2.4 Implementation Details

We ran all treatments of the experiment simultaneously on Prolific in June 2021 with 241 subjects from the US (57 in Baseline, 61 in NoGroup, 62 in SignalFirst and 61 in OneGroup). The

choose how much effort to put in reading the signal, the incentives to do so would be the same for both groups.

experiment was conducted using software programmed by the authors in Javascript and deployed using Qualtrics. All subjects had to successfully answer comprehension questions in which they were tested on the properties of the prior distributions to begin the experiment. The experiment also included a risk measure adopted from the Caltech Cohort Study (Gillen et al. 2019) which was presented to the subjects at the end of the experiment. There were no time limits and on average the experiment lasted for 60 minutes.

All subjects received a base payment of $7.50. They also had the chance to win an additional $20 depending on the accuracy of their answers and up to $2.20 depending on their choice in the risk elicitation task. The percent chance of winning $20 was set to 100 minus the mean squared error of their assessments over the 75 rounds.[11] Earnings for subjects ranged from $7.50 to $29.60 with average earnings of $15.90.

3 Framework

In this section we develop an empirical framework for analyzing our data and a simple theoretical model to help us interpret the results economically and behaviorally. As in our experiment, there are two groups and each member of a group is characterized by some value v_g, where $g = l$ is the low-mean Orange group and $g = h$ is the high-mean Green group. We assume that $v_g \sim N(\mu_g, \sigma^2)$, so that values are distributed normally with mean μ_g for group g and an identical variance of σ^2 for both groups. In the experiment, $\mu_l = 40 < \mu_h = 60$ and $\sigma = 10$.

Again following our experiment, one of the two groups is first randomly selected, each with equal probability. A member of this group is then randomly chosen. The decision maker (a subject in our Baseline treatment) observes the group identity of the randomly drawn person. In addition, she observes a subjective signal about the value of this randomly chosen person (the 0.25 second signal in the experiment). Her task is to make a prediction $\tilde{v}_g$ about the value of this person in order to minimize expected squared error, $\mathbb{E}[(\tilde{v}_g - v_g)^2]$.

In the next subsection, we introduce our measures of accuracy and discrimination. These measures are well defined regardless of the empirical or behavioral model used to analyze the data. In Sections 3.2 and 3.3, we describe the empirical model and its behavioral interpretation. In Sections 3.4 and 3.5, we discuss the Bayesian benchmark and several other counterfactuals of

[11] Note that since the underlying distribution for each group has variance of 100, a subject who ignored the signal in the Baseline and only reported the group mean in every round would be expected to win the $20 with 0 percent chance.

interest. In Section 3.6, we summarize the value of the model.

3.1 Measures of accuracy and discrimination

The experimental data is given by a joint distribution of true and assessed values for each group, $(v_g, \tilde{v}_g)$. In analyzing this data, to highlight the accuracy-discrimination tradeoff that is central to the inference task we are studying, we'll be making use of the following measures.

Mean squared error, $MSE := \frac{1}{2}\mathbb{E}\left[(\tilde{v}_h - v_h)^2\right] + \frac{1}{2}\mathbb{E}\left[(\tilde{v}_l - v_l)^2\right]$. This measure computes the expected squared distance between assessments and actual values, taking into account that observations are equally likely be from the low-mean or high-mean groups. Given the incentive scheme adopted in the experiment, this also directly describes the monetary incentives of our subjects. We use this as our primary measure of (in)accuracy.

Group difference in assessments, $GD := \int \mathbb{E}\left[\tilde{v}_h - \tilde{v}_l \,|\, v_h = v_l = v\right] dF(v)$, where F is the mixture distribution of the low-mean and high-mean groups. The expectation inside the integral computes the expected difference in assessments between the high-mean and the low-mean group at the *same* actual value. For any person with value v from the low-mean group, this captures how much they are disadvantaged by their group identity. Similarly, for members of the high-mean group with value v, this captures their advantage. Then, we aggregate over all values of v using the unconditional distribution over values. We use this as our primary measure of discrimination.

While MSE serves as a natural measure of (in)accuracy in our setting, there is no equivalent obvious measure for discrimination. Nevertheless, by testing whether group difference (GD) is zero, we are testing whether individuals from different groups who are otherwise identical are treated the same. In this sense, group difference corresponds to the standard notion of discrimination used in economics; for instance, as captured by the concept of "equal pay for equal work" in labor economics (mandated by the Equal Pay Act of 1963). In Appendix A, we discuss other measures of discrimination and specifically relate GD to measures used in the machine learning literature.

We estimate both measures, MSE and GD, non-parametrically when comparing these measures across treatments. For MSE, the non-parametric estimate simply computes the sample average of the squared distance between assessments and actual values. For GD, the non-parametric estimate restricts attention to values of v for which we have at least 10 or more observations for each group in the data. We then estimate the difference in predicted values conditional on $v_h = v_l = v$, and finally aggregate for different values of v using the unconditional distribution of v.

3.2 Empirical strategy

We observe the group identity and both the true value v and the prediction $\tilde{v}$, but we do not observe the subjective signal of the decision maker. For simplicity, and unless necessary, we drop the group subscript when discussing the framework.

Suppose that the conditional expectation $\mathbb{E}[\tilde{v} \mid v]$ is linear in v.[12] This is a testable assumption (which we verify in our data) and it is easy to see that it implies[13]

$$\tilde{v} = B + \omega v + (1 - \omega)\mu + \varepsilon, \tag{2}$$

where:

- B is a bias term; more precisely, $B = \mathbb{E}[Bias(v)]$, where $Bias(v) \equiv \mathbb{E}[\tilde{v} \mid v] - v$;
- ω is the weight on the true value v vs. the mean μ;
- ε is an error term satisfying $\mathbb{E}[\varepsilon \mid v] = 0$.

Before we provide behavioral interpretations for each of these terms, it is convenient to represent our empirical strategy using a simple diagram. Figure 2 panel (a) illustrates the decisions of several hypothetical decision makers for some fixed population or group.[14] Each dot represents a pair of true (horizontal axis) and mean predicted values (vertical axis), and the solid fitted line corresponds to estimates of expression (2) using OLS.

First, the bias term B can be visualized by comparing the predicted value at $v = \mu$ with the value of μ. In this example, the predicted value is higher than μ, implying that there is a positive bias, $B > 0$. Second, the slope of the fitted line indicates the estimate of ω. We can compare this

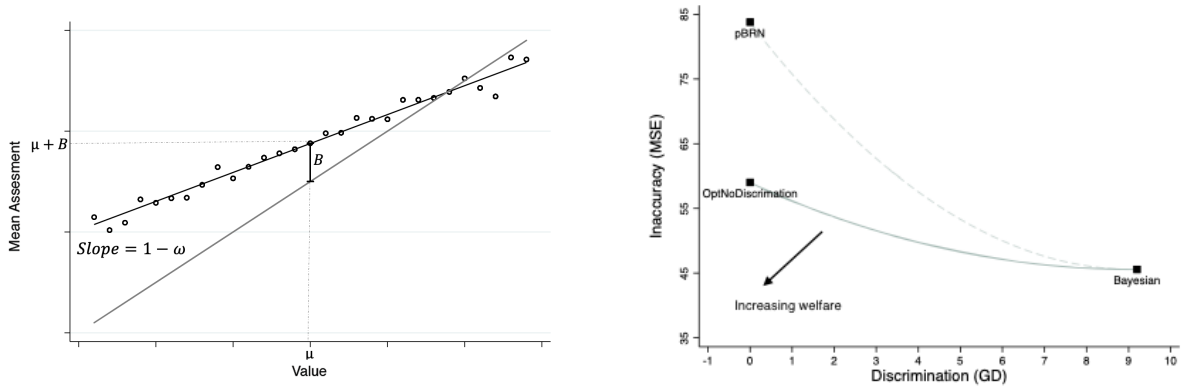
[12] This is true if and only if $(v, \tilde{v})$ are jointly normally distributed. This is why we chose the variable under our control, v, to be (approximately) normally distributed.

[13] Too see how equation (2) is derived from the linearity assumption $\mathbb{E}[\tilde{v} \mid v] = \alpha + \beta v$, note that this assumption implies that

$$\begin{aligned} \tilde{v} &= \mathbb{E}[\tilde{v} \mid v] + \varepsilon \\ &= \alpha + \beta v + \varepsilon, \end{aligned} \tag{1}$$

where $\mathbb{E}[\varepsilon \mid v] = 0$ by construction. Next, note that the bias term defined in the text, $B = \mathbb{E}[\alpha + (\beta - 1)v] = \alpha + (\beta - 1)\mu$, so that replacing $\alpha = B - (\beta - 1)\mu$ in (1) we obtain $\tilde{v} = B + \beta v + (1 - \beta)\mu + \varepsilon$. To get to equation (2), we define $\omega := \beta$.

[14] Alternatively, the data could represent the decisions of a single decision maker who is observed to make several predictions.



(a) (b)

Figure 2: Empirical Strategy *Notes: In Panel (a) dots depict mean assessment, solid black line is best linear fit, and gray line shows 45 degree line (see Section 3.2 and 3.3 for details and interpretation). In Panel (b) solid line is the accuracy-discrimination frontier, the dashed line represents loci of points computed at different weights on the subjective signal, pBRN depicts the perfect base-rate neglect counterfactual, OptNoDiscrimination depicts the counterfactual with highest accuracy (at zero discrimination), Bayesian refers to the counterfactual with highest accuracy (see Sections 3.5 and 3.4 for details).*

slope with the slope of the 45 degree line (gray line), which represents a situation where there is no bias and where decision makers put all the weight on the signal ($\omega = 1$), and therefore no weight on the prior mean. In the example, the slope is less than one and so the decision maker is putting some weight on group information. We will use this type of diagram to present and interpret our results and compare our treatments.

In the previous subsection, we described how our measures of discrimination, MSE and GD, would be non-parametrically estimated to compare across treatments. These measures can also be estimated parametrically using the empirical model. We will rely on these parametric estimates to compute counterfactuals. In particular, the empirical model implies that $MSE_g = B_g^2 + (1-\omega_g^2)\sigma^2 + Var(\varepsilon_g)$ and $GD = B_h - B_l + (1-\omega_h)\mu_h - (1-\omega_l)\mu_l + (\omega_h - \omega_l)\bar{\mu}$, where $\bar{\mu} := .5\mu_l + .5\mu_h = 50$ is the average of the group means. Focusing on $\omega_h = \omega_l = \omega$ as a special case (which is approximately true in all our treatments) is informative as it reveals the range of factors that contribute to our measure of discrimination. In this case, GD reduces to $B_h - B_l + (1-\omega)(\mu_h - \mu_l)$. The first component, $B_h - B_l$, captures the difference in bias between the two groups; the second component, $(1-\omega)(\mu_h - \mu_l)$, is due to the fact that agents put weight $1-\omega$ on the prior means.

3.3 Behavioral interpretation

In order to provide an economic interpretation for the parameters of our empirical model, we consider the following behavioral model.[15] For a given group, we assume that the decision maker observes a noisy, subjective signal

$$s = b^S + v + \varepsilon', \tag{4}$$

where b^S is a bias in the perception of the signal and $\varepsilon' \sim N(0, \xi^2)$ is an error term satisfying $\mathbb{E}[\varepsilon' \mid v] = 0$. In addition, we postulate that the decision maker submits a prediction that is a convex combination of the observed signal and the prior mean of the group, plus possibly a prediction bias term, b^P:

$$\tilde{v} = b^P + \omega s + (1 - \omega)\mu. \tag{5}$$

The linearity assumption is motivated by the fact that the optimal (i.e., Bayesian) prediction is linear as discussed in the next section. Together, equations (4) and (5) imply equation (2), where we can now interpret each term using our economic/behavioral model:

- the bias $B = b^P + \omega b^S$ is a combination of prediction bias and signal-perception bias;
- the weight ω represents the relative weight placed on the signal vs. the prior;
- the error term $\varepsilon = \omega\varepsilon'$ is a function of weight ω and the signal error ε.

In summary, a decision maker in the model is characterized by bias B (resulting from b^P and b^S), a weight ω on the signal vs. the prior, and the variance ξ^2 of her subjective signal. Note that bias B and weight ω can be estimated directly from equation (2). Given ω, the regression error in equation (2) gives a measure of the variance of the subjective signal, ξ^2.

3.4 Bayesian benchmark

A special case of our framework is that of a Bayesian agent who has no bias, i.e., $b^P = b^S = 0$. Since the agent seeks to minimize expected squared error conditional on her signal, $\mathbb{E}[(\tilde{v} - v)^2 \mid s]$,

[15] The standard approach (since Grether (1980)) to studying deviations from Bayesian updating uses the following framework:

$$\frac{p(v = v_1 \mid s)}{p(v = v_2 \mid s)} = \left(\frac{p(v_1)}{p(v_2)}\right)^{\alpha} \left(\frac{p(s \mid v = v_1)}{p(s \mid v = v_2)}\right)^{\beta}, \tag{3}$$

where optimal behavior requires $\alpha = \beta = 1$. If $s \sim \mathcal{N}(v, \xi^2)$, it can be shown that for any value of (α, β), assessments consistent with Equation 3 are given by $\tilde{v} = \omega s + (1 - \omega)\mu$, where $\omega = \frac{\beta\sigma^2}{\beta\sigma^2 + \alpha\xi^2}$. That is, while this framework can account for deviations from optimal weight ω, it doesn't allow for a bias term B.

the optimal prediction is $\tilde{v}^{Bay} = \mathbb{E}[v \mid s]$. In particular, the normality assumptions on both v and s imply that

$$\tilde{v}^{Bay} = \omega^{Bay} s + (1 - \omega^{Bay})\mu,$$

where

$$\omega^{Bay} = \frac{\sigma^2}{\xi^2 + \sigma^2} \tag{6}$$

is the optimal weight on the signal. In particular, the weight on the signal is decreasing in the variance associated with the signal error. By obtaining an estimate of the variance of the signal error, we will be able to compute this optimal weight for a Bayesian agent.

Finally, we will also apply this linear model to the case where the decision maker does not observe group identity but knows that the person is equally likely to be drawn from one of the two populations. In this case, we will take the mean μ above to be the average mean of the two groups. The Bayesian optimal prediction is no longer linear in this case, for the subtle reason that a signal provides information about the population from which a person is being drawn, and, therefore, the optimal weight on the signal vs. the prior depends on the signal itself. However, the optimal prediction (as shown in Appendix B) and the aggregate assessment strategy of our subjects in our data (as shown in Appendix D) are both approximately linear.

3.5 Accuracy-discrimination frontier

Consider a social planner whose objective is to minimize a weighted average of our discrimination and accuracy measures,

$$\kappa GD + (1 - \kappa)MSE. \tag{7}$$

The special case $\kappa = 0$ corresponds to the objective of the agent in our behavioral model (also to subjects' incentives in our experiment). In this case, the optimal prediction is a weighted average of the signal and the mean of the group, where the weight on the signal is given by the optimal Bayesian weight, ω^{Bay}, described in the previous subsection. In Figure 2 panel (b), we plot this Bayesian benchmark for the primitives of our model and a value of the variance of the error term, ξ^2, that is similar to the value we estimate in our data. The figure measures discrimination (GD) in the horizontal axis and (in)accuracy (MSE) in the vertical axis, and the arrow pointing towards the origin indicates the direction of increasing welfare. As mentioned earlier, the Bayesian benchmark results in relatively low (in)accuracy at the expense of treating people with the same value differently.

At the opposite extreme, $\kappa = 1$, the planner only cares about minimizing discrimination, and GD is minimized as long as people from both groups are treated equally. For example, the planner can achieve $GD = 0$ simply by setting $\omega = 1$ and placing all the weight on the signal. This is a case of perfect base-rate neglect, and we depict it as pBRN in Figure 2. In addition, the dashed line in Figure 2 represents the loci of points determined by all values of $\omega \in [\omega^{Bay}, 1]$ in the behavioral model. That is, this dashed line depicts a continuum of strategies ranging from the optimal Bayesian weight on the signal to full weight on the signal, i.e., perfect base-rate neglect.

The perfect base-rate neglect benchmark (with full weight on the signal and, therefore, zero weight on the prior mean), however, achieves zero discrimination at a greater expense than necessary in terms of accuracy. The planner could instead make more efficient assessments by also putting weight on the average mean, $\bar{\mu} = (\mu_h + \mu_l)/2$, still treating both groups equally. This case of maximum accuracy subject to the constraint of zero discrimination (with optimal weights on signal s and average mean $\bar{\mu}$) is depicted by the OptNoDiscrimination benchmark in Figure 2.

Starting from the OptNoDiscrimination benchmark, the planner could increase accuracy by replacing $\bar{\mu}$ with a slightly higher value for the high-mean group and with a slightly lower value for the low-mean group. This change will increase accuracy, though at the expense of treating the groups differently, thus increasing discrimination. This tradeoff between discrimination and accuracy faced by the planner is illustrated by the frontier (solid line) in Figure 2, where OptNoDiscrimination and the Bayesians benchmarks are two extreme points on this frontier. Since the planner also puts weight on $\bar{\mu}$ (in addition to μ_l and μ_h), this frontier falls strictly below the dashed line. In summary, the frontier represents a continuum of outcomes where it is not possible for a planner using linear strategies to increase accuracy (or decrease discrimination) without making things worse off in terms of the other measure.[16]

The frontier depends on the parameters of the problem (which we control in the experiment) and on the variance of the subjective signal (which we can estimate using our behavioral model). While decreases in signal variance move the frontier inward, deviations from Bayesian behavior (in terms of bias B or weight w) generate outcomes that are outside of the frontier.

[16]Formally, the frontier is obtained by minimizing (7) over all linear inference strategies that have the following structure:

$$\tilde{v} = \omega_g s + (1 - \omega_g) \left(\gamma(\frac{\mu_l + \mu_h}{2}) + (1 - \gamma)\mu_g \right),$$

where $\omega_g \in [0, 1]$ and $\gamma \in [0, 1]$. This linear restriction gives us tractability and is consistent with our focus on linear strategies in the behavioral model. Since the frontier depicts solutions subject to this linearity constraint, our results on how far outcomes are from the accuracy-discrimination frontier can be interpreted as presenting a lower bound.

3.6 Value of the model

We conclude this section by summarizing the main purposes of the behavioral model. First, it provides a behavioral interpretation of the relationship between true and predicted values. In particular, assessments are driven by three parameters: bias B, relative weight on signal vs. group information ω, and informativeness of the signal ξ^2. Second, the model provides a Bayesian benchmark that can be compared to the experimental results. Third, the behavioral model provides us with several counterfactuals of interest that can be described by an accuracy-discrimination frontier. Fourth, the model can be used to construct a simple algorithm to correct for errors in inference and improve outcomes in terms of accuracy and discrimination, as we show in Section 4.5.

4 Results

Our results are organized as follows: Section 4.1 provides a first look at our results by comparing treatments in terms of our inaccuracy (MSE) and discrimination (GD) measures. Section 4.2 applies the framework in Section 3 to estimate the parameters of interest and provides several counterfactuals. Section 4.3 uses these estimates to study more closely the differences between the Baseline and NoGroup treatments. Section 4.4 focuses on the SignalFirst and OneGroup treatments to identify the mechanisms driving the inefficiencies observed in the Baseline. Section 4.5 demonstrates how the model can be used to improve outcomes in terms of accuracy and discrimination. Sections 4.6 and 4.7 end by highlighting heterogeneity in individual-level data and learning effects.

Except where otherwise stated, all statistical tests reported in the text are based on linear regressions with errors clustered at the subject level whenever there are multiple observations per subject. Since our measure of discrimination is estimated first conditional on true value and then aggregated over the distribution of true values (as outlined in Section 3.1), we use bootstrapping to make statistical statements on GD. In pooled statistics (but not in individual-level analysis), we remove 10 (out of our 241) subjects who made extreme forecasting errors and were clearly inattentive or confused.[17] Doing this provides a more accurate portrait of the data's central tendencies, but does not change any of our qualitative conclusions (e.g. the ranking of MSE or GD by treatment).

[17]Specifically, we removed the 5% of subjects whose MSE was greater than 200 – the MSE a subject could achieve simply by choosing the unconditional mean of 50 every period, ignoring group- and individual-level information. It is virtually impossible to make such extreme errors unless inattentive or confused, and removing such subjects is typically necessary in experiments using online samples.

4.1 A first look at results

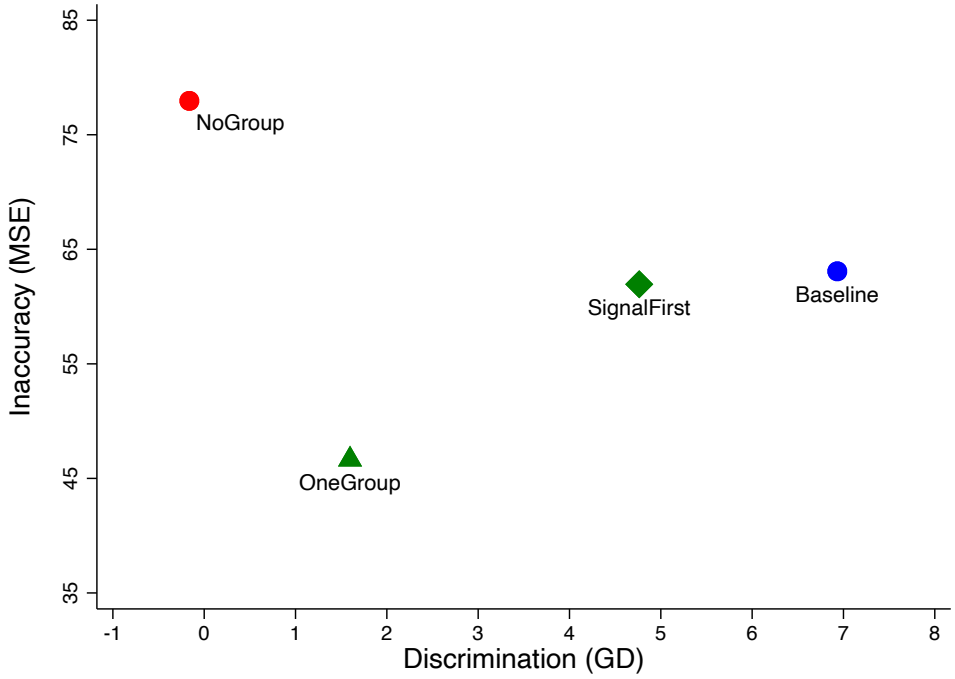


Figure 3: Mean Squared Error and Group Difference by Treatment

In Figure 3 we summarize our results by plotting the average mean squared error (MSE, y-axis) and the average group difference (GD, x-axis) for each treatment as four distinct points (leaving detailed statistical analysis to later sections). These raw cuts of the data, which rely neither on the empirical or behavioral models, effectively summarize our main findings.

First, the right-most location of the Baseline point shows that subjects engage in statistical discrimination and the fact that it is to the southeast of the NoGroup point shows that our subjects follow the comparative statics of statistical discrimination models. Giving subjects access to group information in Baseline allows them to improve their accuracy (reduce MSE) relative to the NoGroup treatment by discriminating on the basis of group (increasing GD).

Second, the accuracy-discrimination tradeoff subjects make in the Baseline treatment is inefficient. This is clear even in the raw data because, without changing any of the substantive details of the problem, we can improve on one or both dimensions of the problem. For instance, simply by compelling subjects to observe the signal prior to learning the group (the SignalFirst treatment), we can cause subjects to discriminate less at *no cost to accuracy*. This suggests that subjects in the Baseline over-discriminate to some degree due to misperception of the signal caused

by foreknowledge of the group.

Third, this inefficiency is greater when subjects are required to assess people from two groups. Simply by assigning different subjects to specialize in assessing different groups (instead of alternating between the two) in the OneGroup treatment, we significantly reduce *both* discrimination and inaccuracy.

The fact that subjects are inefficient in their statistical discrimination raises the possibility that "free lunches" are available in managing discrimination: simply by improving the perceptual and cognitive context of inference, we may be able to significantly reduce discrimination without any cost (and perhaps even with gains) to accuracy. In the remainder of this section, we make these observations more formally and report estimates of the structural model discussed in Section 3 in order to understand what inferential failures lie behind these patterns of inefficient statistical discrimination.

4.2 Overview of empirical estimates

In this section, we apply the framework in Section 3 to estimate the parameters of interest and provide several counterfactuals. We illustrate these estimates and counterfactuals in Table 1 and Figures 4 and 5. In subsequent sections, we discuss and interpret these estimates and counterfactuals.

Table 1: Model Estimates

	ω_l	B_l	ω_h	B_h	$\omega_h = \omega_l$	$B_h = B_l$	ξ_l^2	ξ_h^2	ω_l^{Bay}	ω_h^{Bay}	$\omega_l = \omega_l^{Bay}$	$\omega_h = \omega_h^{Bay}$
Baseline	0.82	-2.11	0.82	1.52		***	79	89	0.56	0.53	***	***
NoGroup	1.00	-0.49	1.02	-0.02			73	79	0.73	0.72	***	***
SignalFirst	0.81	-0.88	0.86	0.50	*	***	85	84	0.54	0.54	***	***
OneGroup	0.92	0.27	0.88	-0.41			53	60	0.65	0.62	***	***

Subscript l (h) denotes low-mean (high-mean) group.

Stars denote the confidence level with which the hypothesis associated with the column can be rejected.

*** 1%, ** 5%, * 10% significance.

First, Table 1 reports estimates of the three key parameters of the structural model described in Section 3.3–bias B, the weight on the subjective signal, ω, and the variance of the subjective signal, ξ^2–obtained by pooling all observations as if they were coming from the same individual. We include one specification for each treatment and each group (high-mean and low-mean, differentiated by

subscript $g \in \{l, m\}$). In Section 4.6, we show that the results also hold when these parameters are estimated at the individual level.

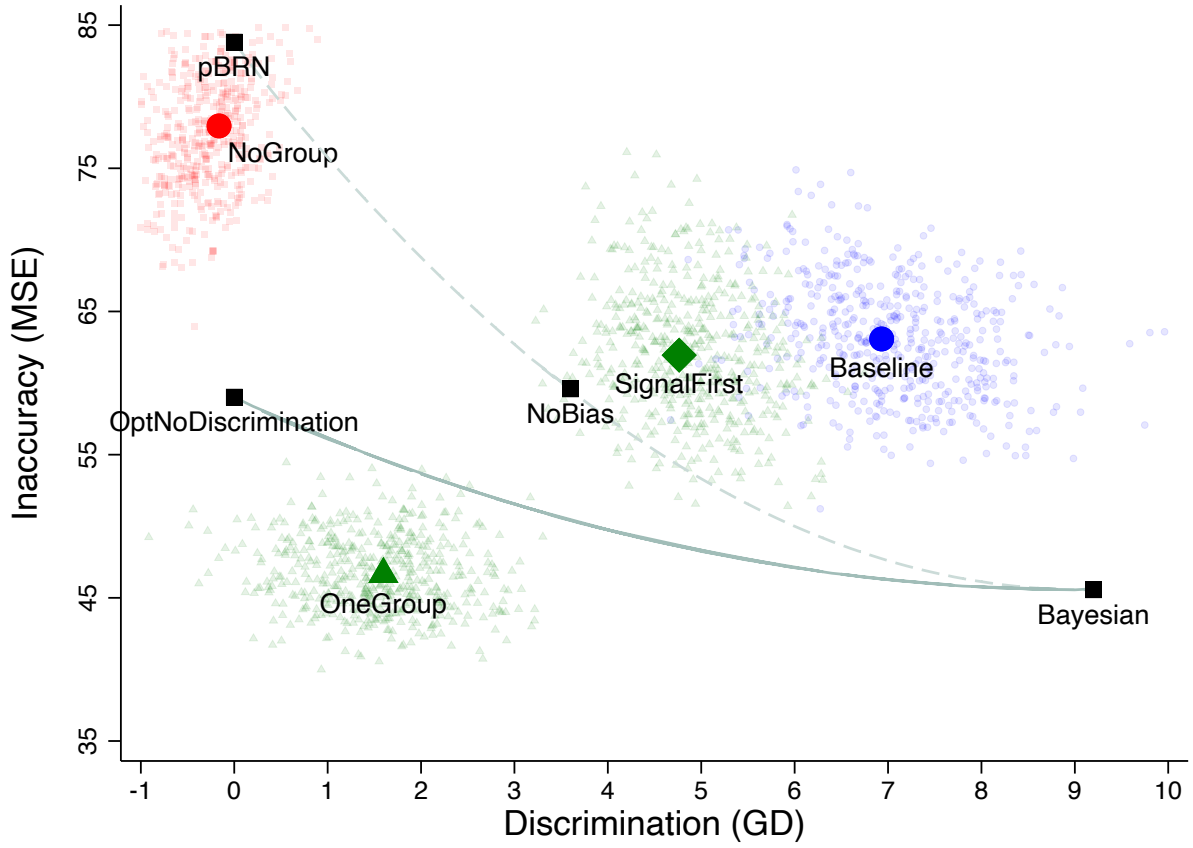


Figure 4: Mean Squared Error and Group Difference by Treatment *Notes: Solid line is the accuracy-discrimination frontier, the dashed line represents loci of points computed at different weights on the subjective signal, pBRN depicts the perfect base-rate neglect counterfactual, OptNoDiscrimination depicts the counterfactual with highest accuracy (at zero discrimination), Bayesian refers to the counterfactual with highest accuracy, and NoBias is the counterfactual where assessments are debiased but there is still base-rate neglect (see Sections 3.5 and 3.4 for details).*

Second, we use the estimated variance of the signal error for the Baseline treatment to construct several counterfactuals of interest (see Sections 3.4 and 3.5 for more details).[18] Figure 4 plots these counterfactuals together with the raw data points that were shown in Figure 3. The solid line depicts the accuracy-discrimination frontier and the dashed line is the loci of points determined by

[18]Recall that all of these theoretical benchmarks set biases $B_h = B_l = 0$ and depend only on the signal variance (in addition to the given primitives of the problem). Because we cannot reject the hypothesis that the variances of the high and low mean groups are the same, we use the average of the estimated variances in Table 1 for the Baseline treatment to construct these counterfactuals. In Appendix E we show that the results are very similar when using the median of all the individual-level variances (as opposed to the variance estimated from the pooled data).

all values of $\omega \in [\omega^{Bay}, 1]$ in the behavioral model. The figure also plots specific counterfactuals such as the Bayesian and OptNoDiscrimination benchmarks (both on the frontier) and the pBRN benchmark (corresponding to perfect base-rate neglect, $\omega = 1$). Finally, the figure also plots a counterfactual, NoBias, that sets the bias term B equal to zero but assumes people use the weights on signal vs. prior as estimated in the Baseline treatment. For a better comparison between the raw data and the theoretical counterfactuals, we also depict bootstrapped values for GD and MSE with lighter colors to visually illustrate the variation in our data.

Figure 5 provides an additional look at the data by plotting subjects' mean assessments against true value for each group and treatment. As explained in Section 3.2, this figure provides a quick way to assess the bias and the weight for each group and treatment. The figure reveals that assessments are approximately linear in value in every treatment and for every group. In Appendix D, we provide additional analysis in support of the linearity assumption.

4.3 Baseline Behavior: Accuracy-Discrimination Tradeoff

In this section, we use the estimates from the empirical model to take a closer look at the differences between the Baseline and NoGroup treatments, and, in particular, to compare the observed accuracy-discrimination tradeoff with several theoretical benchmarks.

We begin by confirming that subjects engage in statistical discrimination in our Baseline treatment and follow the comparative statics of statistical discrimination models. Recall that the only difference between Baseline and NoGroup is that, in the former, subjects are told the group from which the value is drawn. Comparing Baseline vs. NoGroup in Figure 3, we see that subjects discriminate significantly in Baseline by locating significantly to the *right* of NoGroup in the graph ($p < 0.01$). As a consequence, MSE in Baseline drops significantly relative to NoGroup ($p < 0.05$).[19] These results are also easy to discern in Figure 5: holding value constant, there is no difference in assessments by group in NoGroup, but a substantial difference in Baseline.

Result 1 Subjects engage in statistical discrimination and follow its comparative statics, improving accuracy by discriminating on the basis of group.

More importantly, Figure 4 also reveals our main result: although subjects engage in statistical

[19] If we control for elicited risk measure, the difference between the treatments is 11 ($p < 0.10$). If we also include the three extreme subjects from each treatment with MSE >200, the aggregate MSE difference between the NoGroup and Baseline treatments reduce to 7 (which is no more statistically significant). Detailed results are reported in Appendix C).

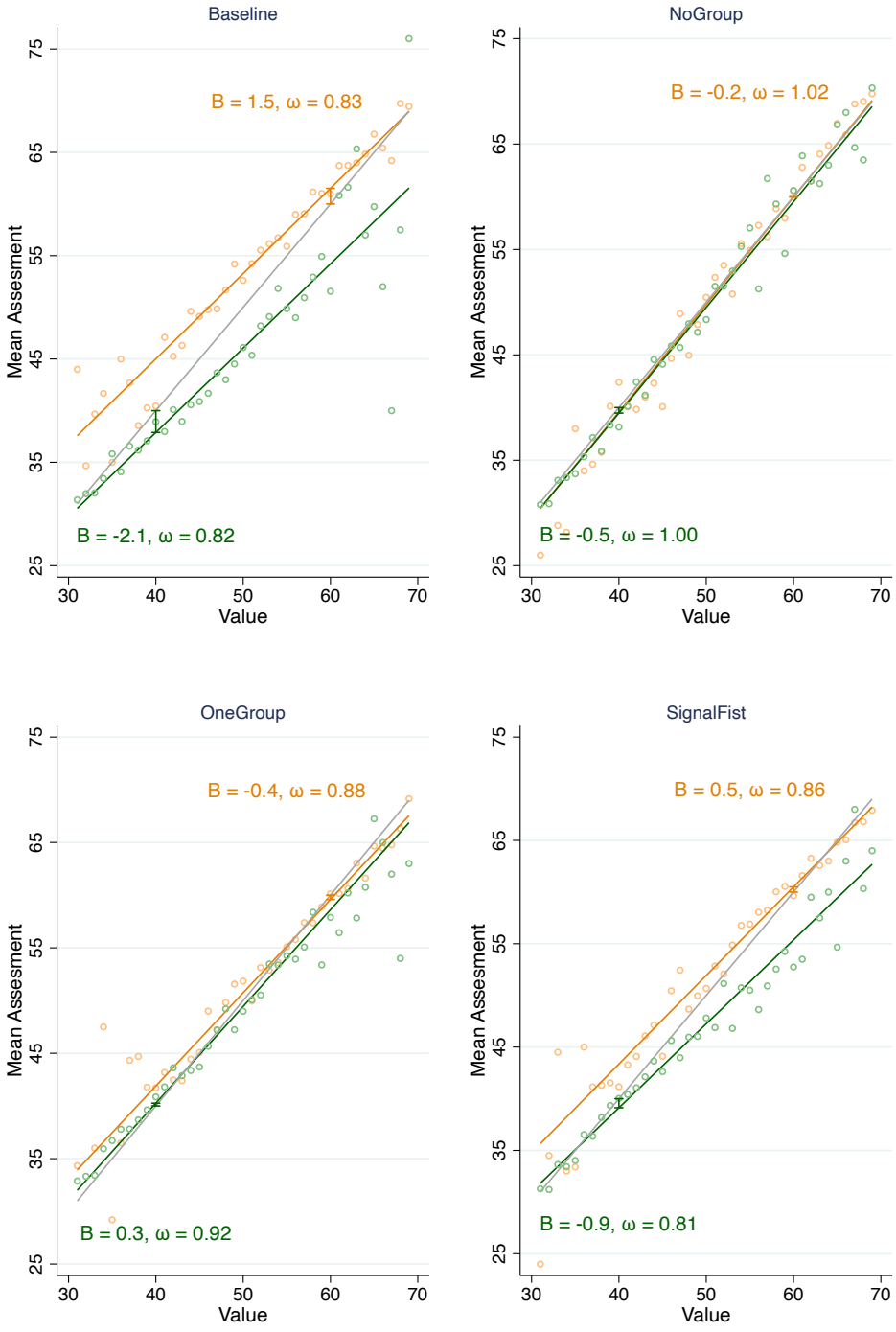


Figure 5: Average Assessment by Value in Each Treatment *Notes: See Section 3.2 for details on interpretation of the Bias and slope. Green (Orange) dots are for low-mean (high-mean) group. Green and Orange lines depict best linear fit by group and treatment; gray line depicts 45 degree line.*

discrimination in Baseline, they do so in an inefficient way, both relative to the Bayesian benchmark (which is the right benchmark for our subjects whose purpose is to minimize MSE) and relative to the accuracy-discrimination frontier. The Baseline point is to the northwest of the Bayesian benchmark, meaning subjects (i) discriminate too little ($p < 0.05$) and (ii) make inefficiently inaccurate assessments ($p < 0.001$). Given our estimates of subjects' variance in discerning the signal, they could make more accurate assessments, discriminate less, or both if they made better use of their information. However, the latter pattern, (ii), is not a necessary consequence of the former, (i). The Baseline point is also significantly higher than the accuracy-discrimination frontier, meaning, simply by correcting statistical errors, subjects could dramatically improve accuracy *without increasing discrimination at all*. Or alternatively subjects could maintain their observed level of accuracy but with *zero discrimination*, simply by making better use of their information.

Result 2 Subjects inefficiently engage in statistical discrimination in the Baseline treatment, relative to the Bayesian benchmark. Importantly, by eliminating mistakes in statistical inference it would be possible to reduce discrimination without reducing accuracy.

Our structural model was designed not only to contextualize the accuracy-discrimination tradeoff, but also to diagnose the source of errors. As discussed in Section 3.3, the model isolates two distinct sources of error. First, ω measures the weight subjects put on the signal relative to the prior: when $\omega = 0$ she puts no weight on her signal and all the weight on the prior, while when $\omega = 1$ the subject puts all the weight on the signal, entirely ignoring the prior. The optimal weight depends on the subjective signal's variance, which we estimate and use to calculate the weight a Bayesian would place on the signal, ω^{Bay} (see equation (6)). Second, B measures the systematic bias in evaluating a specific group. In a rational model (e.g. in a standard statistical discrimination model), $B = 0$. We present estimates of (ω_g, B_g) separately for low-mean and high-mean groups, $g \in \{l, h\}$, and for each treatment in Table 1.

The results reveal several important patterns for the Baseline treatment. First, subjects show significant evidence of *overprecision*, putting similar but in each case excessive weight on the signal relative to the prior. This means that subjects act as if their precision of the subjective signal were higher than what we actually measure it to be. Comparing ω_g to ω_g^{Bay}, we find that the former is larger than the latter for both the low-mean and high-mean groups, indicating that subjects put more weight than is optimal on the signal. As the last two columns of Table 1 show, this deviation from Bayesian benchmarks is highly statistically significant.

Second, subjects in Baseline display significant bias in both groups that go in opposite directions:

subjects systematically underestimate members of the low-mean group and overestimate members of the high-mean group, leading to a net group bias of $B_h - B_l = 3.6$. Both of these biases (and their difference) are highly statistically significant ($p < 0.001$).

These two forces produce the inefficiency in the Baseline treatment. Overprecision tends to lower discrimination by causing agents to put too little weight on group information, while group bias pushes behavior in the opposite direction.

Figure 4 illustrates the effect of each of these two forces, group bias and overprecision, by plotting the counterfactual NoBias benchmark, which represents the case where the weight remains as in the Baseline but the bias is eliminated in both groups. In particular, while eliminating the group bias results in no significant change to accuracy, it actually decreases discrimination by half. Also, the fact that the NoBias benchmark lies roughly halfway in between the Bayesian and pBRN benchmarks illustrates a significant degree of overprecision, as described above. Correcting overprecision (by changing weight ω to optimal ω^{Bay}) would result in the Bayesian benchmark, yielding an additional 25% increase in accuracy at the expense of more than doubling discrimination.

Result 3 Inefficiency in the Baseline treatment relative to the accuracy-discrimination frontier is a joint consequence of (i) overprecision and (ii) group bias. Eliminating group bias decreases discrimination by 50% at no cost to accuracy. In addition, correcting for overprecision increases accuracy by about 25%, but doing so more than doubles discrimination.

Differences in estimates for weight ω and bias B also explain why the NoGroup treatment differs significantly from the Baseline. That lack of group bias in NoGroup ($B_h - B_l$ not statistically different from zero) highlights that the technology we use to represent a signal (a grid with dots shown for .25 seconds) is unbiased, and that the existence of a group bias in Baseline is driven by knowledge of group identity. The fact that there is full weight on the signal and zero weight on the prior average mean of 50 suggests that subjects in NoGroup are naive, since they could significantly increase accuracy by putting some weight on this average mean (recall that the optimal weight on the average mean is represented by the OptNoDiscrimination counterfactual).[20,21]

Result 4 Withholding information about group identity eliminates discrimination, but this comes at a larger-than-necessary cost to accuracy because subjects do not use available prior information

[20] As discussed in Section 3.3, for the NoGroup treatment we estimate a model where subjects put weight on the signal and on the average mean of 50.

[21] The NoGroup point lies slightly below the pBRN benchmark because the estimated subject variance is slightly lower (though statistically not different) for the NoGroup treatment than for Baseline, and the Baseline variance was used to construct the pBRN benchmark.

(the average group mean).

Finally, we discuss the estimated values of the signal variance reported in Table 1. First, as reported in the table, for each treatment we find no statistically significant differences in estimated variances between groups. We also find no differences in estimated variances for the Baseline, NoGroup, and SignalFirst treatments, and the estimated variances are similar in magnitude to the population variance that we picked for the experiment. These findings corroborate three features of our design: (i) the fact that population and signal variances are of similar magnitudes imply that the optimal Bayesian weight on the signal is close to 0.5; this is empirically valuable because it leaves plenty of room to find either over or under-reaction to the prior or signal; (ii) the fact that variances do not differ by group suggests that subjects are paying similar attention to the signals from both groups, which is indeed optimal in our case because population variances are the same for both groups; and (iii) there seems to be little opportunity for subjects to put more or less effort in reading the 0.25 second signal, since otherwise we would have observed lower variance for the subjective signal in the NoGroup treatment, where group information is not provided and the signal is more valuable.

A final observation, is that there is a significant decrease in signal variance in OneGroup ($p < 0.05$)[22], where subjects are restricted to make inferences about only one group. We discuss this result in the next subsection.

4.4 Mechanism

We conducted our additional treatments, OneGroup and SignalFirst, with two purposes in mind. First, it is clear from the previous discussion that the raw data points for the Baseline and NoGroup treatments are far from the accuracy-discrimination frontier. Our additional treatments explore interventions that we hypothesized might move subjects closer to the accuracy-discrimination frontier. Second, the new treatments we consider serve a valuable diagnostic role, allowing us to better understand the sources of the inefficiency we observe in the Baseline treatment without necessarily relying on the structural model.

In SignalFirst, we consider the possibility that foreknowledge of the group interferes with subjects' perception – that knowing the group interferes with the quality of the signal subjects are able to extract from the grid of dots we show them, at the moment they observe it.

The SignalFirst treatment was designed to eliminate or at least reduce this interference, by

[22]We use bootstrapping to make statistical statements on signal variance estimates.

informing subjects of the group only after subjects observe the signal. The SignalFirst point in Figure 4 reveals two key facts relative to Baseline. First, the treatment neither raises nor lowers MSE relative to Baseline – the two are statistically indistinguishable. Second, despite this, subjects discriminate significantly less in SignalFirst than in Baseline ($p < 0.05$). Figure 5 shows that this is true across a range of values, with mean assessments for each group lying closer together in SignalFirst than Baseline. These patterns strongly suggest that at least part of the inefficiency we observe in the Baseline treatment is driven by the influence foreknowledge of the group has on subjects' ability to perceive an unbiased signal.

Looking at structural estimates in Table 1, we get insight into how the SignalFirst treatment achieves this reduction in discrimination. Evidence of overprecision is similar to that in Baseline: ω parameters (and the ω^{Bay} benchmarks against which we judge overprecision) are not statistically different across the two treatments. However estimates for the bias parameters are each strongly attenuated towards zero in SignalFirst relative to Baseline, and the group bias of $B_h - B_l = 1.4$, while still statistically significant ($p < 0.01$), is substantially smaller than in Baseline ($p < 0.01$). Thus, observing the signal before knowing the group "works" to reduce discrimination not by altering overprecision or, equivalently in our context, base-rate neglect but by shrinking the group bias.[23]

Result 5 Group bias is intensified when subjects have foreknowledge of group identity. As a result, when subjects assess the signal before knowing the group, they discriminate significantly less at no cost to accuracy.

Next, we consider the possibility that group bias is driven by the *contrast* between the two groups. To test this, we designed the OneGroup treatment to eliminate subjects' ability to build stereotypes based on group contrasts by assigning each subject to evaluate only one group (high or low) throughout the experiment. Figure 4 reveals a strong reduction in *both* inaccuracy and discrimination ($p < 0.01$ for both MSE and GD) in OneGroup relative to Baseline.

Structural estimates show that inferences suffer from overprecision in both treatments (the last two columns in Table 1 show that ω_g is significantly larger than ω_g^{Bay} in both cases). The ratio of ω_g to ω_g^{Bay} is larger in OneGroup relative to Baseline, suggesting more overprecision in the former treatment.

The most striking difference between the two treatments is reflected in the estimates for the

[23]The contrast between SignalFirst and Baseline is stronger in the second half of the session, which we discuss in Section 4.7.

bias parameters. OneGroup "works" by entirely eliminating group bias: B_l, B_h, and also their difference are all statistically indistinguishable from zero. The result thus suggests that group bias is *entirely* a consequence of stereotypes, enabled by scope provided to subjects to contrast groups across evaluations.

Result 6 Group bias entirely disappears when scope for contrasting groups across rounds is eliminated.

In addition to removing group bias, Figure 4 shows an additional advantage of the group-level specialization instituted in OneGroup. The OneGroup point actually lies below the frontier, indicating that subjects improve on the estimated tradeoff between accuracy and discrimination available in the Baseline treatment. The reason for this is simple and apparent in the estimates from Table 1. Estimates of the variance of the subjective signal are substantially lower in OneGroup than in Baseline, indicating that subjects are able to more precisely extract signals from the evidence provided to them. (Recall that as the variance of the signal decreases, the frontier moves downward–this explains why the OneGroup point is below the Baseline frontier, since the latter is computed using the lower signal precision of the Baseline treatment.) This decrease in variance seems unlikely to be due to contrast effects since variance does not decrease in the NoGroup treatment, where subjects make assessments without seeing group identity. Instead, it is more likely to be a result of the fact that in OneGroup subjects are specialized in assessing values from one population, potentially decreasing the complexity of the task.[24]

4.5 Using the model to improve outcomes

In the previous section, we presented two types of interventions that bring us closer to the accuracy-discrimination frontier. In this section, we show another way in which outcomes can be improved. Specifically, we show how a planner with access to data on true values and assessments can apply a model like ours to improve outcomes in terms of accuracy and discrimination. This type of data might be available for example to firms which combine data on initial assessment of workers at the hiring stage with long run value of workers for the firm as revealed in internal reviews and career trajectories. Likewise, universities can combine initial assessment of students at the application stage with their performance at graduation.

[24]The decrease is variance is consistent with the literature on efficient coding (e.g., Khaw, Li & Woodford (2021) and Frydman & Jin (2021)), which argues and provides experimental evidence that when the prior distribution has lower variance (as in OneGroup, where subjects face one distribution as opposed to a mixture over two distributions) then encoding of information is such that the signal is also more precise.

Our model suggests that an outside observer can construct a simple algorithm to correct for the errors we find: overprecision and group bias. Identification and estimation of these errors allows the observer to improve outcomes at the prediction stage by (i) debiasing assessments, and (ii) readjusting weight on the signal vs prior information.

Here, we illustrate how this can be done with our own data. We divide our data (focusing on the Baseline treatment) into two: a training and a testing set. The training set is used to estimate parameters of the model (B, ω, ξ^2). These estimates are then used to "adjust" predictions in the testing set. By separating the data on which we estimate parameters and apply adjustments, we can test the performance of the model out of sample. As a proof of concept, we focus on two types of adjustments that are intended to: (i) maximize accuracy subject to zero discrimination, i.e. OptNoDiscrimination benchmark; (ii) maximize accuracy subject to no constraint, i.e., Bayesian benchmark. The adjustments are done using the following steps.

1. Given (B, ω, ξ^2), for each observation in the testing set, estimate a (debiased) subjective signal.
    - Each assessment $\tilde{v}$ can be represented as follows: $\tilde{v} = B + \omega s_u + (1-\omega)\mu$, where s_u is an unbiased signal.[25] Thus, s_u can be computed from $\tilde{v}$ using μ, ω, B.
2. The signal variance estimate ξ^2 implies an optimal weight ω for each of the two benchmarks.
    - For the OptNoDiscrimination benchmark, $\omega^{Ond} = \frac{2\sigma^2}{\xi^2+2\sigma^2}$[26]
    - For the Bayesian benchmark (as described in Section 3.4), $\omega^{Bay} = \frac{\sigma^2}{\xi^2+\sigma^2}$.
3. Compute adjusted prediction using estimates for signal s_u and weights ω^{Ond} or ω^{Bay}.
    - Adjusted Bayesian prediction $\tilde{v}^{Bay} = \omega^{Bay} s_u + (1-\omega^{Bay})\mu_g$.
    - Adjusted OptNoDiscrimination prediction $\tilde{v}^{Ond} = \omega^{Ond} s_u + (1-\omega^{Ond})\left(\frac{\mu_l+\mu_h}{2}\right)$.

Table 2 reports results from 500 random repetitions of the procedure described above. There are two observations. First, adjusted predictions in the OptNoDiscrimination column are welfare increasing—generate lower inaccuracy and discrimination—relative to the data. Second, adjusted predictions in the Bayesian column display higher discrimination, but much lower inaccuracy as

[25] From Equations 4 and 5, $\tilde{v} = b^P + \omega s + (1-\omega)\mu = b^P + \omega(b^S + v + \varepsilon') + (1-\omega)\mu = B + \omega s_u + (1-\omega)\mu$, where $s_u := s - b^S = v + \varepsilon'$.

[26] Consistent with our linearity restriction in Section 3.2, ω is chosen to minimize expected squared error of predictions $\tilde{v} = \omega s + (1-\omega)\left(\frac{\mu_l+\mu_h}{2}\right)$.

Table 2: Actual vs. Adjusted Predictions

	Data		OptNoDiscrimination		Bayesian
GD	7.6	$>^{***}$	0.3	$<^{***}$	9.4
MSE	63	$>^{***}$	59	$>^{***}$	44

Values represent mean values from 500 repetitions of the procedure.

$>^{***}$ means equality is violated in less than 1% of repetitions.

expected from our earlier analysis. These results demonstrate how the model can be useful for correcting for group bias and overprecision in applications.

4.6 Individual-level estimates

So far, we have computed counterfactuals using pooled data. In Appendix E, we present the counterparts of Table 1 and Figure 4, reporting median-values from individual-level estimates. All of our main conclusions—on deviations from Bayesianism (overprecision, group bias) and their consequences for accuracy and discrimination—remain true at the individual level. In the main text, we highlight that there is significant heterogeneity across subjects. But, whenever we see differences in the aggregate data, these differences carry over to the first-order stochastic order when considering individual variation.

Figures 6 and 7 plot CDFs of B and $\omega - \omega^{Bay}$ respectively from our structural model, but estimated at the individual subject level. The results reveal significant heterogeneity across subjects, but that the patterns are nonetheless the same as in the pooled analysis. The distribution of the bias parameter B_h for the high-mean group strongly first order stochastically dominates the distribution of B_l for the low-mean group in Baseline and can be distinguished from one another by a Kolmogorov-Smirnov test ($p < 0.001$).[27] The difference in median estimates for this treatment between B_h and B_l is 4.5, even larger than the aggregate estimate of 3.6. By contrast, the two distributions for B_h and B_l are nearly identical in No Group with median estimates very close to 0. The same is also true in OneGroup. In SignalFirst, the two distributions are closer than those in Baseline but can still be distinguished from one another by a Kolmogorov-Smirnov test ($p < 0.001$); the difference in median estimates is 1.8, less than half of what was observed in the Baseline.

[27] In Appendix G we also include the cumulative distribution of group bias $B_h - B_l$ estimated on the individual level. 72 percent of subjects in Baseline display positive group bias. This decreases to 66 percent in SignalFirst and 54 percent in NoGroup.

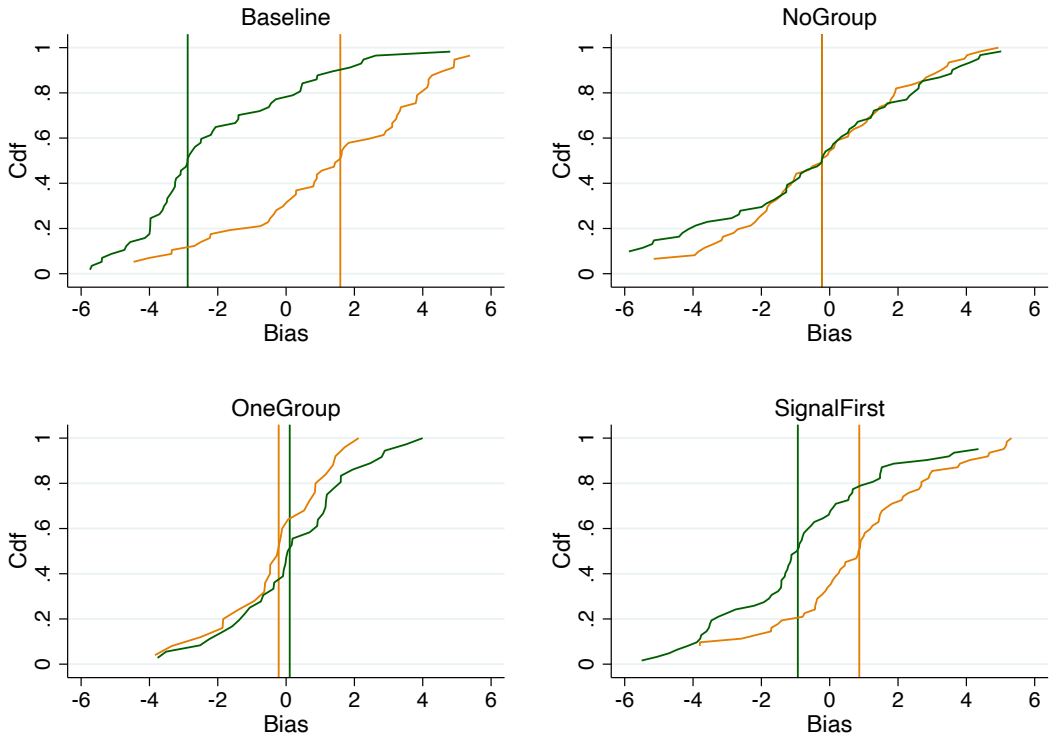


Figure 6: Estimates for Bias Parameter by Group and Treatment *Notes: Green (Orange) line represents estimated bias parameter for low-mean (high-mean) group. Vertical lines denote median value.*

As Figure 7 shows, subjects do not optimally weight the subjective signal and the prior group information. This is true for both groups in all treatments. Solid lines depict the cumulative distribution of the difference between estimated weight on prior ω and Bayesian weight ω^{Bay}: positive values correspond to overprecision, where too much weight is placed on the signal, and negative values correspond to underprecision, or too much weight placed on the prior. More than 85 percent of subjects in all treatments have estimated values that are positive, and overprecision is almost universally observed in the OneGroup and and NoGroup treatments.

4.7 Learning

In our analysis of the data, we have so far abstracted from learning effects. Recall that subjects in our experiment (in all treatments) made 75 rounds of assessments. At the end of each round, subjects were reminded of their decision and received feedback about the true value of the person they were making an assessment about. Such feedback gives subjects the opportunity to recognize patterns in their mistakes, enabling them to adjust their assessment strategy. To explore these

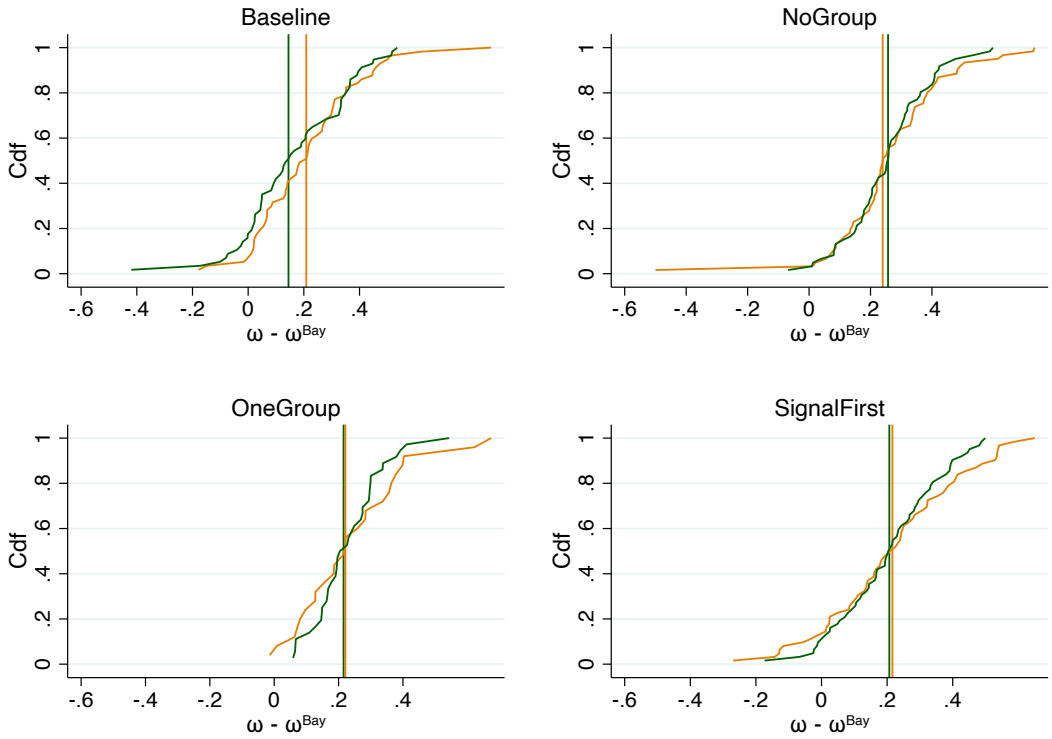


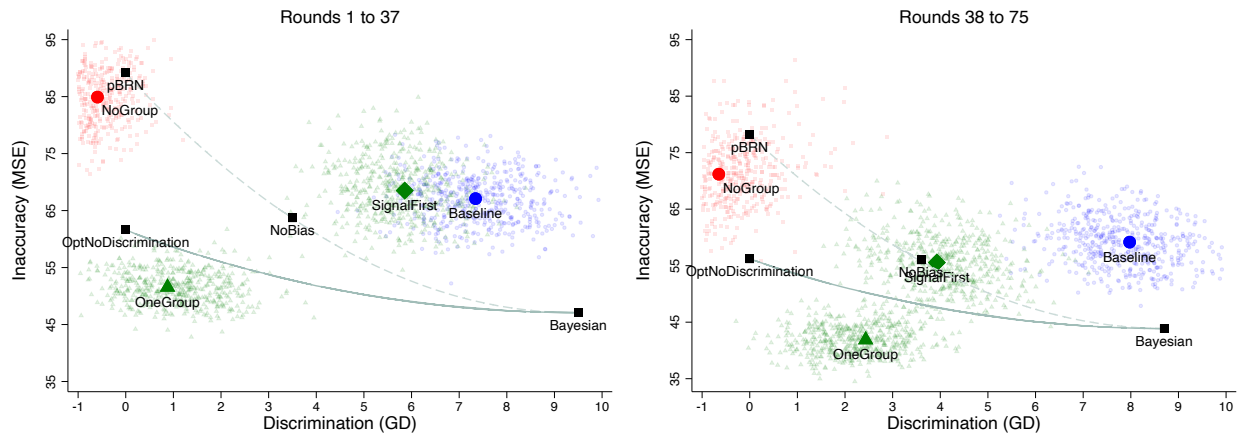
Figure 7: Estimates of Weight vs. Bayesian Weight on Prior by Group and Treatment *Notes: Green (Orage) solid line represents estimated difference between estimated weight on prior ω and Bayesian weight ω^{Bay} for low-mean (high-mean) group. Vertical lines denote median values.*

effects, Figure 8 reproduces Figure 4 separating data for early (1-37) and late (38-75) rounds. In Appendix F we also report model estimates (reproducing Table 1) separately for these early and late rounds.

We start by noting that the two panels in Figure 8 paint a very similar picture in terms of how our treatments compare to each other in terms of accuracy and discrimination, and how these points locate relative to counterfactual benchmarks, and the accuracy-discrimination frontier. Despite these qualitative similarities, there are visually observable learning effects.

The first notable contrast between the two panels is that observations for all treatments—as well as the accuracy-discrimination frontier including counterfactual benchmarks—are shifted downwards for later rounds. This is primarily a consequence of the reduction in signal variance estimates (reported in Tables 6 and 7 of Appendix F), which suggest that subjects are getting better over time at reading the subjective signal.

Second, the two types of mistakes that generate inefficiency in our setting—overprecision and



(a) (b)

Figure 8: Contrasting Early versus Late Rounds *Notes: Solid line is the accuracy-discrimination frontier, the dashed line represents loci of points computed at different weights on the subjective signal, pBRN depicts the perfect base-rate neglect counterfactual, OptNoDiscrimination depicts the counterfactual with highest accuracy (at zero discrimination), Bayesian refers to the counterfactual with highest accuracy (see Sections 3.5 and 3.4 for details).*

group bias—are persistent in later rounds. There is no strong evidence to suggest that subjects are able to learn from their experiences in a way that enables them to correct these errors. This can be seen by comparing the location of the Baseline point to two counterfactual benchmarks: NoBias (where $B = 0$ and the value of ω is set equal to the Baseline estimate) and Bayesian (where $B = 0$ and $\omega = \omega^{Bay}$): (i) the distance between the Baseline point and the NoBias benchmark is as large in later rounds as in earlier rounds indicating that subjects cannot learn to correct group bias; (ii) the NoBias and Bayesian benchmarks remain substantially apart in later rounds indicating that subjects are not able to correct for overprecision. Tables 6 and 7 of Appendix F confirm these observations by showing that the estimate of $B_h - B_l$ in Baseline does not decline and ω remains significantly different from ω^{Bay} in later rounds.

Result 7 While the noisiness of the subjective signal declines over time, the two types of mistakes that generate inefficiency—overprecision and group bias—are persistent.

Arguably the most striking contrast between earlier and later rounds displayed in Figure 8 is the location of the SignalFirst point relative to Baseline and the NoBias benchmark. The decline in discrimination, as we move from Baseline to SignalFirst, is larger in later rounds, so much so that the SignalFirst point almost comes to coincide with the NoBias benchmark. This is a consequence of the decline in the estimates for $B_h - B_l$ in SignalFirst in later rounds relative to earlier rounds (as reported in Tables 6 and 7 of Appendix F).

The comparison between Baseline and SignalFirst in later rounds provides some insight into the nature of the bias term in the long run. Recall that group bias is a combination of prediction bias, b^P, and signal-perception bias, b^S, i.e., $B = b^P + \omega b^S$. A reasonable assumption is that the signal perception bias is zero in the SignalFirst treatment, since this bias captures how knowledge of group identity affects signal perception.[28] Under this assumption, $B = b^P$ in the SignalFirst treatment. But, estimates for B in SignalFirst in later rounds are not statistically different from zero. Moreover, it is natural to assume that the prediction bias term b^P is the same in both the Baseline and SignalFirst treatments. The reason is that this is an error that arises at the assessment stage, and the information available to the subject at this stage is identical in both treatments. Under these assumptions, we can conclude that the group bias observed in the Baseline (in later rounds) is almost entirely driven by signal-perception bias (and not prediction-bias).

Result 8 The decline in discrimination from Baseline to SignalFirst is larger in later rounds. The comparison suggests group bias in later rounds of the Baseline is mostly driven by signal-perception bias (distortion of subjective interpretation of the signal, driven by knowledge of the group).

5 Discussion

Decades of work in behavioral economics has established that human beings are not "intuitive statisticians." That is, we do not have an inborn capacity to process information in a Bayesian way to produce accurate beliefs. Our results suggest that this is no less true of classical statistical discrimination, where statistical errors combine with biased interpretation of subjective information in a distinctive way to produce not only inaccuracy but also inefficiently inflated rates of discrimination. We view these results as providing several lessons for economists and other social scientists, and as suggesting several lines of future research.

First, our results suggest that decision makers discriminate but tend do so far less effectively than statistical discrimination models predict. This means that simply by improving our ability to discriminate *effectively*, we can discriminate *less* and without sacrificing accuracy. Indeed our results suggest that discrimination can be reduced and accuracy simultaneously improved simply by finding ways to combat cognitive and perceptual errors. There is therefore a sort of behavioral "free lunch" available when managing discrimination that comes from finding policies and choice

[28] One may be concerned that a signal-perception bias would still exist in the absence of group information, but B is estimated to be essentially zero in the NoGroup treatment, suggesting that the only source of bias in the signal technology is knowledge of group identity.

environments in which the types of mistakes we've identified are eased.

Our experiment identifies two such policies, but there are surely more to be designed and discovered. First we show (in our SignalFirst treatment) that simply forcing decision makers to evaluate individuals *before* knowing what group they belong to reduces discrimination at no cost to accuracy. This treatment mirrors an important line of empirical work showing the effectiveness of similar "identity-blinded evaluation policies" in the field,[29] but our abstract setting helps us to understand (at least in part) why these types of policies work and what sorts of cognitive mistakes they correct. Second, and more dramatically, we show (in our OneGroup treatment) that enforcing *specialization* by assigning decision makers to assess only members of one group reduces discrimination while increasing accuracy. It is easy to imagine deploying specialization policies that mirror this treatment. Many other policy interventions may also have the potential to correct the inefficiencies identified in this paper, and searching for such policies seems like a natural next step in this research.

Second, our experiment was designed to identify and disentangle the cognitive and perceptual errors driving the mistakes we observe in statistical discrimination. To do this, we introduce a simple conceptual framework and structural model that separately identifies channels that govern the trade off between accuracy and discrimination. Using estimates from this model we characterize two key deviations from rational statistical discrimination.

One of these mistakes is "group bias" – a tendency to crudely account for group differences in a way that increases discrimination while simultaneously decreasing accuracy. This bias is significantly reduced when subjects do not not know a person's group prior to observing idiosyncratic evidence on that person (i.e. in our SignalFirst treatment, especially with experience), suggesting that the bias arises in the interpretation of evidence (in our case, in perception). Notably, the fact that this bias works through perception suggests the importance of a central element of our design: that subjects are asked to combine subjectively evaluated information (in our case, a perceptual task) with objective statistical information to form their assessment. This type of integration (objective statistical information with subjective perceptual information) seems present in some of the most important settings where we expect statistical discrimination to arise.

Group bias disappears altogether in our OneGroup treatment, suggesting that it is ultimately

[29]This includes studies of the effect of "veiling" characteristics of workers' from evaluators, includes the worker's gender (Goldin & Rouse 2000, Krause et al. 2012), ethnicity (Behaghel et al. 2015), criminal history (Agan & Starr 2018, Doleac & Hansen 2020, Sherrard 2021), credit history (Bos et al. 2018, Ballance et al. 2020), salary history (Agan et al. 2021) and gender categorization of jobs (Kuhn & Shen 2021).

driven by "contrast effects," arising when subjects are asked to assess individuals from two or more groups over time. The primacy of contrast effects here connects our findings to Bordalo, Coffman, Gennaioli & Shleifer (2016) which describes and tests for stereotypes that arise because decision makers use salient differences in the contrast between groups to characterize members of groups (a variation on Kahneman and Tversky's representativeness heuristic). Because we are able to identify our group bias as perceptual, our experiment provides important clues about *how* people might arrive at such biased inferences. In evaluating subjective signals, our subjects seem to "look for" (consciously or not) evidence that is "representative" of the group the individual belongs to, resulting in subjective signals that are systematically biased to amplify group differences. This finding also connects directly with evidence from psychology on how people's perceptions can be distorted by what they expect to see (Kelley 1950, Darley & Gross 1983, Lord et al. 1979). Moreover, these results provide an explanation for how people might come to hold systematically inaccurate beliefs about group differences, even when (as in our experiment) they start out with correct priors.

The other mistake our estimates reveal is "overprecision": subjects act as if their noisy subjective perceptions are less noisy than they actually are, causing them to put excess weight on this information in forming evaluations. This mistake reduces discrimination, but at a cost to accuracy. In contrast to group bias, we find that overprecision is robust to the policy interventions we've explored. Finding interventions that are effective at combating this further mistake is an important avenue for future research, though we caution that unlike group bias, overprecision is discrimination-reducing. Removing it will tend to increase accuracy at the cost of also increasing discrimination.

Identifying these psychological channels is important for several reasons. For one thing, we find that these mistakes are highly resistant to learning and do not abate with experience, meaning correcting them likely will demand policy interventions or changes to the procedures we use to make decisions. Identification of the bedrock mistakes driving these inefficiencies may be crucial for identifying and designing effective methods and policies to remove them. For another, our results from Section 4.5 suggest that the framework we've used to identify these channels might also be directly useful for developing algorithms to "debias" assessments ex post (by removing cognitive and perceptual errors) in field applications (e.g. hiring, admissions or criminal evaluations). Applying these methods to our data, we find that we can debias assessments in a way that lowers both inaccuracy and discrimination. Exploring the use of our framework and debiasing algorithms built from it in field applications seems a promising direction for future work.

Finally, our experiment demonstrates the value of complementing rich field data with data

from abstract experimental designs. Because our experiment is abstract, we are able to remove other (surely important) drivers of discrimination, such as"taste based discrimination" or animus, allowing us to clearly interpret the deviations from the rational model we observe as stemming from purely cognitive and perceptual errors. The fact that we find significant discrimination that is not explainable by rational statistical discrimination in such a design, suggests that some apparent taste based discrimination or animus in estimates using field data might be confounded with simple cognitive mistakes. This observation is useful, but it suggests that further research designs are needed to decompose unexplained discrimination in richer contexts into components that are driven by taste versus cognitive/perceptual errors. Doing this is important because the two causes of discrimination are likely to suggest very different sorts of policy interventions to combat discrimination. This will require new research designs that re-introduce scope for taste-based discrimination into designs like ours and iterations on our framework that can separate biases due to taste from those due to mistakes. This seems a particularly promising next step in this research.

References

Agan, A. & Starr, S. (2018), 'Ban the box, criminal records, and racial discrimination: A field experiment', *The Quarterly Journal of Economics* **133**(1), 191–235.

Agan, A. Y., Cowgill, B. & Gee, L. K. (2021), Salary history and employer demand: Evidence from a two-sided audit, Technical report, National Bureau of Economic Research.

Araujo, F. A., Wang, S. W. & Wilson, A. J. (2021), 'The times they are a-changing: Experimenting with dynamic adverse selection', *American Economic Journal: Microeconomics* **13**(4), 1–22.

Arnold, D., Dobbie, W. & Yang, C. S. (2018), 'Racial bias in bail decisions', *The Quarterly Journal of Economics* **133**(4), 1885–1932.

Arrow, K. (1973), *The theory of discrimination*, Princeton University Press.

Ballance, J., Clifford, R. & Shoag, D. (2020), 'no more credit score: Employer credit check bans and signal substitution', *Labour Economics* **63**, 101769.

Barocas, S., Hardt, M. & Narayanan, A. (2019), *Fairness and Machine Learning*, fairmlbook.org. http://www.fairmlbook.org.

Barron, K., Huck, S. & Jehiel, P. (2019), Everyday econometricians: Selection neglect and overoptimism when learning from others, Technical report, WZB Discussion Paper.

Becker, G. S. (1957), *The economcis of discrimination*, University Chicago Press.

Behaghel, L., Crépon, B. & Le Barbanchon, T. (2015), 'Unintended effects of anonymous resumes', *American Economic Journal: Applied Economics* **7**(3), 1–27.

Benjamin, D., Bodoh-Creed, A. & Rabin, M. (2019), Base-rate neglect: Foundations and implications, Technical report, working paper.

Bertrand, M. & Duflo, E. (2017), 'Field experiments on discrimination', *Handbook of economic field experiments* **1**, 309–393.

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

Bordalo, P., Coffman, K., Gennaioli, N. & Shleifer, A. (2016), 'Stereotypes', *The Quarterly Journal of Economics* **131**(4), 1753–1794.

Bos, M., Breza, E. & Liberman, A. (2018), 'The labor market effects of credit market information', *The Review of Financial Studies* **31**(6), 2005–2037.

Cattaneo, M. D., Crump, R. K., Farrell, M. H. & Feng, Y. (2019), 'Binscatter regressions', *arXiv preprint arXiv:1902.09615* .

Charles, K. K. & Guryan, J. (2011), 'Studying discrimination: Fundamental challenges and recent progress', *Annu. Rev. Econ.* **3**(1), 479–511.

Charness, G., Oprea, R. & Yuksel, S. (2021), 'How do people choose between biased information sources? evidence from a laboratory experiment', *Journal of the European Economic Association* **19**(3), 1656–1691.

Darley, J. M. & Gross, P. H. (1983), 'A hypothesis-confirming bias in labeling effects.', *Journal of Personality and Social Psychology* **44**(1), 20.

Doleac, J. L. & Hansen, B. (2020), 'The unintended consequences of ban the box: Statistical discrimination and employment outcomes when criminal histories are hidden', *Journal of Labor Economics* **38**(2), 321–374.

Enke, B. (2020), 'What you see is all there is', *The Quarterly Journal of Economics* **135**(3), 1363–1398.

Esponda, I. & Vespa, E. (2014), 'Hypothetical thinking and information extraction in the laboratory', *American Economic Journal: Microeconomics* **6**(4), 180–202.

Esponda, I., Vespa, E. & Yuksel, S. (2019), Mental models and learning: The case of base rate neglect, Technical report, working paper.

Eyster, E. & Rabin, M. (2005), 'Cursed equilibrium', *Econometrica* **73**(5), 1623–1672.

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

Frydman, C. & Jin, L. J. (2021), 'Efficient coding and risky choice', *Quarterly Journal of Economics, Forthcoming* .

Gillen, B., Snowberg, E. & Yariv, L. (2019), 'Experimenting with measurement error: Techniques with applications to the caltech cohort study', *Journal of Political Economy* **127**(4), 1826–1863.

Goldin, C. & Rouse, C. (2000), 'Orchestrating impartiality: The impact of" blind" auditions on female musicians', *American economic review* **90**(4), 715–741.

Gottlieb, D. & Smetters, K. (2021), 'Lapse-based insurance', *American Economic Review* .

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

Grubb, M. D. (2009), 'Selling to overconfident consumers', *American Economic Review* **99**(5), 1770–1807.

Grubb, M. D. & Osborne, M. (2015), 'Cellular service demand: Biased beliefs, learning, and bill shock', *American Economic Review* **105**(1), 234–71.

Hutchinson, B. & Mitchell, M. (2019), 50 years of test (un) fairness: Lessons for machine learning, *in* 'Proceedings of the Conference on Fairness, Accountability, and Transparency', pp. 49–58.

Kahneman, D. & Tversky, A. (1972), 'On prediction and judgement', *ORI Research monograph* **1**(4).

Kelley, H. H. (1950), *The warm-cold variable in first impressions of persons.*

Khaw, M. W., Li, Z. & Woodford, M. (2021), 'Cognitive imprecision and small-stakes risk aversion', *The review of economic studies* **88**(4), 1979–2013.

Klayman, J. (1995), 'Varieties of confirmation bias', *Psychology of learning and motivation* **32**, 385–418.

Koehler, J. J. (1996), 'The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges', *Behavioral and brain sciences* **19**(1), 1–17.

Krause, A., Rinne, U. & Zimmermann, K. F. (2012), 'Anonymous job applications in europe', *IZA Journal of European Labor Studies* **1**(1), 1–20.

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

Lord, C. G., Ross, L. & Lepper, M. R. (1979), 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.', *Journal of personality and social psychology* **37**(11), 2098.

Lundberg, S. J. & Startz, R. (1983), 'Private discrimination and social intervention in competitive labor market', *The American Economic Review* **73**(3), 340–347.

Mengel, F. & Campos Mercade, P. (2021), 'Irrational statistical discrimination', *Available at SSRN 3843579* .

Mobius, M. M., Niederle, M., Niehaus, P. & Rosenblat, T. S. (2021), Managing self-confidence: Theory and experimental evidence, Technical report, National Bureau of Economic Research.

Mobius, M. M. & Rosenblat, T. S. (2006), 'Why beauty matters', *American Economic Review* **96**(1), 222–235.

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

Narayanan, A. (2018), Translation tutorial: 21 fairness definitions and their politics, *in* 'Proc. Conf. Fairness Accountability Transp., New York, USA', Vol. 1170.

Neumark, D. (2018), 'Experimental research on labor market discrimination', *Journal of Economic Literature* **56**(3), 799–866.

Ngangoué, M. K. & Weizsäcker, G. (2021), 'Learning from unrealized versus realized prices', *American Economic Journal: Microeconomics* **13**(2), 174–201.

Nickerson, R. S. (1998), 'Confirmation bias: A ubiquitous phenomenon in many guises', *Review of general psychology* **2**(2), 175–220.

Oprea, R. & Yuksel, S. (2021), 'Social exchange of motivated beliefs', *Journal of the European Economic Association* .

Phelps, E. S. (1972), 'The statistical theory of racism and sexism', *The American Economic Review* **62**(4), 659–661.

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

Sherrard, R. (2021), ''ban the box'policies and criminal recidivism', *Available at SSRN 3515048* .

Soll, J. B. & Klayman, J. (2004), 'Overconfidence in interval estimates.', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **30**(2), 299.

Weizsäcker, G. (2010), 'Do we follow others when we should? a simple test of rational expectations', *American Economic Review* **100**(5), 2340–60.

Online Appendix for

Behavioral Statistical Discrimination

Ignacio Esponda Ryan Oprea Sevgi Yuksel

CONTENTS:

A Measures of Discrimination

We refer the reader to Barocas, Hardt & Narayanan (2019),Narayanan (2018) andHutchinson & Mitchell (2019) for reviews of this literature. We focus on a criteria of non-discrimination that is most relevant in our setting: *Separation*.[30] In our inference task, this criteria refers to the statistical properties of the joint distribution of $\tilde{v}$ (assessment), v (true value) and g (group identity).

> *Separation*: $\tilde{v} \perp g \,|\, v$. This criterion requires assessments to be independent of the group identity *conditional* on value.

Separation allows for the distribution of assessments to differ by group, but only to the extent that such differences can be justified by actual differences in true values between the groups. Namely, the criterion requires people from different groups with the same underlying true value to be treated the same. This is notion of fairness reflected for example in the slogan "equal pay for equal work" with regards to the gender pay gap.

Note that group difference (as captured by our measure GD) being equal to zero is a necessary (but not sufficient) condition for Separation. In this sense, GD provides us a with a preliminary test of Separation, as well as a simple, easy-to-interpret continuous measure of the degree to which it is violated. However, it is worth noting that, in using GD, we implicitly focus on the first moment (differences in means) when we contrast distribution of assessments between different groups. More complex measures of discrimination based on the Separation criterion can be constructed by incorporating different features (such as second, third moments etc.) of the distribution.[31]

[30] This literature highlights two more criteria for non-discrimination that could be relevant in our setting: Independence, and Sufficiency (closely liked to Calibration which is also commonly discussed in this literature).

> *Independence*: $\tilde{v} \perp g$. This criterion requires assessments to be independent of the group identity. Note that this cannot be a reasonable goal in an inference task where the distribution of values *do* differ by group (as in our setting).
>
> *Sufficiency*: $v \perp g \,|\, \tilde{v}$. This criterion requires values to be independent of the group identity *conditional* on assessments. The criterion requires the distribution of true values to be the same for different groups when we condition on a specific assessment. The Bayesian Benchmark to our inference task satisfies this criterion.

[31] We note, however, that it is far from obvious what type of impact these other features of the distribution should have on a measure of discrimination. Moreover, our take on this issue is highly likely to be context dependent. We have deliberately kept our experimental design simple by abstracting away from the issue of how assessments impact the utility of the individuals who are being evaluated. However, in most applications, we worry about discrimination foremost because of the utility loss it induces on those that are being discriminated against. For example, a manager's evaluation of the candidate might influence the likelihood they are hired for a job; a teacher's evaluation of a student

B Bayesian Benchmark in NoGroup Treatment

While the optimal inference about value v is linear in signal s—under the assumption that $s \sim \mathcal{N}\left(v, \xi^2\right)$—in the Baseline, SignalFirst and OneGroup treatments, this is not the case in the NoGroup treatment where information on group identity is withheld from the subjects. We can use law of iterated expectations to characterize the optimal inference as a function of signal s in this treatment:

$$\tilde{v}^{Bay} = \mathbb{E}(v \,|\, s) = p(g = h \,|\, s)\mathbb{E}(v \,|\, s, g = h) + p(g = l \,|\, l)\mathbb{E}(v \,|\, s, g = l),$$

where $p(g \,|\, s)$ denotes the probability that the person belongs to group g conditional on signal s. By Bayes' rule:

$$p(g \,|\, s) = \frac{\int \frac{1}{\sigma}\phi\left(\frac{v-\mu_g}{\sigma}\right)\frac{1}{\xi}\phi\left(\frac{s-v}{\xi}\right)dv}{\int \frac{1}{\sigma}\phi\left(\frac{v-\mu_h}{\sigma}\right)\frac{1}{\xi}\phi\left(\frac{s-v}{\xi}\right)dv + \int \frac{1}{\sigma}\phi\left(\frac{v-\mu_l}{\sigma}\right)\frac{1}{\xi}\phi\left(\frac{s-v}{\xi}\right)dv}$$

Figure 9 below depicts the Bayesian benchmark for three different values of $\xi^2 \in \{50, 75, 100\}$. The Figure shows that for such values (which cover the range estimated in the experiment), the Bayesian benchmark can be approximated closely with a linear function.

might impact the kind of college they are able to get into, etc. How different features of the distribution of assessments translate into utility differences between different groups will necessarily depend the specifics of the setting.

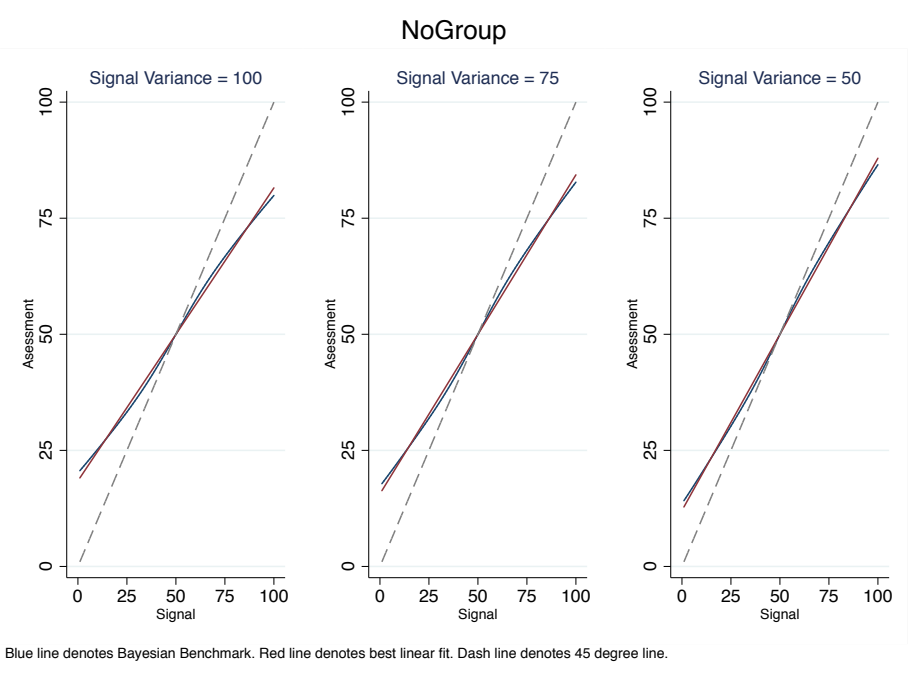


Figure 9: Bayesian Benchmark in NoGroup Treatment

C Statistical Tests

Table 3: OLS estimation (Dependent Variable: Squared Error in Assessments)

	(1)	(2)	(3)	(4)
NoGroup	14.88**	10.60*	8.432	6.766
	(6.045)	(5.887)	(17.35)	(18.14)
OneGroup	-16.43***	-15.34***	-34.38**	-34.99**
	(4.774)	(4.768)	(14.01)	(14.54)
SignalFirst	-1.131	-2.420	-7.115	-7.967
	(6.135)	(6.161)	(16.89)	(17.27)
Risk measure		-0.251***		0.0205
		(0.0763)		(0.165)
Constant	63.08***	76.24***	84.91***	84.77***
	(3.973)	(6.017)	(13.21)	(14.69)
Observations	17325	16950	18075	17700

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Constant shows MSE at Baseline at risk measure of zero.

Dummies shows difference relative to Baseline.

Lower values for risk measure correspond to higher risk aversion.

Risk measure missing for 5 subjects.

(1) and (2): Subjects with MSE less than or equal to 200.

(3) and (4): All subjects.

Table 4: Model Estimates by Treatment (All Rounds)

	Baseline	NoGroup	SignalFirst	OneGroup
ω_l	0.816***	1.003***	0.812***	0.918***
	(0.0308)	(0.0298)	(0.0271)	(0.0217)
ω_h	0.825***	1.019***	0.857***	0.885***
	(0.0311)	(0.0285)	(0.0245)	(0.0313)
B_l	-2.110***	-0.495	-0.875***	0.266
	(0.322)	(0.475)	(0.316)	(0.304)
B_h	1.519***	-0.217	0.500	-0.413
	(0.342)	(0.351)	(0.380)	(0.289)
Observations	4050	4350	4425	4500

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Subjects with MSE less than or equal to 200.

D Testing linearity

In Figure 10, we compare best linear fit with best fractional polynomial fit to demonstrate that the linearity assumption is with little loss in the relevant region of the value distribution. In Figure 11 we use the binscatter methods developed in Cattaneo, Crump, Farrell & Feng (2019). The estimated values show nonparametric estimates for assessment conditional on value for each bin, also displaying confidence bands. We also use the binscatter-based hypothesis testing procedures developed by Cattaneo, Crump, Farrell & Feng (2019) and find that a linear function form cannot be rejected in the Baseline treatment for either the high-mean or low-mean group.

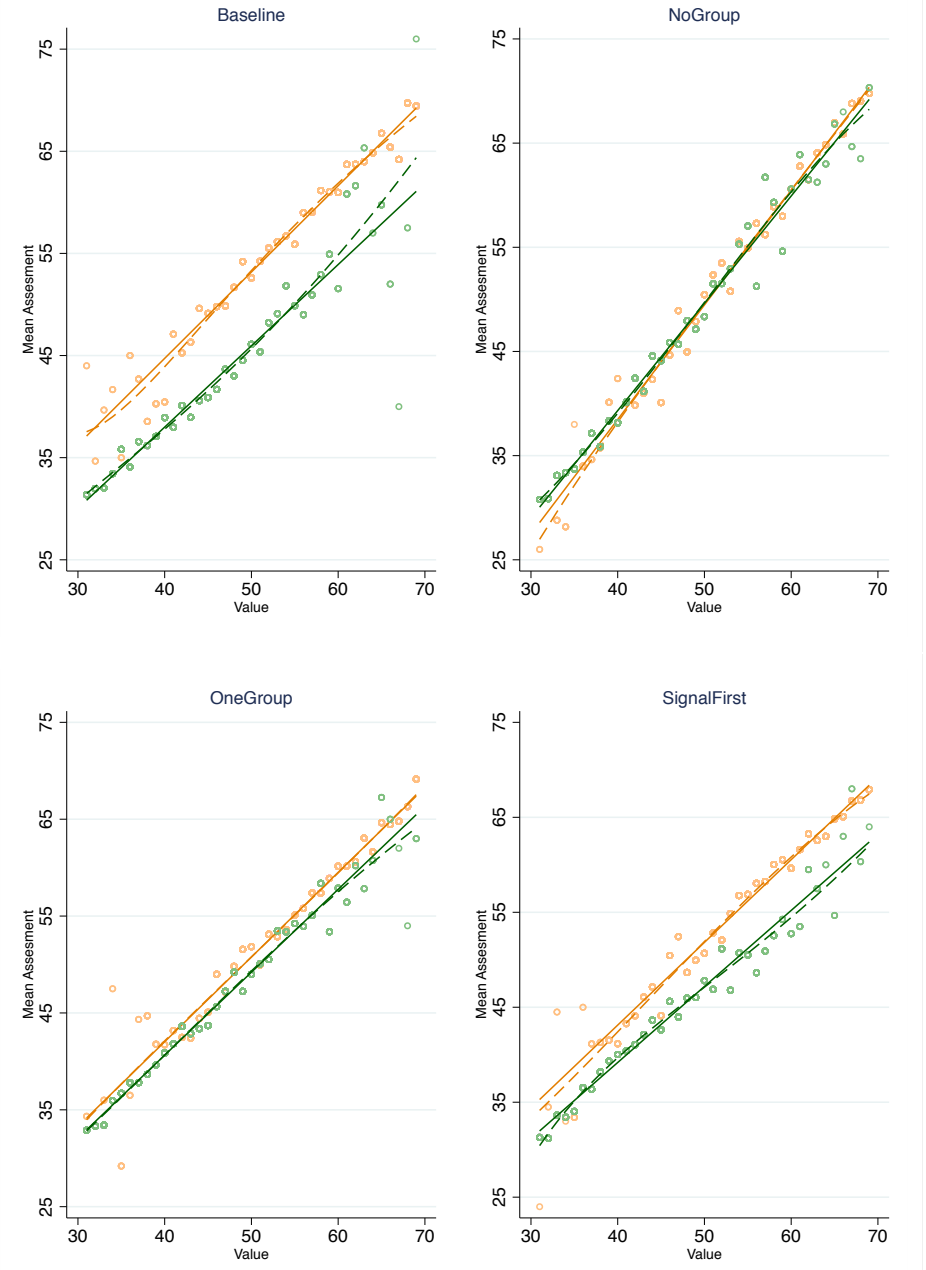


Figure 10: Best Linear Fit vs. Best Fractional Polynomial Fit

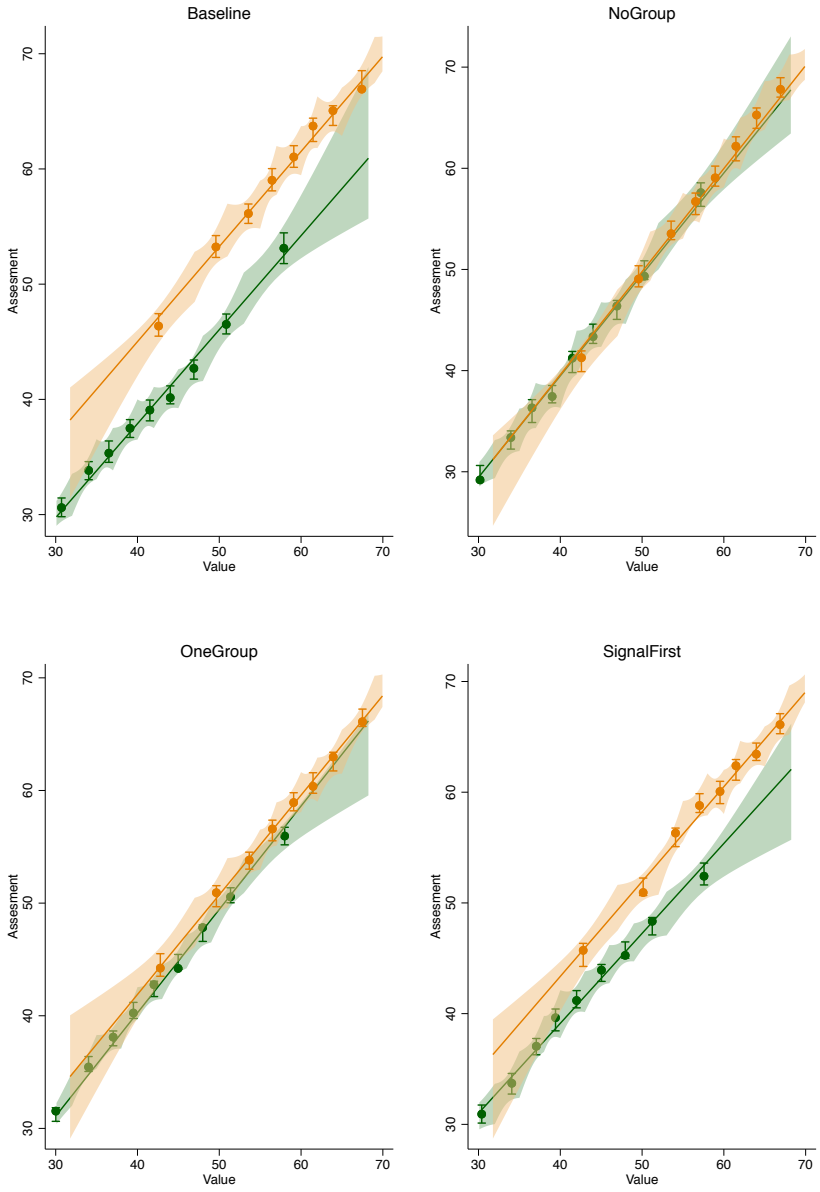


Figure 11: Best Linear Fit and Binned Scatter Plots

E Individual-Level Analysis

Table 5: Model Estimates on Individual-Level (Median Values)

	ω_l	B_l	ω_h	B_h	$\omega_h = \omega_l$	$B_h = B_l$	ξ_l^2	ξ_h^2	ω_l^{Bay}	ω_h^{Bay}	$\omega_l = \omega_l^{Bay}$	$\omega_h = \omega_h^{Bay}$
Baseline	0.80	-2.9	0.80	1.6		***	54	65	0.65	0.60	***	***
NoGroup	1.05	-0.2	1.04	-0.2			57	53	0.78	0.79	***	***
SignalFirst	0.82	-0.9	0.84	0.9		***	50	57	0.67	0.64	***	***
OneGroup	0.92	0.1	0.90	-0.2			42	49	0.70	0.67	***	***

Subscript l (h) denotes low-mean (high-mean) group.

Stars denote the confidence level with which the hypothesis associated with the column (that the distributions are the same) can be rejected (Wilcoxon rank-sum test).

*** 1%, ** 5%, * 10% significance.

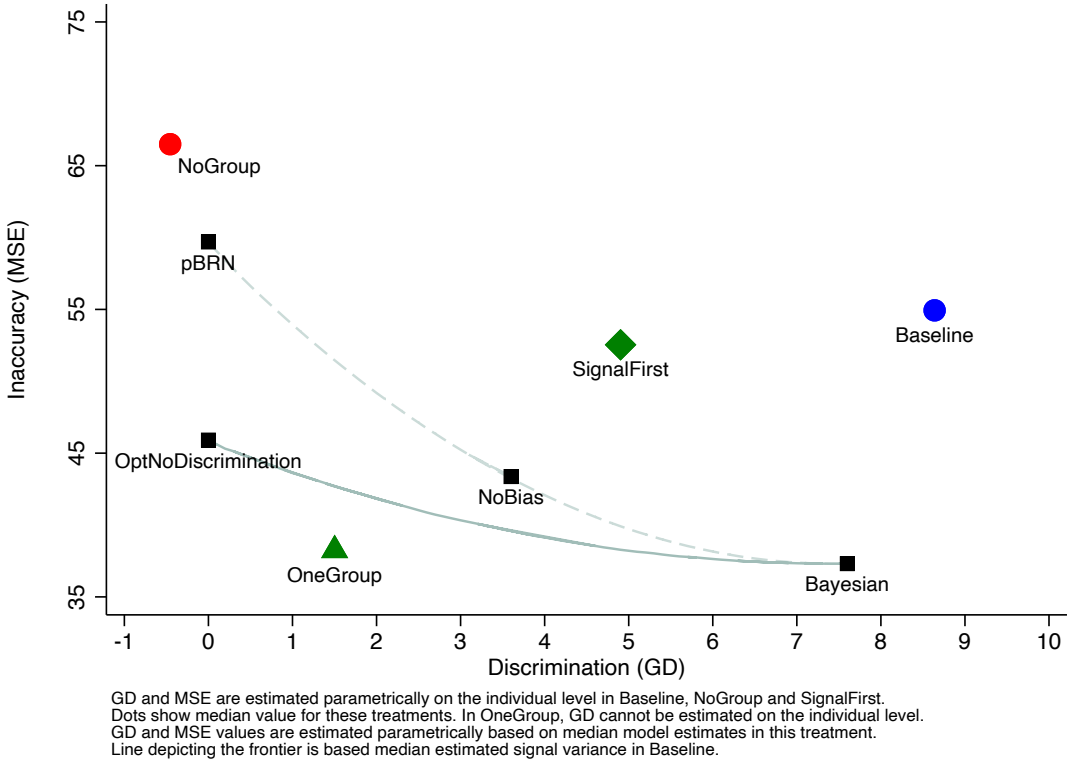


Figure 12: accuracy-Discrimination Tradeoff (Using Median Values)

F Learning

Table 6: Model Estimates for Rounds 1-37

	ω_l	B_l	ω_h	B_h	$\omega_h = \omega_l$	$B_h = B_l$	ξ_l^2	ξ_h^2	ω_l^{Bay}	ω_h^{Bay}	$\omega_l = \omega_l^{Bay}$	$\omega_h = \omega_h^{Bay}$
Baseline	0.83	-2.5	0.81	1.2		***	83	95	0.54	0.51	***	***
NoGroup	1.06	-0.70	1.03	-0.36			77	78	0.72	0.72	***	***
SignalFirst	0.81	-1.5	0.87	0.75		***	91	92	0.52	0.52	***	***
OneGroup	0.93	-0.63	0.93	-0.12			59	60	0.63	0.63	***	***

Subscript l (h) denotes low-mean (high-mean) group.

Stars denote the confidence level with which the hypothesis associated with the column can be rejected.

*** 1%, ** 5%, * 10% significance.

Table 7: Model Estimates for Rounds 38-75

	ω_l	B_l	ω_h	B_h	$\omega_h = \omega_l$	$B_h = B_l$	ξ_l^2	ξ_h^2	ω_l^{Bay}	ω_h^{Bay}	$\omega_l = \omega_l^{Bay}$	$\omega_h = \omega_h^{Bay}$
Baseline	0.80	-1.7	0.84	1.8		***	74	82	0.57	0.55	***	***
NoGroup	1.02	-0.29	0.98	-0.10			68	80	0.75	0.71	***	***
SignalFirst	0.82	-0.33	0.85	0.25			77	75	0.57	0.57	***	***
OneGroup	0.91	-0.16	0.84	0.40	*		47	60	0.68	0.63	***	***

Subscript l (h) denotes low-mean (high-mean) group.

Stars denote the confidence level with which the hypothesis associated with the column can be rejected.

*** 1%, ** 5%, * 10% significance.

For each subject, we estimate the share of early rounds (1-37) in which their assessment overshoots the actual value, i.e. either $\tilde{v} > v > \mu$ or $\tilde{v} < v < \mu$. We study how model estimates change from early to late rounds, contrasting subjects for whom this share is higher than the median (Table 8) in their treatment vs. others (Table 9).

Table 8: Change in Model Estimates by Treatment

	Baseline	NoGroup	SignalFirst	OneGroup
ω_h	0.955***	1.206***	0.998***	1.085***
	(0.0491)	(0.0376)	(0.0500)	(0.0447)
ω_l	-0.979***	-1.149***	-0.985***	-1.051***
	(0.0321)	(0.0306)	(0.0449)	(0.0353)
B_h	0.765	-0.193	1.595***	-0.571
	(0.546)	(0.652)	(0.615)	(1.108)
B_l	3.227***	0.820	1.730***	-0.462
	(0.452)	(0.575)	(0.619)	(0.617)
Change in ω_h in Rounds >37	-0.0194	-0.156**	-0.0834	-0.167***
	(0.0399)	(0.0615)	(0.0608)	(0.0490)
Change in ω_l in Rounds >37	0.126*	0.133***	0.145**	0.0981***
	(0.0678)	(0.0361)	(0.0734)	(0.0349)
Change in B_h in Rounds >37	0.722	-0.303	-0.556	-0.257
	(0.572)	(0.844)	(0.630)	(0.939)
Change in B_l in Rounds >37	-1.551***	0.0969	-1.810**	-0.359
	(0.545)	(0.831)	(0.711)	(0.593)
Observations	1650	2175	2025	1725

Bootstrapped standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Subjects with MSE less than or equal to 200.

Table 9: Change in Model Estimates by Treatment

	Baseline	NoGroup	SignalFirst	OneGroup
ω_h	0.701***	0.904***	0.759***	0.827***
	(0.0507)	(0.0498)	(0.0339)	(0.0411)
ω_l	-0.712***	-0.895***	-0.657***	-0.843***
	(0.0382)	(0.0507)	(0.0330)	(0.0279)
B_h	1.525***	-0.497	0.0984	-0.569
	(0.378)	(0.579)	(0.625)	(0.420)
B_l	1.926***	0.696	1.344***	0.0844
	(0.682)	(0.971)	(0.417)	(0.479)
Change in ω_h in Rounds >37	0.0647	0.00425	0.0251	-0.0440
	(0.0580)	(0.0579)	(0.0390)	(0.0456)
Change in ω_l in Rounds >37	-0.0549	-0.0515	-0.137***	-0.0381
	(0.0420)	(0.0528)	(0.0360)	(0.0234)
Change in B_h in Rounds >37	0.600	0.825	-0.540	0.847
	(0.422)	(0.783)	(0.454)	(0.816)
Change in B_l in Rounds >37	-0.118	-1.006	-0.665*	-0.199
	(0.594)	(1.138)	(0.393)	(0.394)
Observations	2400	2175	2400	2775

Bootstrapped standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Subjects with MSE less than or equal to 200.

G Additional Figures

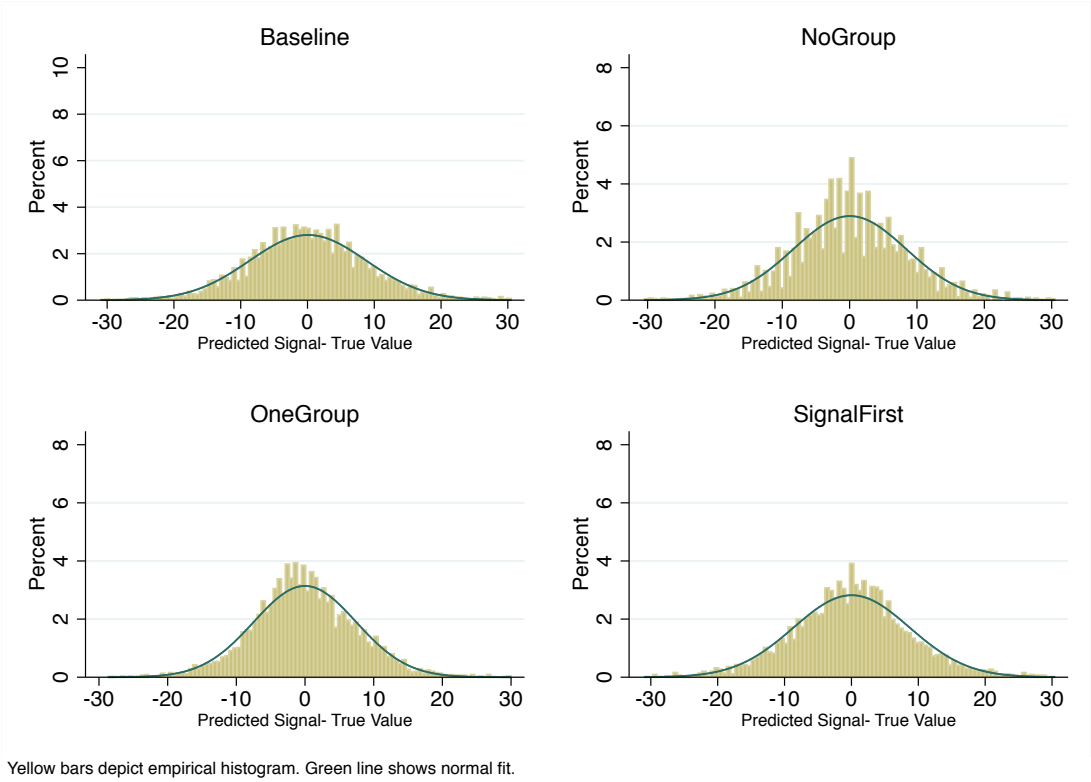


Figure 13: Distribution of Estimated Signal Error by Treatment

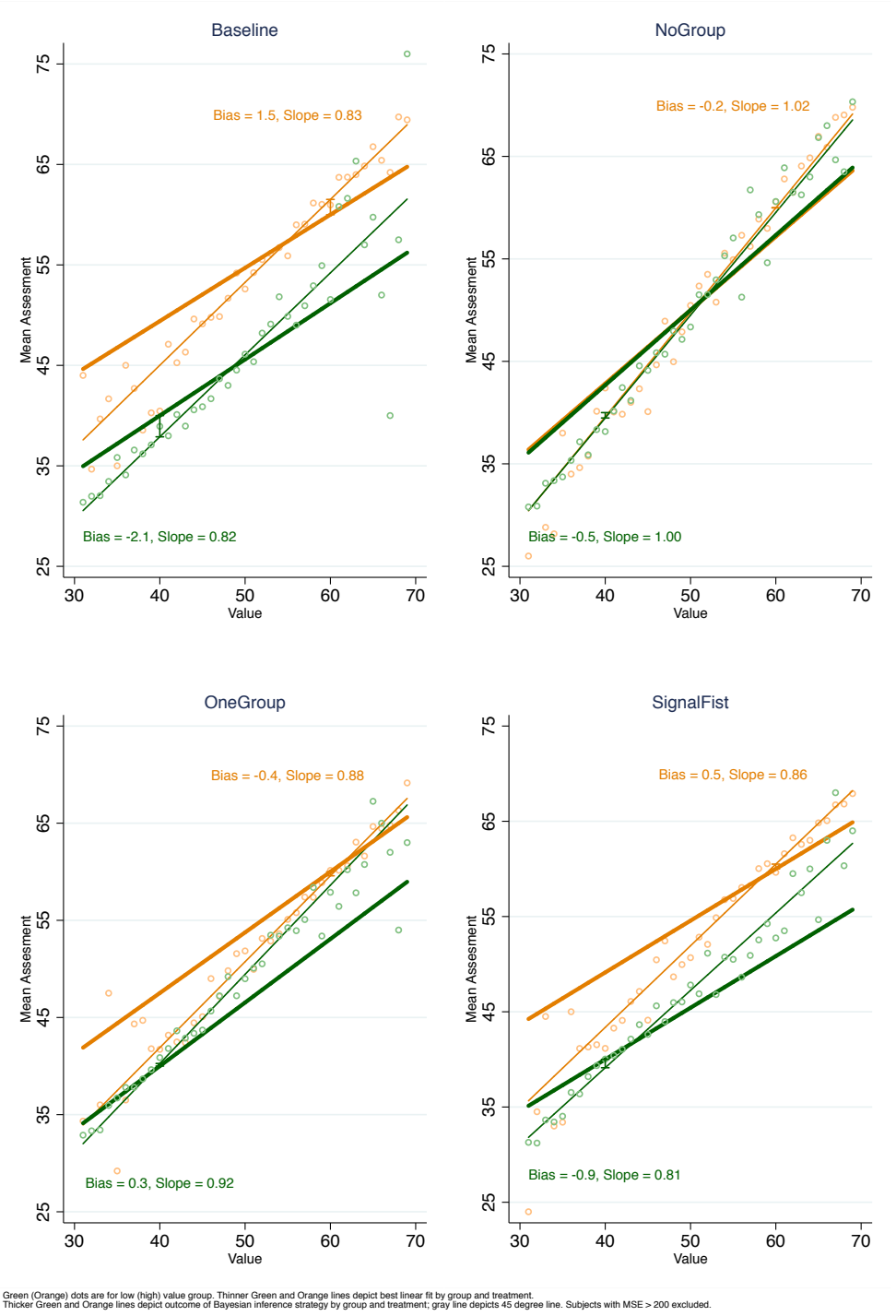


Figure 14: Average Actual and Bayesian Assessment by Value in Each Treatment

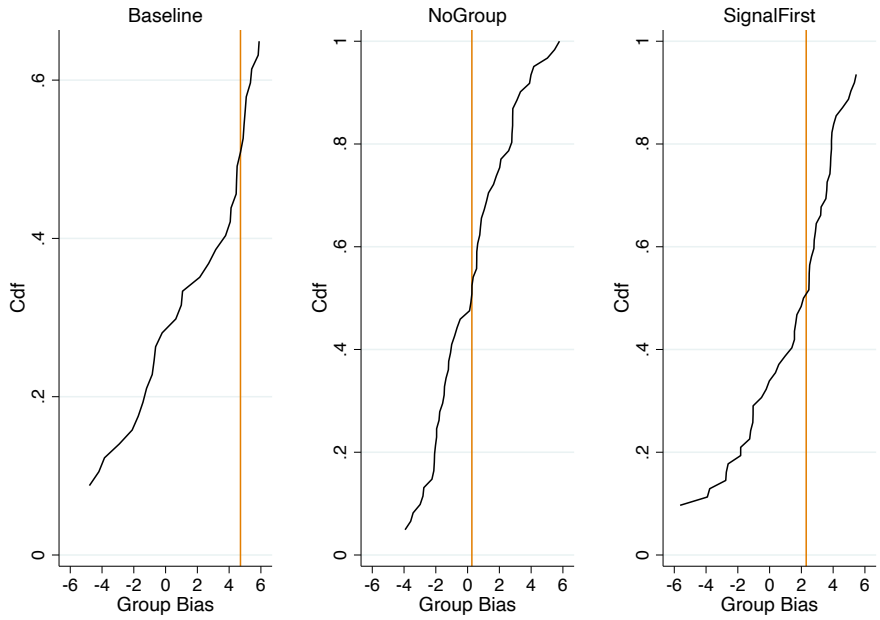


Figure 15: Estimates of Group Bias by Treatment *Notes: Group Bias denotes $B_h - B_l$. Vertical lines denote median values. OneGroup treatment is not included because group bias cannot be estimated on the individual level.*

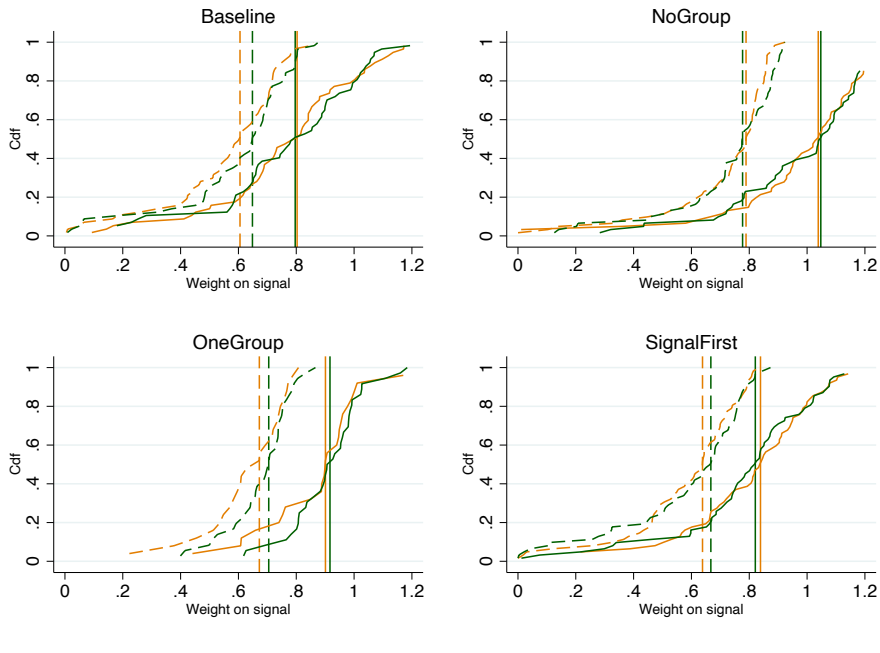
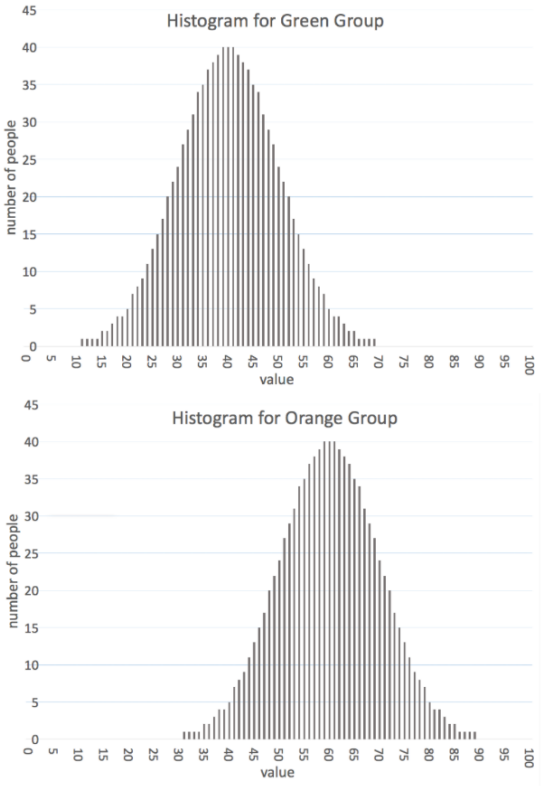


Figure 16: Estimates of Weight on Signal by Group and Treatment *Notes: Green (Orange) solid line represents estimated weight on signal for low-mean (high-mean) group. Green (Oramge) dashed line represents Bayesian weight on signal for low-mean (high-mean) group. Vertical lines denote median values.*

H Instructions for Baseline Treatment

The Data

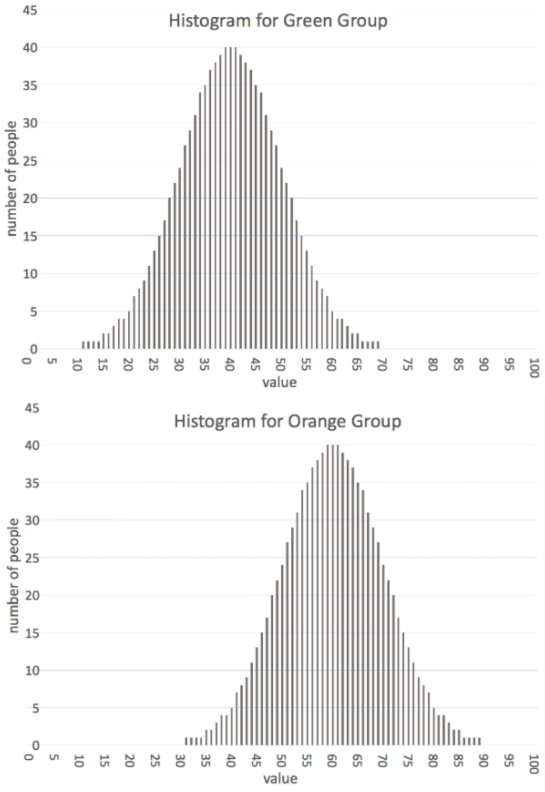
- We have **TWO GROUPS** of people. Each group consists of 1,000 people.
  - We'll refer to the first group as the Green group, and the second group as the Orange group.
- Each person is assigned a **NUMERICAL VALUE** from 0 to 100.
- The following figures, which are called histograms, depict the distribution of values for the Green group (top figure) and the Orange group (bottom figure).



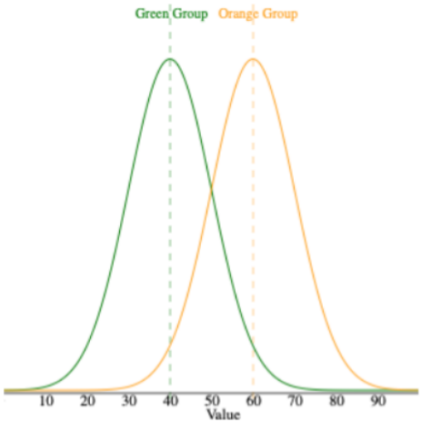
- These figures are read as follows. The horizontal axis shows each possible value, from 0 to 100 . For each possible value, the height of the corresponding bar depicts the number of people who have that value (as shown on the vertical axis).
- For example, there were 24 people in the Green group with value 50 and 24 people in the Orange group with value 50 in the data.

More on the Distributions

- Here is some additional information about these distributions



- The average value for people in the Green group is 40 and the average value for people in the Orange group is 60.
- Each of these two histograms are symmetric around their average. For example, the number of people in the Orange group with a value of 55 (5 less than the average of this group) are the same as the number of people in the Orange group with a value of 65 (5 higher than the average of this group). This is also equal to the number of people in the Green group with a value of 35 (5 less than the average of this group) and the number of people in the Green group with a value of 45 (5 more than the average of this group).
- The standard deviation is a measure of dispersion of a distribution around its average. The standard deviation in our data is the same for people in the Green group and for people in the Orange group, and it is given by 10. In particular, for each of the two distributions:
  - 42% of the people are within 5 points of the average
  - 71% of the people are within 10 points of the average
  - 88% of the people are within 15 points of the average
- Finally, for comparing the distributions, it is convenient to put the two normal distributions in the same figure, as shown below



- The experiment will rely on this data. You will be reminded of the distributions during the experiment.

Your Task

- This experiment consists of 75 rounds. In each round the following steps will take place:

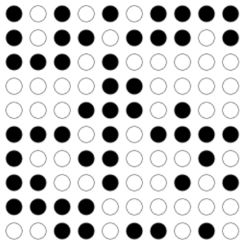
- **1.** The interface will **RANDOMLY** select, with equal chance, **ONE OF THE GROUPS**, either the Green group or the Orange group.

- **2.** The interface will randomly select, with equal chance, **ONE PERSON** out of the 1,000 people from the group selected in step 1. We call this person "the **SELECTED PERSON**" for this round.

- **3.** Your job will be to **GUESS** the actual value for this selected person. That is, you will guess this person's value [from 0 to 100].

- **4.** The interface will show you the **ACTUAL VALUE** of the selected person, but it will show this information in a way that is not easy to see perfectly. In particular:
  - The interface will show you a **SQUARE GRID** with a total of 100 balls, where each ball is either black or white.
  - The selected person's actual value is **EQUAL** to the **NUMBER OF BLACK BALLS** on the screen. The location of these black balls on the grid will be randomly selected.
  - You will only be shown the square grid with balls for a very **SHORT AMOUNT OF TIME**, so it is unlikely you will be able to guess exactly right the value of the selected person.

- Below is an example of a square grid showing the actual value of the selected person. In this example, there are 50 black balls (out of the 100 balls). Therefore, the actual value for this person is 50.



- In the experiment, this square grid will be shown to you for 0.25 seconds.

- **5.** Before seeing this, the interface will tell you the group (green or orange) to which the selected person belongs.

- **Your payoff:** The accuracy of your guesses will determine your chances of winning a prize of $20.
- In each round, you incur a loss that grows the further your guess is from the actual value of the person selected in the round. In particular, this loss is:
  - Loss = (the actual number of this person MINUS your guess)^2
- At the end of the experiment, your percentage chance of winning the prize will be equal to 100 minus your average loss from all the rounds.
- To maximize your chances of winning the prize, you should always **SUBMIT YOUR BEST GUESS** for the value.

- At the end of the round, the interface will **RETURN** this person to its group (green or orange) of 1,000 people, so it is possible that the interface will draw this person again in future rounds.

Summary

- The experiment consists of **75 ROUNDS**.
- In each round, **a GROUP** (green or orange) is randomly selected.
- Then **a PERSON** in the data from that group is randomly selected.
- Your task is to **GUESS** the actual value of the selected person. That is, you will guess this person's value [from 0 to 100].
- Before you make a guess, you will be **TOLD THE GROUP** to which the selected person belongs (green or orange).
- You will also see a **SQUARE GRID** with a total of 100 black and white balls, where the **NUMBER OF BLACK BALLS IS EQUAL TO THE ACTUAL VALUE** of the selected person.
- After the 75 rounds are finished, you will need to answer a few additional questions to complete the experiment.