Contrast-Biased Evaluation

Ignacio Esponda
UCSB

Ryan Oprea
UCSB

Sevgi Yuksel
UCSB

October 25, 2022

Abstract

We provide evidence for a previously undocumented cognitive bias that we call "contrast-biased evaluation" which is particularly relevant to settings of statistical discrimination. Experimental subjects distort their perception of information on individual group members and misinterpret such information to be more representative of the group the individual belongs to in contrast to a reference group. As such, the bias disappears when subjects either (i) receive information before learning of the individual's group or (ii) are prevented from contrasting different groups. We show that this bias can be easily estimated from appropriately constructed datasets and can be distinguished from previously documented inferential biases in the literature. Importantly, we document how removing the bias produces a kind of free lunch in reducing discrimination, making it possible to significantly reduce discrimination without lowering accuracy.

We would like to thank the editors, anonymous referees and numerous seminar participants for helpful comments. Esponda: iesponda@ucsb.edu; Oprea: roprea@gmail.com; Yuksel: sevgi.yuksel@ucsb.edu.

1 Introduction

In this paper we document a novel bias in inference that we call "contrast-biased evaluation": people tend to misperceive individual-level signals (e.g. evidence on an individual's behavior, credentials or performance) as though they are more representative of the group (e.g., a gender, nationality, race) the individual belongs to, relative to a reference group, than they really are. Because this perceptual distortion arises due to contrast effects between groups, it is especially relevant to problems of statistical discrimination which, by their nature, involve inferences about individuals from two or more groups with contrasting background distributions.

Unlike previously documented group contrast effects, contrast-biased evaluation is a distortion in the *perception* of new evidence rather than a distortion in prior beliefs about group members (as in e.g. Bordalo, Coffman, Gennaioli & Shleifer (2016)). Likewise, contrast-biased evaluation is distinct from previously documented inferential biases like over- or under-inference: it occurs not by causing agents to under-weigh the prior (as with base-rate neglect) or the signal (as with conservatism) in the updating of beliefs, but by causing agents to distort the perception of the signal itself. As such, contrast-biased evaluation distorts inference in a distinctive way, biasing perception of evidence in systematically opposite directions for contrasting groups, thereby generating an irrational discriminatory gap in inference which amplifies underlying group differences. We show that this bias can be removed simply by changing the context under which people evaluate evidence on individual group members, resulting in a kind of "free lunch" in reducing discrimination by making it possible to reduce discrimination without any cost to inferential accuracy.

In the first part of the paper (Section 2) we build a model of contrast-biased evaluation, showing how it can arise from *representativeness* (Kahneman & Tversky (1972*b*)), a well-documented mistake in which decision makers evaluate the likelihood of a characteristic or type by considering not only its prevalence within the relevant group (as a Bayesian would do), but also its *relative* prevalence in this group when contrasted to a reference group. Formally, we extend Bordalo et al. (2016) who model how this bias can distort the way people remember or represent the composition of groups, producing irrational stereotypes. In our model, the same mistake instead distorts the way people interpret *new evidence*, causing people to "look for" (or simply better notice) group-representative evidence when evaluating a member of a group. In standard statistical discrimination models with normal priors and signals, this mistake produces a simple additive bias in the evaluation of members of each group. We show in Section 3 that the resulting bias can easily be estimated from suitable data using simple regressions and distinguished from other previously documented

errors in inference such as base-rate neglect.

Our main contribution is a novel experiment (described in Section 4), designed to empirically identify this bias and study its properties. In our main Baseline treatment, we ask subjects to estimate the "type" (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 type (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 on their screen equal to the fictitious person's true type for a split second. The short exposure to the dots means that subjects cannot perfectly observe the true type and therefore receive only a noisy, subjective signal. While subjects are paid based on only the accuracy of their assessments, in our analysis we compare outcomes in terms of both accuracy and discrimination (difference in assessments between individuals of the same type from contrasting groups).

This experiment includes a number of design elements that allow us to cleanly measure the bias for the first time. First, unlike most inference experiments, we provide subjects with a subjective signal (rather than a numerically described statistical signal) that can be mis-perceived or otherwise mis-interpreted, creating scope for perceptual biases that are shut down by design in most previous work. Second, unlike previous experiments on contrast effects (e.g. Bordalo et al. (2016)), we study a setting designed deliberately to minimize misrepresentation or mis-remembering of the prior (i.e. of the groups themselves): we show subjects the prior distributions of types on their screens, give them significant training on how to interpret them and verify that they understand their properties. This allows us to focus on biases occurring due to misinterpretation of the signal rather than misrepresentation of the prior. Finally, we deliberately designed the experiment to be completely abstract, cleanly removing other, non-inferential sources of discrimination (such as animus or taste-based discrimination) that might otherwise confound our measurement of contrast-biased evaluation.

This experimental environment not only serves as a crisp testing ground for contrast-biased evaluation, but also as a natural setting for assessing the descriptive accuracy of models of statistical discrimination. This is no accident. Contrast-biased evaluation occurs precisely in the inference setting described by statistical discrimination models, in which a decision maker makes inferences about a member of one of several possible groups by combining individual and group level information. We therefore are able to use our dataset first to examine the hypothesis of contrast-biased evaluation and next to study the types of inefficiencies the resulting bias generates in statistical discrimination.

First (in Section 5), we document strong evidence of contrast-biased evaluation in our data. Subjects positively bias their estimates of members of the high-mean orange group and negatively bias their estimates of members of the low-mean green group. This bias is large relative to base-rate neglect (a classical error which we also find in our data) and is surprisingly stable, persisting even after dozens of repetitions of the task. Importantly, however, the bias disappears entirely in a diagnostic treatment (OneGroup) in which we ask subjects to make assessments about members of only one group (green or orange) throughout the experiment, indicating that the bias is a consequence of contrast effects. It also disappears with experience in a second treatment (SignalFirst) in which subjects observe the individual-level information, i.e. the signal, (presented as dots) *before* learning of the individual's group, indicating that the bias operates by distorting subjects' perception of the signal (rather than, e.g., by distorting their understanding of the prior). Thus the novel bias we observe in our Baseline treatment matches the empirical fingerprint of contrast-biased evaluation: it is a contrast-biased distortion of perception that is distinct from classical over- and under-inference and is highly robust to experience.

Next (in Section 6), we show that contrast-biased evaluation causes fundamental inefficiencies in statistical discrimination, producing opportunities to costlessly reduce discrimination. Subjects in our Baseline treatment follow the basic comparative statics of the statistical discrimination model, improving their accuracy (although only marginally) relative to a control treatment (NoGroup) in which subjects are not informed of the group, by "discriminating" between members of the two groups. However this discrimination is inefficient in the sense that outcomes are far from the "accuracy-discrimination frontier": given the information subjects are able to extract from perceptual signals, they could improve their accuracy while simultaneously discriminating less between groups simply by avoiding contrast-biased evaluation. Indeed, we show that simple cognitively-inspired interventions which remove the scope for contrast-biased evaluation – like specializing subjects to make assessment about members of only one group, or forcing subjects to evaluate individuals before learning of their group – produce striking reductions in discrimination and in some cases significant improvements in accuracy. To the degree that contrast-biased evaluation arises also in field applications, our experiment suggests that such cognitive interventions might be effective at costlessly reducing discrimination.

Our paper is related to several previous literatures. First, we contribute to a long empirical literature on biases in inference, reviewed in Benjamin (2019), that focuses on mistakes people make in *aggregating* distinct pieces of information (for example, a prior and a signal). The literature shows that subjects regularly fail to combine information in a Bayesian way, often making mistakes

in how much weight they put on different pieces of information in updating their beliefs. In our data, we find evidence of one classical example: base-rate neglect (over-inference), in which subjects put too little weight on their prior (and consequently too much on their signal),[1] perhaps driven by subjects' overconfidence in the quality of their subjective evaluations of the perceptual signal ("overprecision").[2] However, we show that contrast-biased evaluation is a distinct error – a distortion in the perception of information rather than a mistake in how subjects combine different pieces of information – that can be empirically distinguished from classical inferential errors like base-rate neglect.

Second, our work relates to a long literature on confirmation bias (see Nickerson (1998) and Klayman (1995) for reviews) – a mistake whereby prior beliefs distort decision makers' perception of new information. This line of research suggests that, in settings (like ours) in which signals are subjective and ambiguous, 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'.[3] However, the bias we document is conceptually and empirically distinct from confirmation bias: we do not find evidence that subjects are reluctant to update beliefs and form assessments far from the prior mean, a fingerprint of asymmetric updating that is typically associated with

[1]Base-rate neglect is one of the most frequently documented biases in updating (going back to Kahneman & Tversky (1972*a*)). See Benjamin, Bodoh-Creed & Rabin (2019) and Esponda, Vespa & Yuksel (2022) for recent perspectives on this bias. Conservatism (under-inference) is another commonly identified mistake in the literature and refers to subjects putting too little weight on the signal. Recently, Mobius, Niederle, Niehaus & Rosenblat (2022) provide evidence of this in a context in which subjects form beliefs about their own performance, and suggest that this bias could be due to ego utility.

[2]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. See Soll & Klayman (2004), Grubb (2009) and Grubb & Osborne (2015) for further documentation and discussion of overprecision.

[3]Kelley (1950), Darley & Gross (1983), and Lord, 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 (2022). This also connects to a literature that studies the importance of "first impressions," persistence of initial beliefs in the face 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". Similarly, a recent literature also documents how people distort their evaluation of information to justify self-serving decisions (Gneezy et al. 2020, Saccardo & Serra-Garcia 2022). These experiments study investment advise in which the advisor receives a commission that depends on their recommendation. Treatments vary the time at which advisors learn about their own incentives relative to evaluating investment options. The papers show that advice is more self-serving when advisors learn about their incentives before evaluating the investment options. This is a timing effect that we also find in our SignalFirst treatment, but it is driven by a very different behavioral mechanism.

confirmation bias.[4] Moreover, contrast-biased evaluation is distinct from confirmation bias because it is generated not by objective characteristics of the subject's prior (as with confirmation bias) but by the contrast between the prior and other distributions that are, strictly speaking, irrelevant to inference. As a result, the bias we document appears in our Baseline treatment but disappears in our OneGroup treatment – two treatments that should not produce different behaviors in standard models of confirmation bias. As a result, contrast-biased evaluation (unlike confirmation bias) is a perceptual distortion that we should expect to be particularly relevant to the context of statistical discrimination.

Third, we build on a theoretical and empirical literature studying how contrast effects and *representativeness* produce distorted beliefs. The representativeness heuristic (see Tversky & Kahneman (1983) for an early formulation) is a mental shortcut that simplifies probabilistic assessments about heterogeneous groups, and, as shown in recent literature, can lead to the formation of stereotypes. Stereotypes are beliefs that contain a "kernel of truth," as they are rooted in true differences between groups, but overweight the prevalence of 'representative' types in each population.[5] Bordalo et al. (2016) model and empirically demonstrate how representativeness-based recall can distort prior beliefs about different groups.[6] By contrast, we study a setting in which beliefs about group differences are tightly controlled and purposely highly salient. Our contribution is to show that representativeness and contrast effects can instead severely distort the way people perceive *new information* in updating contexts.[7] That is, building on the framework of Bordalo et al. (2016), we

[4]Recently, Sarsons (2017) used data on physicians referrals of surgeons to study whether gender influences the way new information is interpreted. The data reveals clear asymmetries in how physicians' beliefs about a female or male surgeon's ability changes with new information. However, it is not possible in that data to determine whether these asymmetries are driven by incorrect priors about men and women or by differences in how new information is interpreted.

[5]For example, because people in Florida are, on average, older than those in California, one might incorrectly overestimate the share of old people in Florida and the share of young people in California.

[6]There is also a related literature on sequential contrast effects studying how inferences can be impacted by recent (theoretically irrelevant) prior experiences. This is documented in speed dating Bhargava & Fisman (2014), investor behavior Hartzmark & Shue (2018), and candidate evaluation Radbruch & Schiprowski (2021), Kessler, Low & Shan (2022). In our experiment, there is similar scope for these kinds of sequential effects in all of our treatments and so such effects, if present, should not be affected by our treatments. The fact that the bias we measure is highly responsive to treatment and group identity suggests instead that we are measuring a bias generated by contrasts between prior distributions.

[7]Bordalo, Gennaioli & Shleifer (2018) present a model of forecast errors based on representiveness. Agents distort their beliefs by overweighting the likelihood of future outcomes that are more representative of recently observed news relative to past expectations. Such forecasts are excessively volatile, overreact to news, and are subject to predictable reversals.

show theoretically and empirically how contrasts between two groups with different distributions can impact the perception of a distinct piece of new information (rather than the perception of the group distributions themselves) during the updating process.

Fourth, our findings are relevant to an empirical and experimental literature examining how behavioral biases influence discrimination.[8] This particular literature has almost universally focused on documenting "irrational" statistical discrimination driven by *incorrect* prior beliefs about group differences.[9] Recent examples include Arnold, Dobbie & Yang (2018), who show that racial bias in bail decisions is possibly driven by judges' exaggerated beliefs about the relative danger of releasing black defendants, and Bohren, Haggag, Imas & Pope (2019), who outline implications of incorrect beliefs for identification of the source of discrimination (they also document discrimination against Americans partly based on wrong stereotypes).[10] This literature has also produced evidence that providing statistical information can decrease or eliminate statistical discrimination driven by inaccurate beliefs about groups (e.g., Reuben, Sapienza & Zingales (2014), Bohren et al. (2019)). Mengel & Campos Mercade (2021) use an experiment to highlight the implications of non-Bayesian updating (particularly conservatism) on discrimination.[11] While inaccurate prior beliefs and non-Bayesian updating are clearly important sources of discrimination in many settings, our contribution is to instead document biases in the perception of individual-level evidence.

The remainder of the paper is organized as follows. Section 2 presents our theoretical framework and Section 3 our empirical framework. Section 4 describes the experimental design. Section 5 presents results on contrast-biased evaluation and Section 6 documents its effects on statistical

[8]See Charles & Guryan (2011), Bertrand & Duflo (2017), Neumark (2018), and Bohren, Hull & Imas (2022) for recent reviews on the extensive literature studying discrimination, its potential causes, and policies intended to counteract it.

[9]One important exception is Bartoš, Bauer, Chytilová & Matějka (2016) which studies how allocation of costly attention affects discrimination. Our experimental design minimizes scope for costly attention mechanisms (because subjects observe signals for only a fraction of a second, giving them little control over attention) and theoretically equalizes incentives to attend to information across most of our treatments.

[10]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. More recently, focusing on discrimination against women in a hiring context, Coffman, Exley & Niederle (2021) identify beliefs about average group differences as a key driver of discrimination, Barron et al. (2022) document, in addition, *implicit* discrimination where gender-biased decisions are rationalized by adjusting beliefs about which signals are more predictive of performance.

[11]Like us, Mengel & Campos Mercade (2021) find that subjects behave as if their subjective perceptions are more accurate than they actually are. However, because Mengel & Campos Mercade (2021) provide subjective information about the prior rather than the signal, this results in conservatism rather than base-rate neglect.

discrimination. Section 7 concludes by discussing the implications of our results.

2 A model of contrast-biased evaluation

In this section we present a model of contrast-biased evaluation. We first describe the kind of inference task in which this new bias can arise. We then show how *representativeness*—defined via the contrast of different groups with distinct prior distributions—can *distort* the agent's perception of the signal and give rise to contrast-biased evaluation.

A decision maker (e.g. a hiring manager or admissions officer) must estimate the *type*, t, (e.g, a characteristic, score or measure of ability or fit) of an individual. The individual belongs to a *group*, g, (e.g. a gender, nationality or ethnic group), which is observed by the decision maker. The individual's type is drawn from a distribution $f(t|g)$ that depends on the group she belongs to. The decision maker observes a noisy *signal*, s, (e.g. a resume or an application package or the candidate's interview performance) of the individual's type, distributed according to $h(s|t,g)$. The decision maker's task is to form an estimate $\hat{t}$ of t that is accurate in the sense that it minimizes the expectation of $(\hat{t}-t)^2$.

By Bayes' rule, posterior beliefs on t given s and g are characterized by $f^p(t|s,g) = \frac{h(s|t,g)f(t|g)}{\int h(s|t,g)f(t|g)dt}$. An implication of this is that the optimal estimate, which corresponds to the posterior expected type, can be written as $\int f^p(t|s,g)tdt$.

Anticipating the setting of our experiment, we focus on a case in which (i) the type and signal distributions are Gaussian, (ii) the type distributions have the same variance, and (iii) the signal is an unbiased estimate of the type: $t \sim \mathcal{N}(\mu_g, \sigma^2)$, $s \sim \mathcal{N}(t, \xi_g^2)$. In this standard setting, the optimal estimate is a linear combination of the group's prior mean and the realized signal s,

$$\hat{t} = \omega_g \mu_g + (1-\omega_g)s, \tag{1}$$

where ω_g is the weight on the prior mean. For an agent who correctly perceives the prior and signal distributions, the optimal Bayesian weight on the prior mean is

$$\omega_g^{Bay} = \xi_g^2/(\xi_g^2+\sigma^2), \tag{2}$$

which is increasing in the relative precision of the prior vs. the signal distribution. As discussed in the introduction, the literature has documented several inferential biases that are characterized by a failure to attach Bayesian weights to priors and signals, including, notably, base-rate neglect, in

which the weight on the prior in equation (1) is too low, $\omega_g < \omega_g^{Bay}$.[12,13] By contrast, our interest is on biases that distort, not the way people combine existing information, but the way people perceive that information in the first place.

To model contrast-driven misperceptions, we draw from Gennaioli & Shleifer (2010)'s and Bordalo et al. (2016)'s work on stereotypes, which is in turn based on Kahneman and Tversky's representativeness heuristic.[14] Bordalo et al. (2016) show how a decision maker can form distorted beliefs about a target group by overweighting its representative types. Representative types are those that are observed more frequently in the target group relative to the reference group.

Formally, Bordalo et al. (2016)'s representativeness measure $R(t, g, -g) := \frac{f(t \mid g)}{f(t \mid -g)}$ captures how representative type t is of group g given reference group $-g$. The model describes how, through representativeness-based recall, the decision maker can distort beliefs about the distribution of types in group g, misrepresenting f by $\tilde{f}$ in the following way:

$$\tilde{f}(t \mid g) = \kappa f(t \mid g)(R(t, g, -g))^{\gamma^p},$$

where κ is a normalization factor and $\gamma^p \geq 0$ is a parameter that captures how susceptible the decision maker is to representativeness. When $\gamma^p = 0$, the agent does not suffer any distortions due to representativeness and $\tilde{f} = f$. But when $\gamma^p > 0$, the agent overweights the likelihood of types that are representative of group g (relative to $-g$) and $\tilde{f} \neq f$.

We are interested in a distinct but complementary possible effect of representativeness: that it distorts the evaluation of *new evidence* rather than beliefs about the distribution of types within a group – the perception of the noisy signal s instead of the prior f. Intuitively, the idea is that when the decision maker is faced with new evidence– when she reads a resume, observes behavior, or evaluates job performance – she is more likely to interpret the evidence to be more representative of

[12] The standard approach to studying deviations from Bayesian updating (since Grether (1980)) focuses on distortions in likelihood ratios. Two parameters capture potential deviations in how the prior likelihood ratio and the signal likelihood ratio is used to form a posterior likelihood ratio (α and β below, respectively):

$$\frac{p(t = t_1 \mid s)}{p(t = t_2 \mid s)} = \left(\frac{p(t_1)}{p(t_2)}\right)^{\alpha} \left(\frac{p(s \mid t = t_1)}{p(s \mid t = t_2)}\right)^{\beta},$$

Given signal s, the agent forms Bayesian beliefs about the relative likelihood of type t being equal to t_1 vs. t_2 when $\alpha = \beta = 1$. In Online Appendix B, we show that when $s \sim \mathcal{N}(t, \xi^2)$, for any value of (α, β), assessments consistent with the equation above are given by $\hat{t} = \omega\mu_g + (1 - \omega)s$, where $\omega = \frac{\alpha\xi^2}{\beta\sigma^2 + \alpha\xi^2}$. That is, while this framework can account for deviations from optimal weight ω, it doesn't allow for average bias in assessments with $\mathbb{E}[\hat{t}] \neq \mu_g$.

[13] Base-rate neglect could be driven by overconfidence in one's ability to read the signal, leading to a perceived signal variance $\hat{\xi}_g^2 < \xi_g^2$ and thus a perceived optimal weight $\omega_g = \hat{\xi}_g^2/(\hat{\xi_g}^2 + \sigma^2) < \omega_g^{Bay}$. This phenomenon is also known as overprecision in the literature (Moore & Healy 2008).

[14] See Kahneman & Tversky (1972*a*,*b*), and Tversky & Kahneman (1983).

the group the individual belongs to. In terms of the model, she misperceives the signal, s, as being more similar to what is representative of the individual's group relative to the reference group.

Let $y(s \mid g) := \int h(s \mid t, g) f(t \mid, g) dt$ be the distribution of the signals conditional on only g. We assume that the perception of the signal (conditional on true type) is distorted to be more representative of the group the individual belongs to (relative to a reference group $-g$). In particular, instead of drawing signals from the distribution $h(s \mid t, g)$, the agent draws signals from the distorted distribution

$$\tilde{h}(s \mid t, g) = \kappa h(s \mid t, g)(R(s, g, -g))^{\gamma^s}$$

where as before κ is a normalizing factor, $R(s, g, -g) = \frac{y(s \mid g)}{y(s \mid -g)}$ describes how representative signal s is of group g, and $\gamma^s \geq 0$ is a measure of the distortion in the signal distribution due to representativeness. Again, when $\gamma^s = 0$ the decision maker's evaluation of evidence is unaffected by representativeness and $\tilde{h} = h$. When $\gamma^s > 0$, the decision maker's evaluation of evidence is distorted such that she is more likely to observe signals that are representative of group g relative to group $-g$ and, as a result, $\tilde{h} \neq h$.

We call this bias "contrast-biased evaluation" because it (i) originates from *contrasts* between different groups, and (ii) is a distortion of the *evaluation* of new evidence rather than of the representation of the prior distribution.

To isolate the impact of contrast-biased evaluation, we assume that the decision maker distorts only the signal, and not the prior distribution.[15] We show in Online Appendix A that contrast-biased evaluation biases the mean of the signal distribution, such that the subjectively observed signal is $s \sim \mathcal{N}(t + \Delta_g, \xi_g^2)$ with

$$\Delta_g = \gamma^s \frac{\xi_g^2}{\xi_g^2 + \sigma^2}(\mu_g - \mu_{-g}),$$

where γ^s is a measure of how much the decision maker distorts her perception of the signal based on representativeness, ξ_g^2 denotes the variance of the signal, σ^2 denotes the variance of the prior, and μ_g and μ_{-g} denote the mean value of group g and reference group $-g$, respectively.[16]

Importantly, the decision maker is unaware of the fact that her perception of the prior signal is distorted, and therefore, when combining the subjective signal with prior information about

[15] In Online Appendix A, we consider a more general model that also allows for distortions in the prior.

[16] The impact of representativess as a distortion in means when using the normal distribution has also been shown in Bordalo, Gennaioli & Shleifer (2018).

the group, the decision maker uses equation (1), where the only difference is that s is now a signal realized from the distorted distribution of signals. By rewriting the distorted signal as $s = t+\Delta_g^s+\epsilon_g$, where $\epsilon \sim \mathcal{N}(0, \xi_g^2)$ is idiosyncratic signal noise, it follows from equation (1) that if the true type is t, then the predicted type is a random variable given by

$$\hat{t} = (1-\omega_g)\Delta_g^s + \omega_g\mu_g + (1-\omega_g)t + (1-\omega_g)\epsilon_g. \tag{3}$$

Equation (3) will form the basis of our empirical strategy, outlined in the next section.

3 Empirical strategy and hypotheses

Consider a dataset (matching our experiment below) in which we observe a decision maker's estimate $\hat{t}$, the true type t and the group identity g , but we do not observe the subjective signal of the decision maker. We begin by postulating a simple empirical model that is agnostic to the underlying model of decision making. We then link this empirical model to the model of contrast-biased evaluation described in Section 2.

For the empirical model, our only assumption is that the expected predicted type conditional on true type is linear in the true type, i.e., for each group g, $\mathbb{E}[\hat{t} \mid t]$ is linear in t.[17] This is a testable assumption (which we verify in our data) and it implies[18]

$$\hat{t} = B_g + \omega_g\mu_g + (1-\omega_g)t + \varepsilon_g, \tag{4}$$

where:

- B_g is a group bias term; more precisely, $B_g = \mathbb{E}[Bias(t)]$, where $Bias(t) \equiv \mathbb{E}[\hat{t} \mid t] - t$;
- ω_g is the weight on the group's prior mean μ_g vs. the true type t;

[17] This is true if and only if $(t, \hat{t})$ are jointly normally distributed for each group g. This is why we chose the variable under our control, t, to be (approximately) normally distributed in our experimental design, described below.

[18] To see how equation (4) is derived from the linearity assumption $\mathbb{E}[\hat{t} \mid t] = \alpha_g + \beta_g t$, note that this assumption implies that

$$\begin{aligned}\hat{t} &= \mathbb{E}[\hat{t} \mid t] + \varepsilon_g \\ &= \alpha_g + \beta t + \varepsilon_g,\end{aligned}$$

where $\mathbb{E}[\varepsilon_g \mid t] = 0$ by construction. Next, note that the bias term defined in the text, $B_g = \mathbb{E}[\alpha_g + (\beta_g - 1)t] = \alpha_g + (\beta_g - 1)\mu_g$, so that replacing $\alpha_g = B_g - (\beta_g - 1)\mu_g$ in the first equation we obtain $\hat{t} = B_g + (1-\beta_g)\mu_g + \beta_g t + \varepsilon_g$. To get to equation (4), we define $\omega_g := 1 - \beta_g$.

- ε_g is an error term satisfying $\mathbb{E}[\varepsilon_g \mid t] = 0$.

This theoretically-agnostic empirical model is useful because, as long as the linearity assumption is satisfied, it allows us to characterize agent behavior using two reduced-form parameters: (i) a group bias term and (ii) the weight on the prior group mean vs. the true type. These terms can be estimated using simple OLS regressions (details provided in Table 1).

The model of contrast-biased evaluation in the previous section provides a structural economic interpretation of these parameters, and, as we discuss in Section 6, it will also be convenient for conducting counterfactual analysis. Comparing equations (3) and (4), it follows that the group bias term captures the effects of contrast-biased evaluation, i.e.,

$$\begin{aligned} B_g &= (1-\omega_g)\Delta_g \\ &= (1-\omega_g)\gamma^s \frac{\xi_g^2}{\xi_g^2+\sigma^2}(\mu_g - \mu_{-g}). \end{aligned} \tag{5}$$

In addition, the weight ω_g on prior mean vs. true type is also the weight placed by the agent on the prior mean vs. her observed signal. Finally, comparison of the equations (3) and (4) reveal that the error term in the empirical model is a direct function of the signal error: $\varepsilon_g = (1-\omega_g)\epsilon_g$. This allows us to estimate the variance of the agent's observed signal, giving a measure of the precision of her information. Thus, from these empirical parameters we can directly derive structural estimates of Δ_g, ω_g and ξ_g^2, the key parameters of our theoretical model.

The model of contrast-biased evaluation makes a distinctive set of predictions about the group bias term that we use to derive the main hypotheses we seek to test in the experiment. First, as discussed previously, contrast-biased evaluation is differentiated from the classical inferential biases of over-inference (base-rate neglect) and under-inference (conservatism) in that it leads to the signal being perceived with a bias. By contrast, base-rate neglect and conservatism occur as distortions on the weight ω_g placed on the prior mean vs. the signal. Perhaps most importantly, contrast-driven evaluation makes distinctive predictions about the sign of the group bias term B_g: If the agent faces members of a relatively high mean group h and low mean group l (i.e., $\mu_h > \mu_l$), B_h will be a positive bias and B_l will be a negative bias.

H1 When the decision maker evaluates members of two groups with different means, the group bias term is positive for the high-mean group and negative for the low-mean group, i.e., $B_h > 0$ and $B_l < 0$.

Thus, estimates that reveal zero group bias terms, or that reveal bias terms that do not have

opposite signs in the right direction for the two groups would each serve as evidence against contrast-biased evaluation in the data. The latter pattern would be evidence of some other distortion, including for example distortions inherent in the signal technology, such as positively or negatively biased signals.

Second, contrast-biased evaluation, as suggested by our labeling of the bias, only occurs when groups with different underlying distributions are contrasted (implicitly) to each other. Thus, such a bias cannot arise in the absence of a second group. This allows us to state a second hypothesis.

H2 When the decision maker evaluates only one group g without access to information about a second group $-g$, $B_g = 0$.

Thus, if the signal bias terms B_g were similar in settings (i) in which subjects are aware of only one group and (ii) are aware of two groups, we would have evidence against contrast-biased evaluation. We would instead have evidence of some alternative bias in evaluation or perception.

Third, contrast-biased evaluation is not the only way that contrast effects might distort beliefs in inference settings like ours. For instance, as in Bordalo et al. (2016), contrast effects can cause decision makers to mis-characterize or mis-remember (i.e., stereotype) the characteristics of the groups themselves, producing a distortion in the prior distribution $f(t \mid g)$ rather than the signal distribution $h(s \mid t, g)$. Similarly, contrast effects might distort the recollection of evidence after it has been observed, producing a bias similar to ours but in the memory of evidence rather than in its perception. Indeed, as we show in Online Appendix A, these types of recollective distortions will generate biased estimates that satisfy the first two hypotheses. However, relative to these distortions, contrast-biased evaluation makes a distinctive prediction that separates it from these alternatives. In our formulation of the inference problem above, the decision makes learns the group identity of individuals *before* observing evidence directly about their type. This ordering should have no consequence if contrast effects operate by distorting the prior or the memory of evidence: under both such mechanisms we would expect to find biased estimates regardless of the timing of events. By contrast, because contrast-biased evaluation involves a misperception of the signal that is guided by group information, it should disappear in an alternative ordering in which the decision maker learns of the group *after* she has already evaluated evidence. We state this as a third hypothesis.

H3 When (i) an agent evaluates signal s prior to learning the individual's group g, $B_g = 0$ but (ii) when she evaluates the signal after learning the individual's group, $B_g \neq 0$.

Thus, if B_g were invariant to the order in which we present the group identity and the signal, we would have evidence against contrast-biased evaluation as modeled above. Results consistent with the first two hypotheses which also violate the third would be evidence in favor of contrast effects that operate through memory rather than perception, as discussed above.[19]

Finally, contrast-biased evaluation makes no prediction about the extent of over or under-inference, a bias that, as we have mentioned, is of a different nature. Importantly, all of our hypothesis hold for any weight that agents place on the prior mean vs. the signal, regardless of whether this weight is optimal. It is worth highlighting, however, that as Equation 5 indicates, the bias that results from contrast-biased evaluation increases with the weight on the signal, implying that base-rate neglect will exacerbate the impact of contrast-biased evaluation on estimates. Thus, while these biases are of a distinct nature, their interaction and separation will be important for explaining deviations from optimal behavior and predicting counterfactual behavior, as we demonstrate in Section 6.

4 Experimental Design

In our experiment subjects face a series of inference tasks. In each task, they are asked estimate the type of a fictitious individual by combining (i) statistical information about the type distribution associated the individual's group, and (ii) imperfect subjective information directly about the individual. Our goal is to cleanly measure contrast-biased evaluation in a simple and stylized design that has key features of many statistical discrimination settings, and to examine its implications for the tradeoff between accuracy and discrimination described by statistical discrimination models. As such, we designed the experiment around the model, empirical framework and hypotheses described in Sections 2 and 3. In Section 4.1, we present the inference task used in the experiment by describing our Baseline treatment. In Section 4.2, we describe how this task is used in our sessions and how it varies across treatments. In Section 4.3, we explain how this design allows us to achieve the empirical goals of our paper. Finally, in Section 4.4, we describe details on the implementation of the experiment.

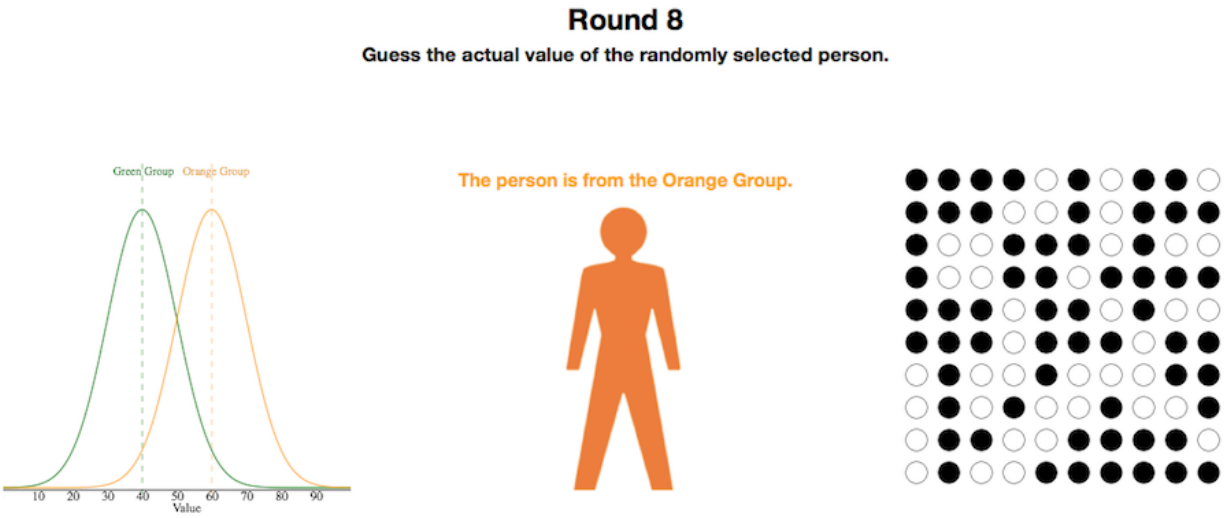


Figure 1: Screenshot from the Inference Task as Employed in Baseline. *Notes: The square grid showing the actual score of the person (number of black dots) disappears after 0.25 seconds.*

4.1 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") of "types" that are approximately normally distributed and differ only in their mean.[20] The "high-mean group" (called the "Orange group" in the experiment) has a mean type of 60, the "low-mean group" (the "Green group") has a mean type of 40 and both have a standard deviation of 10. The computer then randomly draws the "type" of one fictitious "person" from the distribution of the selected group.[21] The subject's task is to assess or guess the type of the person the computer has selected by inputting a number between 0 and 100, and they are paid based on the accuracy of this assessment (as described in Section 4.4).

In our Baseline treatment, subjects are tasked with combining two types of information to make their assessment. Figure 1 shows a screenshot of the task.[22]

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.

[19] As presented in the next section, our experimental design will deliberately minimize the possibility that contrast effects impact inference through these alternative channels.

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

[21] In the experimental interface "types" are referred to as "values".

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

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 always equal to the actual type 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 type 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.[23]

4.2 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 information and signal). 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 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.[24]

[23] Providing feedback about true type at the end of each round gives subjects an opportunity to recognize patterns in their mistakes, enabling them to potentially adjust their inference strategy.

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

OneGroup: Like Baseline, but instead of randomly observing members of the low-mean and high-mean groups over the course of the 75 tasks, subjects are assigned to always observe members of one group only (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.

4.3 Understanding the Design

We designed the experiment to identify contrast-biased evaluation empirically and examine its properties, guided by the theoretical and empirical considerations in Sections 2 and 3. Here we highlight how the design facilitates this identification.

First, we deliberately designed an inference problem with *subjective*, perceptual signals rather than the objective, statistical signals usually employed in inference experiments. We did this because contrast-biased evaluation involves distorted perception or evaluation of new evidence and so it is important that we include signals that are rich enough to be subjectively misinterpreted. This led us to show subjects a matrix of dots as a signal rather than, e.g., a number with a known distribution as in a typical inference experiment. Showing that signal quickly (in a quarter of a second) ensures that it will be perceived noisily and also that there is very little scope for using deliberate effort (á la a costly information acquisition model) to reduce this noise.

Second, on the other hand, we deliberately designed an inference problem with an objective, salient set of prior distributions. In particular we trained subjects on the properties of the prior distribution, quizzed them on these properties and visually reminded them of these distributions on the screen throughout decision making. We did this to minimize the possibility that contrast-effects can impact inference through other channels than contrast-biased evaluation. For instance, stereotypes, as documented in Bordalo et al. (2016), rely on misrepresentation of prior distributions, and we make this unlikely by making the priors objective and salient throughout choice. We did this not because we think stereotypes are unimportant in applications but rather to give us clearer and less confounded measurement of contrast-biased evaluation, our paper's main novelty.

Third, unlike most inference experiments, we studied a setting with normal (Gaussian) prior distributions (typically inference experiments focus on simpler binary settings). We did this to facilitate simple recovery of parameters using the empirical framework described in Section 3. Specifically (i) normality of types and (ii) normality of the signal error jointly guarantee that the conditional expectation $\mathbb{E}[\hat{t} \mid t]$ is linear in the type t. While only (i) is controllable by us in the

design, we verify below that subjects' inferences are consistent with the linearity assumption. This linearity allows us to cleanly estimate B_g, the bias from contrast-biased evaluation, separate it from over- and under-inference, and test that the properties predicted for it in Hypothesis 1 are satisfied.

Fourth, we included treatment OneGroup to test Hypothesis 2. As that hypothesis specifies, a distinctive implication of contrast-biased evaluation that distinguishes it from other perceptual distortions is that it should disappear in environments in which subjects lack a comparison group against which to assess representativeness. The OneGroup treatment implements this, allowing us to test whether the bias is an outgrowth of contrast effects.

Fifth, we included treatment SignalFirst to test Hypothesis 3. That hypothesis states that (unlike other contrast effects like stereotyping), contrast-biased evaluation should disappear when a subject learns the group *after* observing the signal. Contrast-driven evaluation requires the decision maker to distort her perception or interpretation of evidence, guided by representativeness, and this cannot happen if the agent does not know the group at the moment of she observes the signal. The SignalFirst treatment switches the order of evaluation relative to Baseline, allowing us to look for this distinctive comparative static.

Sixth, we included the NoGroup treatment to allow us to study the basic comparative static of statistical discrimination models and study how it is distorted by biases like contrast-biased evaluation and base-rate neglect.

Seventh, we included a large number of periods of repetition (75) of the inference task for two reasons. First, this allows us to study the persistence of contrast-biased evaluation and other biases in the presence of feedback. Specifically, we are able focus our analysis on the last half of the experiment after subjects already have dozens of periods of experience with the task, likely improving the external validity of our conclusions. Second, it allows us to estimate the parameters of the model described in Section 3 also at the individual subject level.

Finally, we made a deliberate decision to study a completely abstract problem involving fictional groups (orange and green) and individuals. We did this for identification reasons. Settings involving real group labels (e.g. 'men' and 'women') introduce a risk that unmeasured taste-based discrimination (or deeply seated stereotypes) might confound our measurements of contrast-biased evaluation. Contrast-biased evaluation is a purely cognitive bias, and as such can be cleanly measured in a purely abstract setting in which these confounding preferences and beliefs cannot interfere with measurement.

4.4 Implementation Details

We ran all treatments of the experiment simultaneously on Prolific on June 19th 2021 with 241 subjects from the US (57 in Baseline, 61 in NoGroup, 62 in SignalFirst and 61 in OneGroup). The 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.[25] Earnings for subjects ranged from $7.50 to $29.60 with average earnings of $15.90.

5 Evidence on Contrast-Driven Evaluation

In this section, we report estimates from our statistical model (Section 3) and use these estimates to test the hypothesis of contrast-biased evaluation. We report systematic evidence of contrast-biased evaluation (a new inferential error) and also of base-rate neglect (an old one).

In order to focus our analysis on the behavior of subjects experienced with the interface and the decision problem, we report results using the last half of the session (after period 37); we report results from the full dataset in Appendix H and flag differences (i.e., evidence of learning) where relevant in the text. 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. For measures computed at the aggregate level or derived from regression estimates (such as Δ_g or ξ_g^2), we use bootstrapping to make statistical statements. 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.[26] Doing this provides a

[25] 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.

[26] Specifically, we removed the 5% of subjects whose MSE (mean squared error of assessment) was greater than 200

more accurate portrait of the data's central tendencies, but does not change any of our qualitative conclusions.

Figure 2 gives us a first view of the data by plotting mean assessments as a function of true type, and corresponding linear fits for each group (the low group in green and the high group in orange) in each treatment. The size of the circles represent the relative frequency with which a specific type (within each group/treatment) was observed in the data. The plots reveal approximate linearity in assessments for both groups in all treatments. In Appendix F, we test and fail to reject the hypothesis that assessments are linear in type. The observation that average assessments are sharply increasing with the true type reassures us that subjects are attentive to the signals and willing to learn from them.

These raw views of the data preview our main findings. Key parameters of the empirical model—B and ω from equation (4)—can be visually identified in the graphs: the slope of the best linear fit corresponds to $1-\omega$ and the vertical distance between the best linear fit and the 45 degree line at $t = \mu$ visualizes average group bias B in assessments. In the NoGroup treatment, where statistical discrimination is impossible, linear fits are identical across groups and coincide with the 45-degree line, which suggests that there is no group bias and all the weight is placed on the subjective signal (i.e., perfect base-rate neglect). When subjects are shown group information (in the Baseline treatment) they show clear evidence of subjects engaging in statistical discrimination: conditional on type, members of each group are evaluated very differently (i.e., best linear fits are vertically separated for the two groups) and the slopes of linear fits are similar across the groups and clearly different from 1 (indicating positive ω values). Most importantly, this discrimination is biased: negative and positive deviations from the 45-degree line at means of 40 and 60 (measuring bias term, B) in the low-mean green and high-mean orange groups respectively, are consistent with hypothesis H1. These biases disappear altogether in OneGroup and SignalFirst as predicted by H2 and H3 respectively, suggesting the bias is driven by contrast-effects and is mostly a bias in the perception of the signal. The raw data thus matches the empirical finger-print of contrast-biased evaluation, as described in Section 3.

Table 1 reports estimates of OLS parameters (group bias B and weight on prior ω), structural parameters derived from these estimates (weight-adjusted bias Δ, variance of the subjective signal, ξ^2 and optimal weight on signal ω^{Bay}) obtained by pooling all observations as if they were coming

– 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.

from the same individual. These measures are estimated separately for each treatment and each group (high-mean and low-mean, differentiated by subscript $g \in \{l, h\}$). Below, we show that similar results hold when these parameters are estimated instead at the individual level.

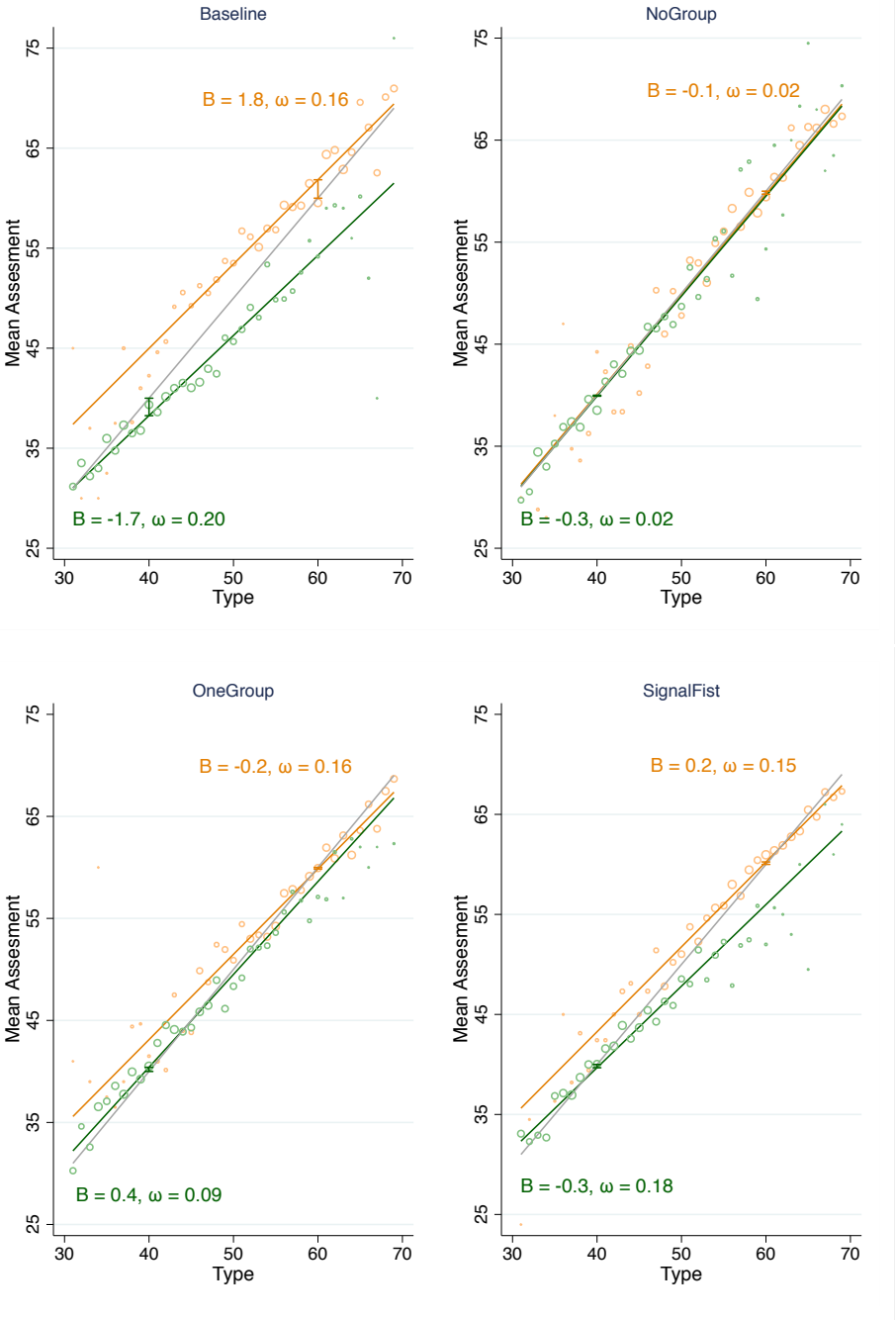
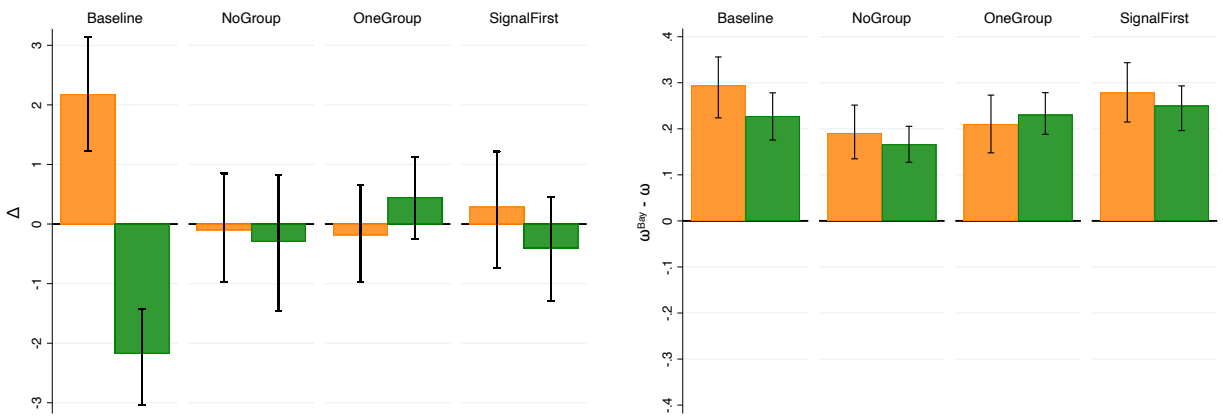


Figure 2: Average Assessment by Type in Each Treatment *Notes: The size of the circles represent the relative frequency with which a specific type (within each group/treatment) is observed in the data. Green (Orange) dots are for low-mean (high-mean) group. Green and Orange lines depict best linear fit by group and treatment; gray line depicts 45 degree line. Empirical strategy is described in Table 1. See Sections 2 and 3 for further discussion and interpretation of B and ω.*

Table 1: Model Estimates

	Baseline	NoGroup	SignalFirst	OneGroup
Regression estimates:				
ω_l	0.199***	0.0192	0.184***	0.0898***
	(0.0393)	(0.0330)	(0.0302)	(0.0218)
ω_h	0.157***	0.0196	0.151***	0.164***
	(0.0331)	(0.0360)	(0.0268)	(0.0339)
B_l	-1.743***	-0.289	-0.327	0.408
	(0.314)	(0.589)	(0.368)	(0.317)
B_h	1.836***	-0.0990	0.245	-0.156
	(0.402)	(0.466)	(0.404)	(0.377)
Estimates derived from ω and B:				
Δ_l	-2.17***	-0.30	-0.40	0.45
Δ_h	2.18***	-0.10	0.30	-0.19
ξ_l^2	74***	68***	77***	47***
ξ_h^2	82***	80***	75***	60***
ω_l^{Bay}	0.43***	0.25***	0.43***	0.32***
ω_h^{Bay}	0.45***	0.29***	0.43***	0.37***
Tests:				
H_0: $\omega_l = \omega_h$	0.318	0.995	0.341	0.072
H_0: $B_l = B_h$	0.000	0.741	0.317	0.257
H_0: $\Delta_l = \Delta_h$	0.000	0.350	0.164	0.134
H_0: $\omega_l = \omega_l^{Bay}$	0.000	0.000	0.000	0.000
H_0: $\omega_h = \omega_h^{Bay}$	0.000	0.000	0.000	0.000
Observations	2052	2204	2242	2280

Notes: For each treatment and group $g \in \{l, h\}$, we estimate B_g and ω_g using OLS on the following specification: $y_{g,i} = B_g + \omega_g x_{g,i} + \varepsilon_{g,i}$, where i denotes each distinct observation, $y_{g,i} \equiv \hat{t}_{g,i} - t_{g,i}$, and $x_{g,i} \equiv \mu_g - t_{g,i}$. This specification is derived by subtracting $t_{g,i}$ from both sides of equation (4). Given estimates for B_g and ω_g, we back out Δ_g using equation (5) and estimate ξ_g^2 by identifying the error associated with the signal using $\varepsilon_{i,g} = (1-\omega_g)\epsilon_{i,g}$ and then taking the sample average of $\epsilon_{i,g}^2$. Given ξ_g^2, ω_g^{Bay} is derived from equation (2). See Sections 2 and 3 for further discussion. Standard errors (clustered at the subject level) are reported in parentheses. ***1%, **5%, *10% significance. Rows on tests report p-value associated with test of each hypothesis. Statistical assessments on estimates derived from ω and B use bootstrapping.



(a) Contrast-Biased Evalution

(b) Base-Rate Neglect

Figure 3: Estimates of Contrast-Driven Evaluation (Δ) and Base-Rate Neglect ($\omega^{Bay} - \omega$) by Treatment and Group *Notes: Empirical strategy is described in Table 1. See Sections 2 and 3 for further discussion. Vertical lines denote 95 percent confidence intervals.*

Estimates of the group bias terms B_l and B_h in Table 1, confirm that in the Baseline treatment (i) the biases in each group (h and l) are large, significant and symmetric, (ii) of opposite sign and therefore (iii) widely separated, producing a large (and statistically significant) irrational discriminatory gap in the assessment of individuals from the two groups. This pattern of group-driven bias is characteristic of contrast-biased evaluation, producing support for hypothesis H1.

Relative to the bias described by the model in Section 2, these group bias terms are attenuated by a factor of $(1 - \omega)$ as reflected in Equation 5. Table 1 corrects for this to produce structural estimates of the key contrast-biased evaluation parameter from the model: Δ_h (weight-adjusted bias for the high-mean group, in orange) and Δ_l (weight-adjusted bias for the low-mean group, in green). We visualize these parameters in the left panel of Figure 3. Again, we find that (i) the two biases are large and symmetric, (ii) of opposite sign and therefore (iii) widely separated. These results suggest that when faced with members of the low-mean group, the signals subjects infer are downward biased; when faced with members of the high-mean group they are instead upward biased. As Table 1 shows, for the Baseline treatment, (i) each of these bias terms is significantly different from zero and (ii) the difference between the two, $\Delta_h - \Delta_l$ is as well.

Result 1. *In the Baseline treatment, subjects' estimates are (i) significantly biased downwards for the low-mean group and (ii) biased upwards for the high-mean group producing (iii) a significant irrational discriminatory gap in assessments. The results are thus consistent with Hypothesis H1.*

In contrast to the Baseline treatment, Table 1 and Figure 3 shows that the bias (whether

measured by reduced form parameter B or structural parameter Δ) virtually disappears when subjects are unable to contrast the two groups with one another (the OneGroup treatment). As the error bars in Figure 3 suggest and formal tests in Table 1 confirm, in the OneGroup treatment Δ_h, Δ_l and their difference, $\Delta_h - \Delta_l$ are all statistically indistinguishable from zero (as are B_h, B_l and $B_h - B_l$). This suggests that the bias we estimate in the Baseline treatment is a consequence of a contrast effect between groups. The disappearance of the bias when contrast effects are removed is another characteristic feature of contrast-biased evaluation and provides support for hypothesis H2.

Result 2. *In the OneGroup treatment, the bias measured in the Baseline treatment – and the discriminatory gap in assessments it produces – disappear. The results are thus consistent with H2: the bias measured in the Baseline treatment is an outgrowth of contrast effects.*

Similarly, Figure 3 and Table 1 show that the bias terms (again whether measured by B or Δ) disappear when subjects observe the signal prior to learning the group (the SignalFirst treatment). This suggests that the bias we estimate in the Baseline treatment operates by distorting subjects' evaluation of perceptual evidence (rather than by distorting memory of the prior or signal) The sharp disappearance of the bias term in the SignalFirst treatment comes about with experience. Estimates using *all* 75 periods (see Appendix H for detailed analysis) shows a sharp attenuation of the bias term in SignalFirst relative to Baseline but not a complete disappearance. This suggests that subjects may suffer also from distortions in memory of the signal (at the point where inferences are made) or the prior in early rounds of SignalFirst, but such distortions (unlike the distortion in signal perception) is transitory and mostly disappears with experience. The data thus matches the third main characteristic of contrast-driven evaluation, articulated in H3.

Result 3. *In the SignalFirst treatment, the bias measured in the Baseline treatment – and the discriminatory gap in assessments it produces – mostly disappear with experience. The results are thus consistent with H3: for experienced subjects, the bias works almost entirely by distorting subjects' perception or evaluation of the signal.*

Together, then, our results provide strong evidence of contrast-biased evaluation: the bias we measure matches the distinctive empirical fingerprint of contrast-biased evaluation described in Section 3. In addition to this new bias, Table 1 also provides evidence of a classical error that also distorts assessments in our data: base-rate neglect. In the Table we provide estimates of the weight, ω subjects place on the prior mean μ, (relative to the signal s) for the low/green group (ω_l) and the high/orange group (ω_h). To complement, the table also gives the weight a Bayesian would place on the prior mean (ω_l^{Bay} and ω_h^{Bay}) given the estimated variance of subjects' perception of

the signal (ξ^2). The results in Table 1 reveal that subjects put significantly more weight on their subjective signals (and less weight on the prior) than a Bayesian would given the noisiness of their evaluations. In other words, subjects act as if their perception of the signal is more accurate than it really is.

The right panel of Figure 3 plots the difference between $\omega^{Bay} - \omega$ for each group and treatment. Positive values show evidence of base-rate neglect (over-inference or under-weighting of the prior) while negative values show evidence of conservatism (under-inference or over-weighting of the prior). The results are universally positive and therefore show clear evidence of base-rate neglect. Moreover, in contrast to our results on contrast-biased evaluation in the left panel, estimates of $\omega^{Bay} - \omega$ in the right panel reveal a remarkably consistent level of base-rate neglect across groups and treatments.

Result 4. *In addition to contrast-biased evaluation, there is significant evidence of base-rate neglect for both groups in all four treatments. Unlike with contrast-biased evaluation, the degree of base-rate neglect is mostly unaffected by group or treatment.*

We make three further observations that help us to interpret and contextualize these results.

First, the NoGroup treatment (in which subjects are forced to make assessments without being informed of the group) serves as a useful sanity check on our estimates. The fact that contrast-driven evaluation does not arise in NoGroup (neither B_h, B_l, Δ_h or Δ_l are statistically different from zero) reassures us that the technology we use to represent signals (a grid with dots shown for .25 seconds) is unbiased, and that the existence of the bias we observe in Baseline is indeed driven by knowledge of group identity in a setting where there are two groups (rather than, for instance, contrast effects that arise by comparison of signals from sequentially observed candidates). Indeed, estimates of ω from Table 1 show that subjects in NoGroup ignore the prior entirely – even without group information, subjects could significantly increase accuracy by putting some weight on the global mean of 50.[27]. However, the level of base-rate neglect ($\omega^{Bay} - \omega$) is similar to the other three treatments.

Second, Table 1 also reports values of the signal variance – the noisiness with which we estimate subjects perceived the signal. We find no differences in estimated variances for the Baseline,

[27]For the NoGroup treatment, we also estimate a linear model where subjects put weight on the signal and on the average mean of 50. The Bayesian 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 prior vs. the signal depends on the signal itself. However, the optimal prediction (as shown in Online Appendix D) and the aggregate assessment strategy of our subjects in our data (as shown in Online Appendix F) are both approximately linear.

NoGroup, and SignalFirst treatments, and the estimated variances are similar in magnitude to the population variance that we picked for the experiment.[28] However, we find one surprising result from our analysis of signal variances: a significant decrease in signal variance in OneGroup relative to Baseline ($p = 0.001$ for $g = l$ and $p = 0.038$ for $g = h$), suggesting that when subjects are restricted to make inferences about only one group they perceive signals about members of that group not only without bias but also with less noise. We discuss this finding in more detail in Section 6.[29]

Finally, we re-conducted our estimation at the individual subject level (i.e., we estimated the full raft of parameters for each subject individually) and obtained similar results. Figure 4 plots CDFs of individual-level Δ_g estimates (in Online Appendix G we provide similar CDFs of $\omega^{Bay} - \omega$ estimates); results are qualitatively similar if we replace Δ with B in this exercise. Although there is significant heterogeneity in estimates across subjects, (i) Δ_h strongly first-order stochastically dominates Δ_l in the Baseline treatment and (ii) the two distributions converge significantly in SignalFirst and especially OneGroup. Equality of the distributions can be rejected by a Kolmogorov-Smirnov test in Baseline and SignalFirst with $p = 0.000$ and $p = 0.019$, respectively. This is not the case for OneGroup ($p = 0.868$) or NoGroup ($p = 0.929$). Differences between Δ_h and Δ_l in SignalFirst, while much smaller than in Baseline, are distinguishable from zero. This suggests that while contrast-biased evaluation is responsible for most of the bias in Baseline, at least for some subjects, contrast effects also bias assessments to a smaller degree even when signals are observed prior to knowledge of group identity.[30] By contrast, Figure 9 in Online Appendix shows that $\omega^{Bay} - \omega$ is overwhelmingly positive (indicating base-rate neglect) for the vast majority of subjects, varying little across groups and treatments.[31]

[28] These findings reassure us on 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 significantly lower variance for the subjective signal in the NoGroup treatment, where group information is not provided and the signal is therefore more valuable.

[29] The biggest effect of learning in our dataset is a reduction in signal variance over time. Relative to the last-half sample studied here, estimates of ξ^2 in the full sample are larger. Intuitively, subjects learn to form less noisy (but not less biased) perceptions of the signal over time.

[30] In Appendix G we also include the cumulative distribution of $\Delta_h - \Delta_l$ estimated on the individual level. For 75 percent of subjects in Baseline $\Delta_h > \Delta_l$. This decreases to 63 percent in SignalFirst and 49 percent in NoGroup.

[31] In Appendix G, we present the counterparts of Table 1 and Figure 5, reporting median-values from individual-level estimates.

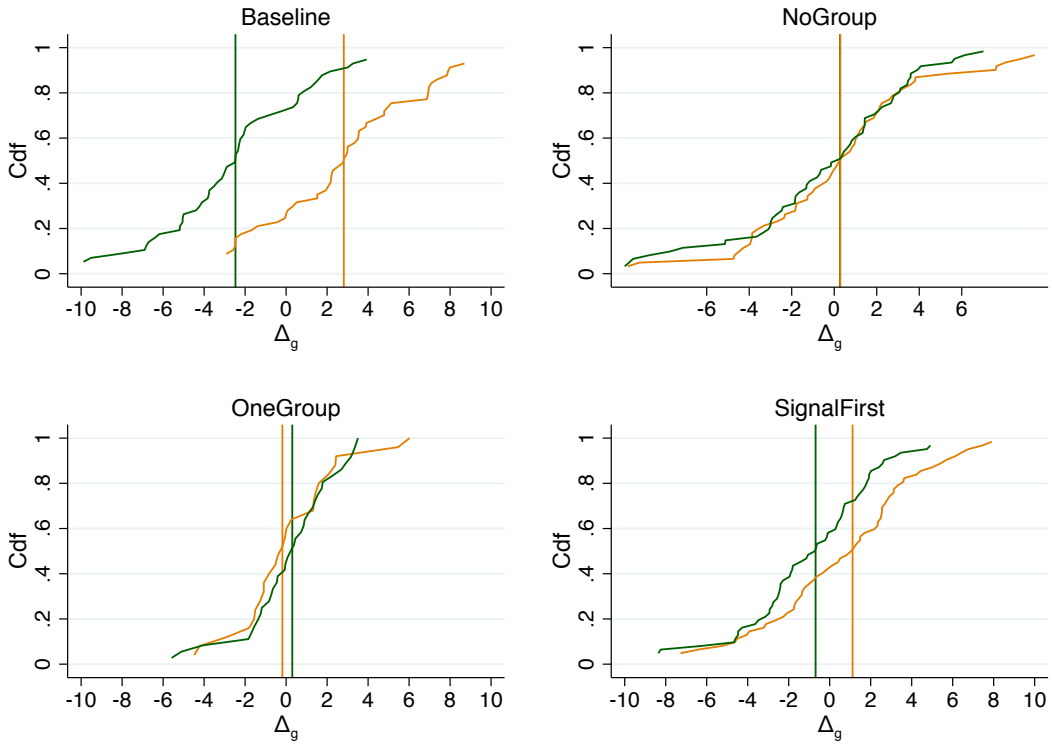


Figure 4: Estimates of Contrast-Driven Evaluation Δ by Group and Treatment *Notes: Green (Orange) line represents estimated bias parameter for low-mean (high-mean) group. Vertical lines denote median value. Empirical strategy is described in Table 1. See Sections 2 and 3 for further discussion.*

6 The Accuracy-Discrimination Tradeoff

A key consequence of contrast-biased evaluation is that it produces an irrational amplification of the gap in decision makers' mean assessments between two groups. While this bias has implications in many contexts, in this section we focus on arguably the most important application: statistical discrimination, where contrasts between groups seem especially salient.

In rational statistical discrimination, a decision maker optimally conditions her estimates on the group identity of the individual being evaluated to maximize accuracy. An implication of this is that any deviation from this benchmark (for example, any reduction in discrimination) has an accuracy cost. In this section, we demonstrate how contrast-biased evaluation generates an inefficiency in assessments that instead produces scope for simultaneous improvement in both accuracy and discrimination.

First, we define formal measures of accuracy and discrimination. We will measure (in)accuracy using mean squared error (MSE), which is also the measure we used to incentivize subjects in the experiment.[32] We will measure discrimination by looking at group differences (GD), defined as the expected difference in assessments between members of the high-mean and low-mean groups when they are of the same actual type.[33] GD measures the extent to which individuals from different groups who are otherwise identical are assessed differently on average. 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).[34]

Assuming that the signal variance ξ^2 and the weight on the prior mean ω do not differ by group (which, as shown in the previous section, is roughly true in the data of our experiment), these measures simplify to:

$$MSE = \omega^2\sigma^2 + (1-\omega^2)(\xi^2 + \frac{\Delta_l^2 + \Delta_h^2}{2}) \tag{6}$$

and

$$GD = \omega(\mu_h - \mu_l) + (1-\omega)(\Delta_h - \Delta_l). \tag{7}$$

Expressions for these measures clearly demonstrate that contrast-biased evaluation is costly in terms of *both* accuracy and discrimination, since $\Delta_l \neq 0$, $\Delta_h \neq 0$ and $\Delta_h - \Delta_l > 0$. For MSE, Δ_l and Δ_h have the same effect as an increase in the signal variance, and the impact of this bias increases with the weight placed on the signal. For GD, the first component, $\omega(\mu_h - \omega_l)$, is the standard term capturing the fact that a decision maker, for the sake of accuracy, partly relies on the prior means in forming assessments, thus leading to discriminatory outcomes. But the second component, $(1-\omega)(\Delta_h - \Delta_l)$, is an irrational amplification of the discriminatory gap due to contrast-biased evaluation.

[32] Formally, $MSE := \frac{1}{2}\mathbb{E}\left[(\hat{t}_h - t_h)^2\right] + \frac{1}{2}\mathbb{E}\left[(\hat{t}_l - t_l)^2\right]$. This measure computes the expected squared distance between assessments and actual scores, given the assumption that observations are equally likely be from the low-mean or high-mean groups.

[33] Formally, $GD := \int \mathbb{E}\left[\hat{t}_h - \hat{t}_l \mid t_h = t_l = t\right] f(t)dt$, where f is the mixture distribution of the low-mean and high-mean groups. In practice, the distribution f will not end up mattering much because the weights ω_l, ω_h are similar across groups implying the integrand to change little with t .

[34] In Appendix C, we discuss other measures of discrimination and specifically relate GD to measures used in the machine learning literature. In Appendix E (Tables 3 to 5), to contextualize treatment differences further, we report the implied likelihood of correctly identifying the higher type among two randomly selected individuals in each treatment.

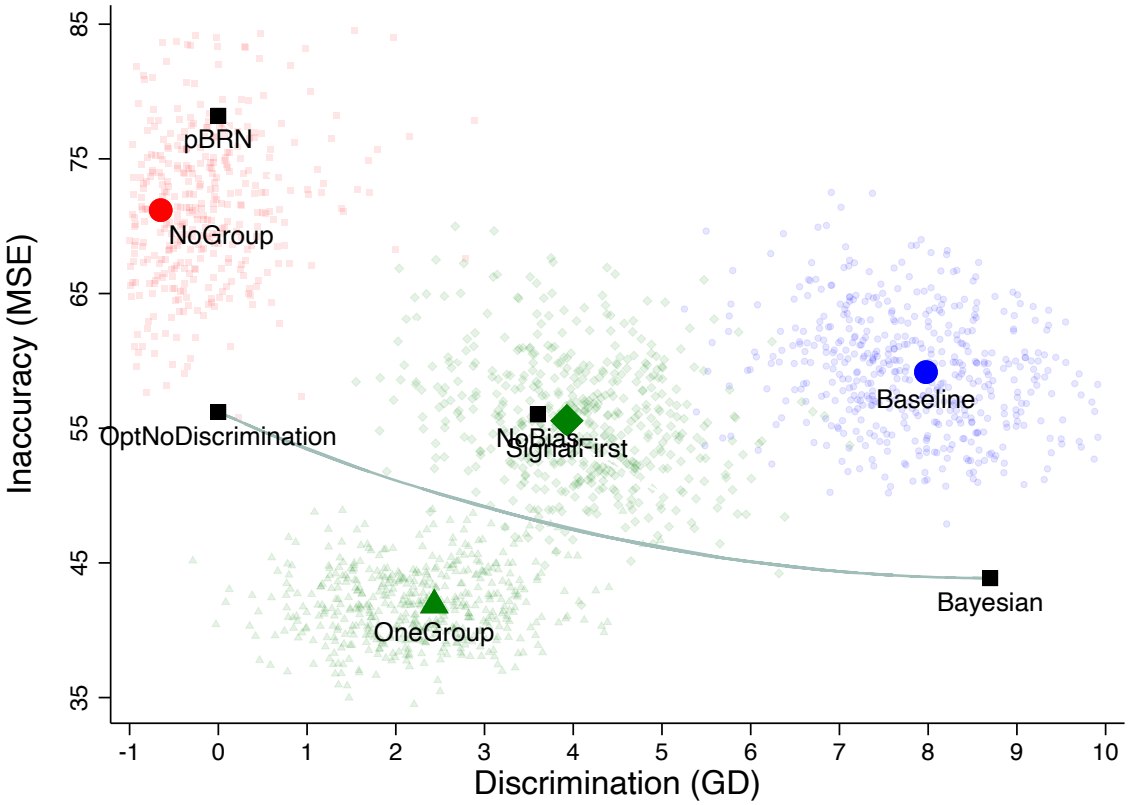


Figure 5: Mean Squared Error and Group Difference by Treatment *Notes: Translucent dots show bootstrapped values. Solid line is the accuracy-discrimination frontier, pBRN depicts the perfect base-rate neglect counterfactual, OptNoDiscrimination depicts the counterfactual with highest accuracy (at zero 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.*

Figure 5 plots the mean inaccuracy of subjects' assessments (mean squared error) on the y-axis against the mean level of discrimination between members of the two groups (group difference in assessment) on the x-axis for each of our four treatments.[35] Translucent dots show bootstrapped values to visualize the degree of variation in the data.

Over this, we overlay several important theoretically-driven counterfactuals. All of these theoretical benchmarks set biases $\Delta_h = \Delta_l = 0$ and are calculated using the estimated signal variance

[35]We use non-parametric estimates of MSE and GD when comparing these measures across treatments, but use parametric estimates to construct counterfactual benchmarks. 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 t 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 $t_h = t_l = t$, and finally aggregate for different values of t using the mixed distribution of t.

in the Baseline treatment (in addition to the primitives implemented in the experiment).[36]

- **Bayesian**: represents the highest accuracy level (lowest MSE) outcome achievable and the level of discrimination (GD) associated with this outcome. The inferential strategy producing this benchmark is described by Equations 1 and 2.
- **OptNoDiscrimination**: represents the highest accuracy level (lowest MSE) outcome achievable subject to a zero discrimination constraint (i.e. GD=0).
- **pBRN**: represents the outcome produced by an inferential strategy that only uses the signal (i.e. $\omega = 0$ in Equation 1), ignoring the prior. Because this strategy suffers from perfect base-rate neglect, it also produces zero discrimination. However, the contrast with the OptNoDiscrimination benchmark reveals the inefficiency associated with this strategy.[37]
- **NoBias**: represents the outcome corresponding to an inferential strategy that places the same weight on the prior vs. the signal as estimated for the Baseline treatment. Because Δ_h and Δ_l are assumed to be zero, this corresponds to a benchmark in which we shut down contrast-biased evaluation but keep base-rate neglect (as measured in the Baseline treatment).

Bridging the Bayesian and OptNoDiscrimination counterfactuals is the *accuracy-discrimination* frontier, showing the locus of MSE-GD pairs that could be achieved by a social planner whose objective is to minimize $\chi GD + (1-\chi)MSE$ for some $\chi \in [0,1]$.[38] The Bayesian counterfactual corresponds to the solution when $\chi = 0$, the OptNoDiscrimination counterfactual corresponds to the solution when $\chi = 1$, and the accuracy-discrimination frontier depicts solutions for intermediate values of χ.

The Figure shows, first, that subjects engage in statistical discrimination in our Baseline treatment and follow the comparative statics of statistical discrimination models. Recall that the only

[36] Because we cannot reject the hypothesis that the signal variances associated with the high and low mean groups are the same, we use the average of the estimated variances in Table 1 for the Baseline. In Appendix G 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).

[37] Even subject to a zero discrimination constraint, accuracy of assessments can be improved by putting some weight on the average mean, $\bar{\mu} = (\mu_h + \mu_l)/2$, as achieved with OptNoDiscrimination.

[38] Formally, the frontier is obtained by minimizing $\chi GD + (1-\chi)MSE$ over all linear inference strategies that have the following structure:

$$\hat{t} = \omega \left(\alpha(\frac{\mu_l + \mu_h}{2}) + (1-\alpha)\mu_g \right) + (1-\omega)s,$$

where $\omega \in [0,1]$ and $\alpha \in [0,1]$. 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.

difference between Baseline and NoGroup is that, in the former, subjects are told the group from which the type is drawn. Comparing Baseline vs. NoGroup in Figure 5, we see that subjects discriminate significantly more in Baseline by locating to the *right* of NoGroup in the graph ($p = 0.000$) and that MSE in Baseline drops relative to NoGroup ($p = 0.090$).[39]

However, Figure 5 also reveals that this discrimination is *inefficient*, both relative to the Bayesian benchmark (which is the right benchmark for our subjects whose incentive is to minimize MSE) and relative to the accuracy-discrimination frontier. The Baseline point is to the north of the Bayesian benchmark, meaning subjects (i) discriminate at similar levels to the Bayesian bechmark ($p = 0.164$), but (ii) make inefficiently inaccurate assessments ($p = 0.000$). The Baseline point is also significantly far from 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 5. *Subjects engage in statistical discrimination and follow its comparative statics, improving accuracy by discriminating on the basis of group. However, discrimination observed in the Baseline treatment is inefficient: by eliminating mistakes in inference it would be possible to reduce discrimination without reducing accuracy.*

This inefficiency is a product of the interaction of the two key errors identified in the previous section (contrast-driven evaluation and base-rate neglect) which both reduce accuracy, but have opposite effects on discrimination: while base-rate neglect tends to lower discrimination by causing agents to put too little weight on group information, contrast-driven evaluation pushes behavior in the opposite direction. Figure 5 illustrates the effect of each of these two errors by plotting the counterfactual NoBias benchmark, which represents the case where the weight remains as in the Baseline but contrast-driven evaluation is eliminated. In particular, while eliminating contrast-driven evaluation results in a small improvement to accuracy, it actually decreases discrimination by half. Also, the fact that the NoBias benchmark lies roughly halfway between the Bayesian and pBRN benchmarks, indicates a significant degree of base-rate neglect in the Baseline treatment. Correcting base-rate neglect (by changing weight ω to optimal ω^{Bay}) would result in achievement of the Bayesian benchmark, yielding an additional 25% increase in accuracy at the expense of a more than doubling of discrimination.[40]

[39] See Appendix E for details. As reported in Appendix H, the drop is slightly larger when we look at all rounds ($p = 0.015$).

[40] 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 as noted above the Baseline

Result 6. *Inefficiency in the Baseline treatment relative to the accuracy-discrimination frontier is a joint consequence of (i) base-rate neglect and (ii) contrast-driven evaluation. Eliminating contrast-driven evaluation decreases discrimination by 50% at no cost to accuracy. In addition, correcting for base-rate increases accuracy by about 25%, but doing so more than doubles the amount of discrimination.*

To whatever degree contrast-biased evaluation arises in real discrimination contexts, our treatments, NoGroup, SignalFirst and OneGroup (which allowed us to identify contrast-driven evaluation in the previous section), double as candidate *behavioral interventions* for improving outcomes. We therefore re-evaluate these treatments, focusing on how these treatments improve outcomes in terms of accuracy and discrimination given the inefficiencies we observe in the Baseline treatment.

Interpreted as a behavioral intervention, the NoGroup treatment mirrors "identity-blinded evaluation policies," an approach to discrimination that has been of great interest to applied economists in the last few decades.[41] In the field, these policies, among other things, remove scope for taste-based discrimination. Here (where scope for taste-based discrimination is already removed by design), we identify the impact such policies have purely on the quality of inference. Withholding information on group identity completely eliminates discrimination by (i) removing scope for statistical discrimination and (ii) removing contrast-driven evaluation. The negative implication is only a marginal increase in inaccuracy. Our results show that, due to the inferential mistakes measured in the Baseline treatment, the NoGroup treatment presents a more attractive policy option than classical models would suggest: at least in our data, it eliminates irrational discrimination and (because of errors in inference) sacrifices little in terms of accuracy in the process.

Result 7. *Withholding group identity fully eliminates discrimination with only a small cost to accuracy.*

The SignalFirst treatment presents a more subtle, and arguably more attractive intervention. Simply requiring subjects to observe the signal *before* they learn the group leads to a significant reduction (a significant leftward shift relative to Baseline, $p = 0.002$) in discrimination without any cost to accuracy.

The plot also shows why this intervention works. The SignalFirst dot is identical to the NoBias

variance was used to construct the pBRN benchmark.

[41] 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).

counterfactual in the plot, suggesting that the intervention works entirely by eliminating contrast-driven evaluation. Indeed, results from the estimates reported in the previous subsection show just this. By forcing subjects to see the signal before learning group information, SignalFirst eliminates the bias altogether. However, the intervention has no complementary effect on the severity of base-rate neglect, which is resistant to this treatment.

Result 8. *Forcing subjects to evaluate evidence prior to learning the group causes subjects to discriminate significantly less at no cost to accuracy.*

A third intervention – though one that may be more difficult to implement in practice – is to task evaluators to specialize in assessing members of only one group, removing scope for contrast effects. The OneGroup treatment implements such an intervention by assigning each subject to evaluate only one group (high/orange or low/green) throughout the experiment. Figure 5 reveals a resulting strong reduction in *both* inaccuracy and discrimination ($p = 0.001$ for MSE and $p = 0.000$ for GD) in OneGroup relative to Baseline. Indeed, the OneGroup point lies *below* the Bayesian accuracy-discrimination frontier that we calibrated using the Baseline signal variance.

Once again, estimates in Table 1 and Figure 3 shed light on how this dramatic improvement in inference is achieved. As with the SignalFirst treatment, OneGroup completely eliminates the bias generated by contrast-biased evaluation: removing scope for contrast effects causes subjects to evaluate the signal in an unbiased way. Also like SignalFirst, OneGroup has no corresponding effect on base-rate neglect: $\omega - \omega^{Bay}$ is similar. However, unlike SignalFirst, estimates for signal variance are substantially lower in OneGroup than in Baseline ($p = 0.016$ for low-mean group and $p = 0.051$ for high-mean group), indicating that subjects are able to extract more precise 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. The decrease in signal variance in OneGroup seems unlikely to be due to the removal of contrast effects (our initial motivation for implementing this treatment) since variance does not decrease in the NoGroup treatment, where subjects similarly make assessments without scope for contrasting groups. 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. The decrease in variance is broadly 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.

In summary, this kind of specialization-in-assessment is a particularly effective intervention in our setting: it eliminates contrast-biased evaluation while also improving the precision of signals. To the extent that these results carry over to field settings, such policies might lead to improvements in important applications (e.g, hiring decisions). The results suggest that (keeping all else constant about the informational environment) specialization can lead to more accurate assessments. But more research is needed to study how specialization can be implemented in the field. In our experiment, we are able to entirely remove contrast effects via specialization by making the subjects completely unaware of the existence of a contrasting group or its properties. This is ideal for testing the hypothesis of contrast-driven evaluation (as we do in the previous section), but it is not clear that this can be implemented in the most important discrimination contexts in the field. For instance, specializing evaluators of e.g. resumes to examine only members of one gender may not produce an effect like the one we measure here because evaluators are aware that another gender exists and may have pre-existing beliefs about how types across these groups contrast with one another, plausibly producing contrast-biased evaluation.[42] This type of intervention seems most likely to be effective in settings in which evaluators do not have strong pre-existing beliefs about how the type (i.e., the measure being evaluated) is differentially distributed across groups. On the other hand, the surprising boost in the precision of signal evaluation we observe in the OneGroup treatment may well apply more generally, meaning this intervention may nonetheless be valuable for improving statistical discrimination.

Result 9. *Tasking subjects to specialize in evaluating only one group causes subjects to discriminate significantly less and simultaneously significantly improves the accuracy of their assessments.*

A final candidate intervention is to explicitly use our model and suitable data to directly debias decision-makers afflicted by contrast-biased evaluation. In particular, if a policy maker has access to data on (i) true types and (ii) assessments, she can use estimates from a model like ours to directly improve outcomes.[43] In Online Appendix J, we describe an algorithm to do this and evaluate its effectiveness using our data. Forming a training sample using half of our data, we estimate (Δ, ω, ξ^2) and use them to adjust predictions in a testing sample consisting of the other half of

[42] Bohnet, Van Geen & Bazerman (2016) provides evidence against specialization, documenting separate evaluation of job candidates to induce greater reliance on group stereotypes relative to joint evaluation.

[43] This type of data might be available, for example, to firms that can combine data on initial assessment of workers at the hiring stage with long run productivity of workers, as revealed in internal reviews and career trajectories. However, in many applications, there can be challenges associated with constructing a data set of this kind for two reasons: (i) biases in initial assessments can have long lasting effects, making it difficult to identify true type; (ii) representation of different groups or types in the data set can be impacted by initial assessments creating selection problems. See Bohren, Hull & Imas (2022) for discussion on related issues.

our data. In the Appendix, we show that a policy maker targeting zero discrimination (i.e. the OptNoDiscrimination benchmark) can adjust assessments to reduce discrimination and inaccuracy. We also show that a policy maker who simply wants to maximize accuracy (i.e., targeting the Bayesian benchmark) can use this debiasing to significantly improve accuracy.

7 Discussion

We provide experimental evidence that contrast effects can severely distort the way people evaluate new evidence – an error that we call "contrast-biased evaluation." Subjects misperceive information on individual group members to be more representative of the group the individual belongs to (relative to an irrelevant reference group) than it really is. This error produces an irrational amplification in how discriminatory assessments are between individuals from different groups. Importantly, we show that this bias is highly resistant to learning, continuing to distort inference even after dozens of periods of experience and feedback. However, we also show that simple cognitively-driven interventions are highly effective at removing this bias, simultaneously reducing discrimination and improving the accuracy of individual judgements.

This error, driven as it is by contrast effects, is particularly relevant to settings of statistical discrimination, in which assessment of individual group members are necessarily formed in the shadow of some other reference group. As such our results may have relevance to perennial efforts to reduce discrimination in naturally occurring settings like employment, education, law and policing. In many such settings, decision makers must process complex information about individuals that is open to interpretation, producing scope for the sorts of evaluative distortions we document here. To the degree this bias indeed arises outside of the lab, our paper offers new tools for combating discrimination using relatively "light touch" policies that can effectively reduce discrimination without reducing the accuracy of judgements.[44]

When should we expect this bias to occur outside of the laboratory, amplifying discrimination in the field? Ultimately, this is an empirical question – and one well-worth investigating. However we

[44]Of course, the availability of the interventions used in our experiment will vary widely across field contexts. For instance, in many contexts it will not be possible to have decision makers assess evidence about individuals before knowing the group the individual belongs to. Likewise, decision makers will not always have information sufficient to use our model to debias decision making. Finally, it is not clear whether (or when) specializing decision makers to make assessments about members of only one group will replicate the treatment effect of our OneGroup treatment – it may be that even then, knowledge of the existence of another group will weaken or even eliminate the treatment effect we observe.

can provide some informed speculation. For this, it is important to emphasize two key features of our experiment: (i) it focuses on a specific perceptual task (estimation of the number of dots observed on a screen) and (ii) it involves two abstract groups (orange vs. green). The first feature creates scope for misperception while the second feature makes contrasts between prior distributions of the two groups clear and salient. As such, the magnitudes we measure for contrast-biased evaluation (and its severity relative to other biases) should hardly be expected to be universal, but rather should be expected to depend on how strongly the real world environment deviates from these characteristics. We conjecture that contrast-biased evaluation is likely to be more important in settings where (i) information on individuals is complex or highly subjective and therefore open to selective (mis-)interpretation and (ii) contrast between different groups are highly salient. To the degree that individual signals derive from hard evidence (where there is no scope for misinterpretation) or contrasts between groups are not as pronounced or salient, we would expect to see attenuation or even elimination of contrast-biased evaluation relative to what we report here.

Similarly, it is an empirical question how important contrast-biased evaluation is relative to alternative drivers of discrimination considered in the literature. Our experiment was designed to isolate contrast-biased evaluation and understand its nature, and we therefore purposely shut down scope for surely-important competing drivers of discrimination like incorrect beliefs about group differences, taste-based discrimination, or animus. Doing this allowed us to investigate the nature of this bias with a level of clarity that would be difficult or impossible using field evidence. But in order to understand the importance of this error relative to other drivers of discrimination, we must move outside of the lab, into the contexts where the full suite of determinants of discrimination are allowed to appear with their natural strengths. For this kind of investigation, our description of the psychological mechanism underlying the bias and our documentation of treatment effects may provide important diagnostic guidance on how to identify contrast-biased evaluation and its relative importance in driving discrimination in the field.

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.

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

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., Ditlmann, R., Gehrig, S. & Schweighofer-Kodritsch, S. (2022), 'Explicit and implicit belief-based gender discrimination: A hiring experiment'.

Bartoš, V., Bauer, M., Chytilová, J. & Matějka, F. (2016), 'Attention discrimination: Theory and field experiments with monitoring information acquisition', *American Economic Review* **106**(6), 1437–75.

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.

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

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

Bhargava, S. & Fisman, R. (2014), 'Contrast effects in sequential decisions: Evidence from speed dating', *Review of Economics and Statistics* **96**(3), 444–457.

Bohnet, I., Van Geen, A. & Bazerman, M. (2016), 'When performance trumps gender bias: Joint vs. separate evaluation', *Management Science* **62**(5), 1225–1234.

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

Bohren, J. A., Hull, P. & Imas, A. (2022), Systemic discrimination: Theory and measurement, 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.

Bordalo, P., Gennaioli, N. & Shleifer, A. (2018), 'Diagnostic expectations and credit cycles', *The Journal of Finance* **73**(1), 199–227.

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.

Coffman, K. B., Exley, C. L. & Niederle, M. (2021), 'The role of beliefs in driving gender discrimination', *Management Science* **67**(6), 3551–3569.

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. & Yuksel, S. (2022), Mental models and learning: The case of base rate neglect, Technical report.

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* .

Gennaioli, N. & Shleifer, A. (2010), 'What comes to mind', *The Quarterly Journal of Economics* **125**(4), 1399–1433.

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.

Gneezy, U., Saccardo, S., Serra-Garcia, M. & van Veldhuizen, R. (2020), 'Bribing the self', *Games and Economic Behavior* **120**, 311–324.

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

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.

Hartzmark, S. M. & Shue, K. (2018), 'A tough act to follow: Contrast effects in financial markets', *The Journal of Finance* **73**(4), 1567–1613.

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*a*), 'On prediction and judgement', *ORI Research monograph* **1**(4).

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

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

Kessler, J. B., Low, C. & Shan, X. (2022), 'Lowering the playing field: Discrimination through sequential spillover effects'.

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.

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.

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

Mobius, M. M., Niederle, M., Niehaus, P. & Rosenblat, T. S. (2022), 'Managing self-confidence: Theory and experimental evidence', *Management Science* .

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.

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

Oprea, R. & Yuksel, S. (2022), 'Social exchange of motivated beliefs', *Journal of the European Economic Association* **20**(2), 667–699.

Radbruch, J. & Schiprowski, A. (2021), 'Interview sequences and the formation of subjective assessments'.

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.

Saccardo, S. & Serra-Garcia, M. (2022), Cognitive flexibility or moral commitment? evidence of anticipated belief distortion,, Technical report, working paper.

Sarsons, H. (2017), 'Interpreting signals in the labor market: evidence from medical referrals'.

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.

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

Online Appendix for

Contrast-Biased Evaluation

Ignacio Esponda Ryan Oprea Sevgi Yuksel

CONTENTS:

A Contrast-Driven Evaluation

Assume $f(t \mid g)$ is normally distributed with mean μ_g and variance σ^2. The normal prior implies the following odds ratio between any two values t_1 and t_2:

$$r(t_1, t_2) = \frac{f(t_1)}{f(t_2)} = e^{-\frac{1}{2\sigma^2}\left((t_1-\mu)^2-(t_1-\mu)^2\right)} = e^{-\frac{1}{2\sigma^2}\left((t_1^2-t_2^2-2\mu(t_1-t_2)\right)} = e^{-\frac{t_1-t_2}{2\sigma^2}((t_1+t_2)-2\mu)} \tag{8}$$

This implies that if for any (t_1, t_2) the odds ratio can be written as above, then the distribution of t is normally distributed with mean μ and variance σ^2.

Lemma 1. *For any γ, $\tilde{f}(t \mid g) := \kappa f(t \mid g) \left(\frac{f(t \mid g)}{f(t \mid -g)}\right)^\gamma$, is normally distributed with distorted mean $\tilde{\mu}_g = \mu_g + \gamma(\mu_g - \mu_{-g})$ and variance σ^2.*

Proof. Without loss, focus on the likelihood ratios:

$$\begin{aligned} \frac{\tilde{f}(t_1 \mid g)}{\tilde{f}(t_2 \mid g)} &= \left(\frac{f(v_1 \mid g)}{f(v_2 \mid g)}\right)\left(\frac{f(v_1 \mid g)}{f(v_1 \mid -g)}\right)^\gamma \left(\frac{f(v_2 \mid -g)}{f(v_2 \mid g)}\right)^\gamma \\ &= e^{\frac{1}{2\sigma^2}\left(-(t_1-\mu_g)^2+(t_2-\mu_g)^2-\gamma(t_1-\mu_g)^2+\gamma(t_2-\mu_g)^2-\gamma(t_2-\mu_{-g})^2+\gamma(t_1-\mu_{-g})^2\right)} \\ &= e^{\frac{1}{2\sigma^2}\left(t_2^2-t_1^2+2(t_1-t_2)(\mu_g+\gamma(\mu_g-\mu_{-g})\right)} \end{aligned} \tag{9}$$

Note that when $\gamma = 0$, we get the correct likelihood ratio corresponding to $\frac{f(t_1 \mid g)}{f(t_2 \mid g)}$ (as in Equation 8). For $\gamma > 0$, we get the likelihood ratio associated with a normal distribution with the same standard deviation but a distorted mean where $\tilde{\mu}_g = \mu_g + \gamma(\mu_g - \mu_{-g})$. □

Lemma 2. *Assume $h(s \mid t)$ and $f(t \mid g)$ are normally distributed with mean t and variance ξ^2 in the first case and μ_g and σ^2 in the second case. For any γ, $\tilde{h}(s \mid t, g) := \kappa h(s \mid t) \left(\frac{y(s \mid g)}{y(s \mid -g)}\right)^\gamma$ with $y(s \mid g) := \int h(s \mid t) f(t \mid g) dt$ is normally distributed with distorted mean $t + \frac{\gamma\xi^2}{\delta^2}(\mu_g - \mu_{-g}))$ where $\delta^2 = \sigma^2 + \xi^2$ and variance ξ^2.*

Proof. Note that $s = \mu_g + \epsilon_1 + \epsilon_2$ where $\epsilon_1 \sim \mathcal{N}(0, \sigma^2)$ and $\epsilon_2 \sim \mathcal{N}(0, \xi^2)$. So $y(s \mid g)$ is normally distributed with mean μ_g and variance $\delta^2 := \sigma^2 + \xi^2$. Again, focus on the likelihood ratios:

$$\begin{aligned} \frac{\tilde{h}(s_1 \mid t,g)}{\tilde{h}(s_2 \mid t,g)} &= \left(\frac{h(s_1 \mid t)}{h(s_2 \mid t)}\right)\left(\frac{y(s_1 \mid g)}{y(s_1 \mid -g)}\right)^\gamma \left(\frac{y(s_2 \mid -g)}{y(s_2 \mid g)}\right)^\gamma \\ &= e^{\frac{1}{2\xi^2}\left(-(s_1-t)^2+(s_2-t)^2+\frac{\gamma\xi^2}{\delta^2}\left(-(s_1-\mu_g)^2+(s_2-\mu_g)^2-(s_2-\mu_{-g})^2+(s_1-\mu_{-g})^2\right)\right)} \\ &= e^{\frac{1}{2\xi^2}\left(s_2^2-s_1^2+2(s_1-s_2)(t+\frac{\gamma\xi^2}{\delta^2}(\mu_g-\mu_{-g})\right)} \end{aligned} \tag{10}$$

Note that when $\gamma = 0$, we get the correct likelihood ratio corresponding to $\frac{h(s_1 \mid t)}{h(s_2 \mid t)}$, for $\gamma > 0$, we get the likelihood ratio associated with a normal distribution with the same standard deviation but a distorted mean where the mean is not t (as it should be), but $t + \frac{\gamma\xi^2}{\delta^2}(\mu_g - \mu_{-g}))$. □

In general, consider three distinct ways representativeness can distort beliefs in the updating process.

Distortion 1 (Distortion of the prior): As in Bordalo et al. 2016, representativeness may distort the prior (i.e., the memory of or characterization of each group's distribution). Let $\gamma^1 \geq 0$ be a measure of this distortion as described above.

Distortion 2 (Recollective Signal Distortion): If agents learn group identify after evaluating the signal, the recollection of the signal (at the updating stage) could be impacted by group identity. The idea is that agents might suffer from a form of "associative memory" where they remember the image in a way that is more representative of that group. Let $\gamma^2 \geq 0$ be a measure of this distortion as described above.

Distortion 3 (Evaluative Signal Distortion, referred to as *Contrast-biased evaluation* in the paper): If agents know group identify before evaluating the signal, the perception of the signal might be impacted by that knowledge. The idea is that the agent "looks for" evidence that is more representative of the group (high signals for high-mean group, low signals for low-mean group). Let $\gamma^3 \geq 0$ is a measure of this distortion as described above.

Assume $h(s \mid t)$ and $f(t \mid g)$ are normally distributed with mean t and variance ξ^2 in the first case and μ_g and σ^2 in the second case. By Lemmas 1 and 2, the agent will update under distorted prior $\tilde{f}(t \mid g)$ which is normal with mean $\tilde{\mu}_g = \mu_g + \gamma^1(\mu_g - \mu_{-g})$ and variance σ^2 and distorted subjective signal $\tilde{h}(s \mid t)$ which is normal with mean $\tilde{t} = t + (\gamma^1 + \gamma^2)\frac{\xi^2}{\sigma^2+\xi^2}(\mu_g - \mu_{-g})$ and variance ξ^2.[45]

[45]This computation assumes that the signal distortions (2 and 3) are unaffected by the prior distortion (1). It is possible that the distorted prior also influences the signal distortion. This could be modelled as prior distortion taking place first and then feeding into the how representativeness of different signals are computed (formally, f would be replaced by $\tilde{f}$ in Lemma 2).

B Under or Over Inference

The aim of this section to formally show how classical deviations from Bayesian updating—under inference (conservatism) or over inference (base-rate neglect)—imply deviations in terms of weight on prior ω, but do not result in biased inference, as captured by B (or Δ).

The standard approach (since Grether (1980), recently reviewed in Benjamin (2019)) to studying deviations from Bayesian updating uses the following framework which characterizes distortions in the posterior likelihood ratio:

$$\frac{p(t=t_1 \mid s)}{p(t=t_2 \mid s)} = \left(\frac{p(t_1)}{p(t_2)}\right)^{\alpha} \left(\frac{p(s \mid t=t_1)}{p(s \mid t=t_2)}\right)^{\beta}, \tag{11}$$

where optimal behavior requires $\alpha = \beta = 1$.

Lemma 3. *If $s \sim \mathcal{N}(t, \xi^2)$ and $t \sim \mathcal{N}(\mu, \sigma^2)$, it can be shown that for any value of (α, β), assessments consistent with Equation 11 are given by $\hat{t} = \omega\mu + (1-\omega)s$, where $\omega = \frac{\alpha\xi^2}{\beta\sigma^2+\alpha\xi^2}$. That is, ω increases with α and decreases with β, but cannot account for a bias term B.*

Proof. Consider a behavioral agent who, conditional on signal s (with g representing the distribution of s), forms a posteriors odds ratios as below for any (t_1, t_2).

$$\begin{aligned}
\hat{r}(t_1,t_2) &= \left(\frac{f(t_1)}{f(t_2)}\right)^{\alpha} \left(\frac{g(s \mid t=t_1)}{g(s \mid t=t_2)}\right)^{\beta} \\
&= \left(e^{-\frac{t_1-t_2}{2\sigma^2}((t_1+t_2)-2\mu)}\right)^{\alpha} \left(e^{-\frac{t_1-t_2}{2\xi^2}((t_1+t_2)-2s)}\right)^{\beta} \\
&= e^{-\frac{\alpha\xi^2(t_1-t_2)}{2\sigma^2\xi^2}((t_1+t_2)-2\mu)} e^{-\frac{\beta\sigma^2(t_1-t_2)}{2\sigma^2\xi^2}((t_1+t_2)-2s)} \\
&= e^{-\frac{(\alpha\xi^2+\beta\sigma^2)(t_1-t_2)}{2\sigma^2\xi^2}\left((t_1+t_2)-2\left(\frac{\alpha\xi^2\mu+\beta\sigma^2 s}{\alpha\xi^2+\beta\sigma^2}\right)\right)}
\end{aligned} \tag{12}$$

This is equivalent to a normal distribution with mean $\frac{\alpha\xi^2\mu+\beta\sigma^2 s}{\alpha\xi^2+\beta\sigma^2}$ and variance $\frac{\sigma^2\xi^2}{\alpha\xi^2+\beta\sigma^2}$. Note that the optimal assessment is the mean of this distribution. □

C Measures of Discrimination

We refer the reader to Barocas, Hardt & Narayanan (2019), Narayanan (2018) and Hutchinson & Mitchell (2019) for reviews of this literature. We focus on a criteria of non-discrimination that is most relevant in our setting: *Separation*.[46] In our inference task, this criteria refers to the statistical properties of the joint distribution of $\tilde{t}$ (assessment), t (true type) and g (group identity).

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

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 types between the groups. Namely, the criterion requires people from different groups with the same underlying true type to be treated the same. This is a 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.[47]

[46] 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*: $\tilde{t} \perp g$. 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 types *do* differ by group (as in our setting).
>
> *Sufficiency*: $t \perp g \,|\, \tilde{t}$. This criterion requires types to be independent of the group identity *conditional* on assessments. The criterion requires the distribution of true types to be the same for different groups when we condition on a specific assessment. The Bayesian Benchmark to our inference task satisfies this criterion.

[47] 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

D Bayesian Benchmark in NoGroup Treatment

While the optimal inference about type t is linear in signal s—under the assumption that $s \sim \mathcal{N}\left(t, \xi^2\right)$—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:

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

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

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

Figure 6 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.

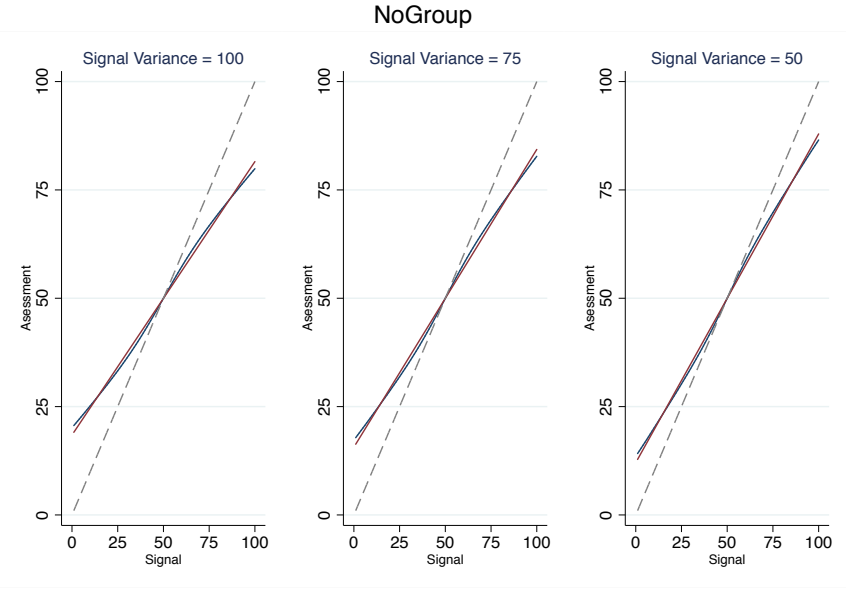


Figure 6: Bayesian Benchmark in NoGroup Treatment

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.

E Statistical Tests and Further Analysis

Table 2: OLS estimation (Dependent Variable: Squared Error in Assessments), Rounds 38-75

	(1)	(2)	(3)	(4)
NoGroup	12.02*	8.667	6.929	6.446
	(7.069)	(6.979)	(19.06)	(19.92)
OneGroup	-17.29***	-16.12***	-37.20**	-37.89**
	(5.107)	(5.141)	(14.76)	(15.36)
SignalFirst	-3.611	-4.335	-9.337	-9.541
	(6.301)	(6.364)	(18.09)	(18.52)
Risk measure		-0.217***		0.0804
		(0.0821)		(0.179)
Constant	59.16***	70.31***	82.34***	78.86***
	(4.308)	(6.446)	(14.15)	(15.75)
Observations	8778	8588	9158	8968

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.

In Tables 3 to 5, given estimates for (ω, Δ, ξ^2) we compute the likelihood with which agent i receives a higher assessment than agent j for different values of $t_i - t_j$. We compare cases in which (i) h vs. l: i is from the high-mean group and j is from low-mean group; (ii) *same*: i and j are from the same group; (i) l vs. h: i is from the low-mean group and j is from high-mean group.

To simplify the analysis, given that differences in ω and ξ^2 are limited between the groups (see Table 1), we set $\omega_h = \omega_l$ and $\xi_h^2 = \xi_l^2$ and use average estimated value of the two groups in our computations. By Equation 3, for agent i to receive a higher assessment than agent j, the following must hold:

$$(1-\omega)\Delta_{g_i} + \omega\mu_{g_i} + (1-\omega)t_i + (1-\omega)\varepsilon_i > (1-\omega)\Delta_{g_j} + \omega\mu_{g_j} + (1-\omega)t_j + (1-\omega)\varepsilon_j.$$

The likelihood of this happening is equal to $\Phi\left(\frac{(\Delta_{g_i}-\Delta_{g_j})+\frac{\omega}{1-\omega}(\mu_{g_i}-\mu_{g_j})+(t_i-t_j))}{\sqrt{2}\xi}\right)$. Note that this is increasing in $\Delta_{g_i} - \Delta_{g_j}$.

Table 3: Likelihood of Higher Assessment when $t_i - t_j = 0$.

	h vs. l	*same*	l vs. h
Baseline	.76	.50	.24
NoGroup	.52	.50	.48
SignalFirst	.65	.50	.35
OneGroup	.59	.50	.41

Table 4: Likelihood of Higher Assessment when $t_i - t_j = 5$.

	h vs. l	*same*	l vs. h
Baseline	.86	.66	.38
NoGroup	.67	.66	.64
SignalFirst	.78	.66	.51
OneGroup	.76	.69	.60

Table 5: Likelihood of Higher Assessment when $t_j - t_j = 10$.

	h vs. l	*same*	l vs. h
Baseline	.93	.79	.54
NoGroup	.81	.79	.78
SignalFirst	.88	.83	.66
OneGroup	.88	.79	.77

F Testing linearity

In Figure 7, 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 8 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 ($p = 0.652$ and $p = 0.118$, respectively).

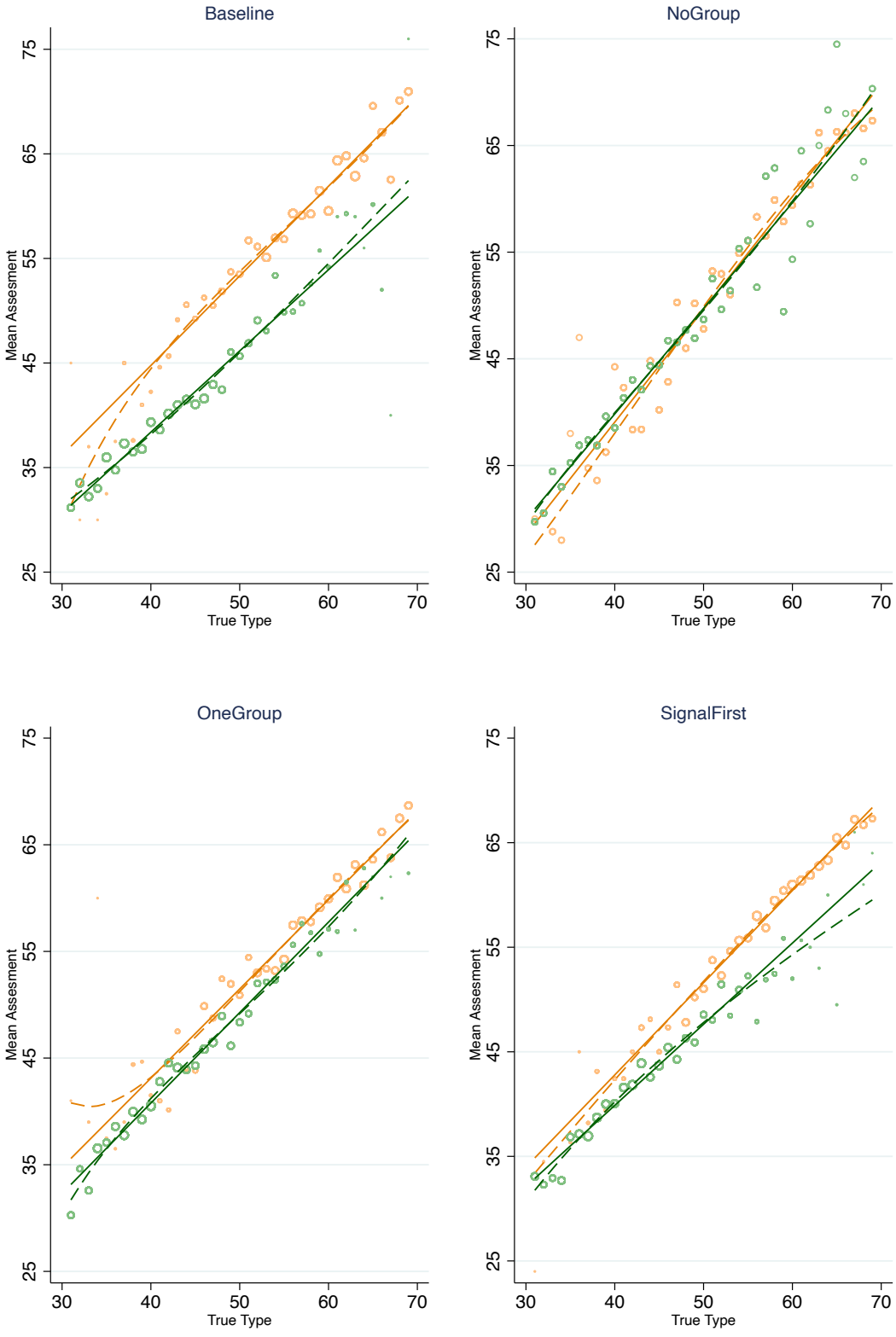


Figure 7: Best Linear Fit vs. Best Fractional Polynomial Fit *Notes: Size of the dots represent relative frequency (by group and treatment) of observing each true type. Solid lines depict best linear fit by group and treatment. Dashed lines depict best fractional polynomial fit.*

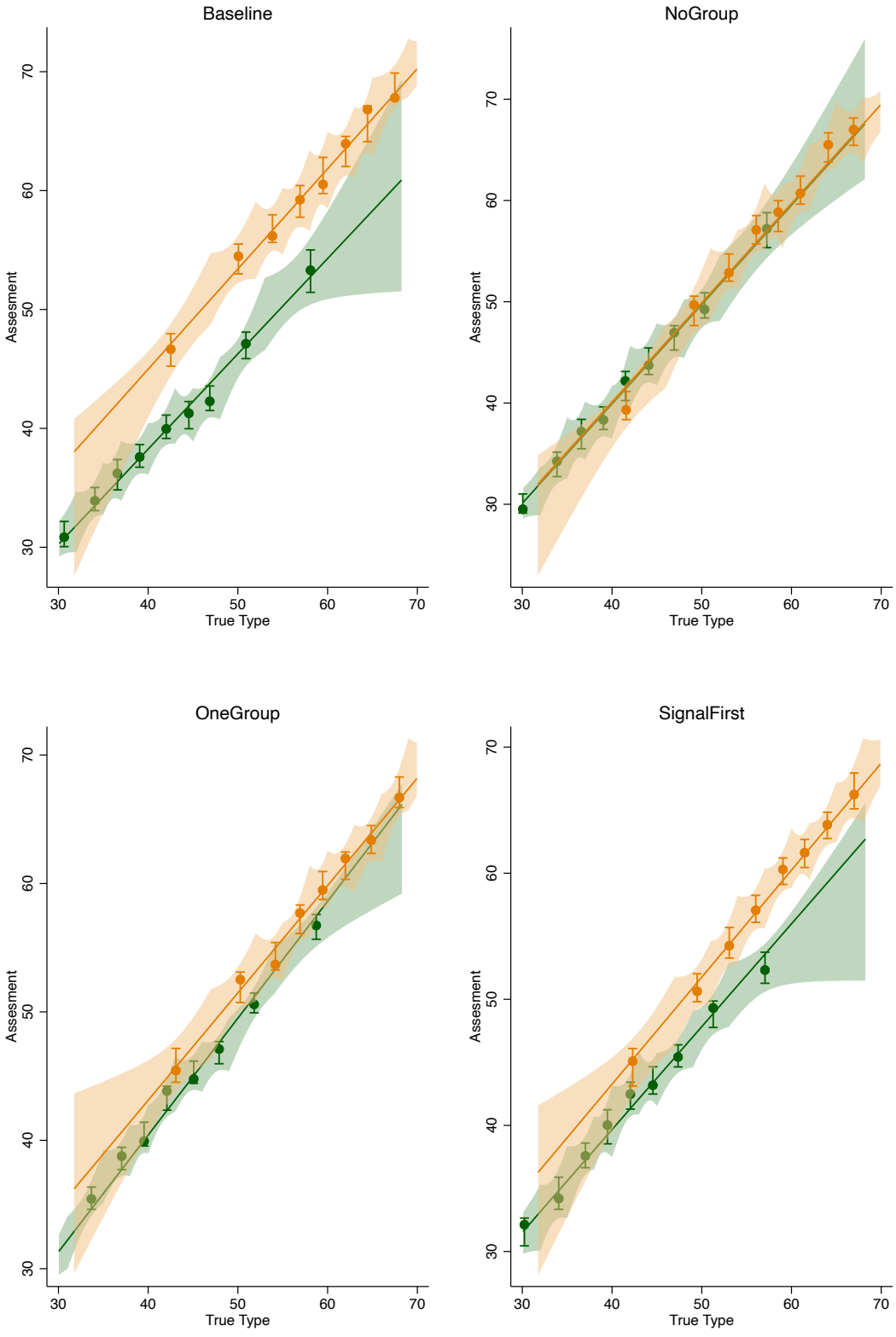


Figure 8: Best Linear Fit and Binned Scatter Plots *Notes: Dots show binned scatter plots depicting nonparamatric estimates of guess given true type in each bin. Green and Orange lines depict best linear fit by group and treatment. Black lines depict 95% confidence intervals. The shaded area depicts 95% confidence bands. All are computed with Stata package using binscatter methods developed in Cattaneo, Crump, Farrell & Feng (2019).*

G Individual-Level Analysis

Table 6: Model Estimates (Median Values)

	Baseline	NoGroup	SignalFirst	OneGroup
Regression estimates:				
ω_l	0.209	-0.009	0.150	0.095
ω_h	0.204	0.057	0.146	0.172
B_l	-2.265	0.266	-0.122	0.243
B_h	2.191	0.265	1.084	-0.176
Estimates derived from ω and B:				
Δ_l	-2.47	0.27	-0.66	0.30
Δ_h	2.53	0.50	1.04	-0.18
ξ_l^2	46	37	38	34
ξ_h^2	48	46	46	42
ω_l^{Bay}	0.31	0.16	0.27	0.26
ω_h^{Bay}	0.32	0.19	0.31	0.29
Tests:				
H_0: $\omega_l = \omega_h$	0.628	0.385	0.824	0.232
H_0: $B_l = B_h$	0.000	0.929	0.053	0.868
H_0: $\Delta_l = \Delta_h$	0.000	0.929	0.019	0.868
H_0: $\omega_l = \omega_l^{Bay}$	0.000	0.000	0.000	0.000
H_0: $\omega_h = \omega_h^{Bay}$	0.000	0.000	0.000	0.006

Rows on tests report p-value associated with test of each hypothesis.

Test are on distributions of individual values (Kolmogorov-Smirnov).

Notes: Empirical strategy is described in Table 1 (but here implemented on the individual level). See Sections 2 and 3 for further discussion.

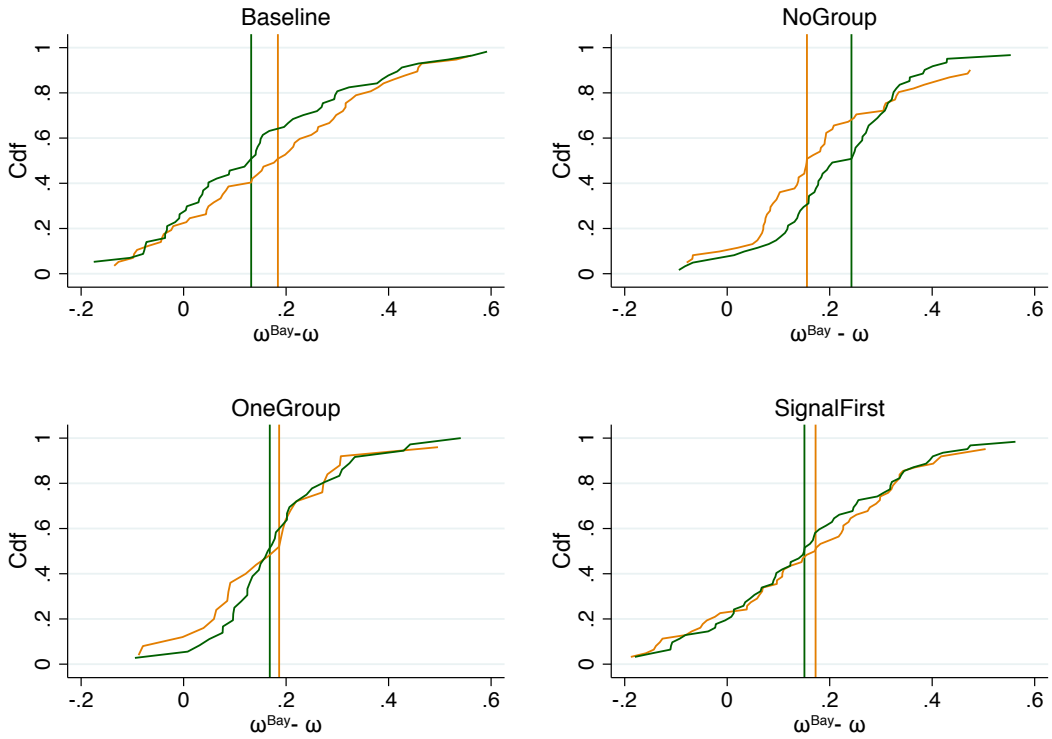


Figure 9: Estimates of Bayesian Weight vs. Weight on Prior by Group and Treatment *Notes: Green (Orange) solid line represents difference between Bayesian weight on prior ω^{Bay} and estimated weight on prior ω for low-mean (high-mean) group. Vertical lines denote median values. Empirical strategy is described in Table 1. See Sections 2 and 3 for further discussion.*

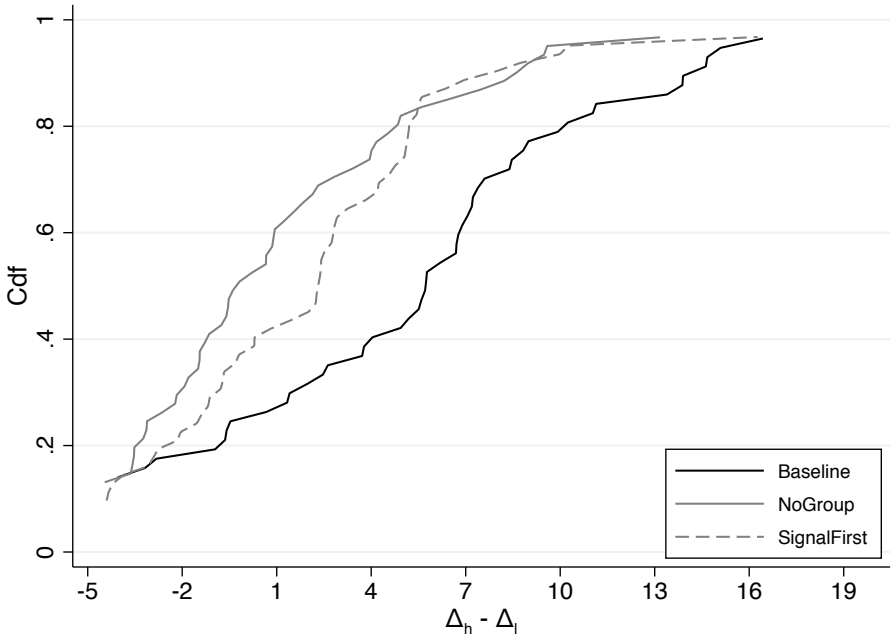


Figure 10: Distribution of Individual Estimates of $\Delta_h - \Delta_l$ by Treatment *Notes: OneGroup treatment is not included because $\Delta_h - \Delta_l$ cannot be estimated on the individual level.*

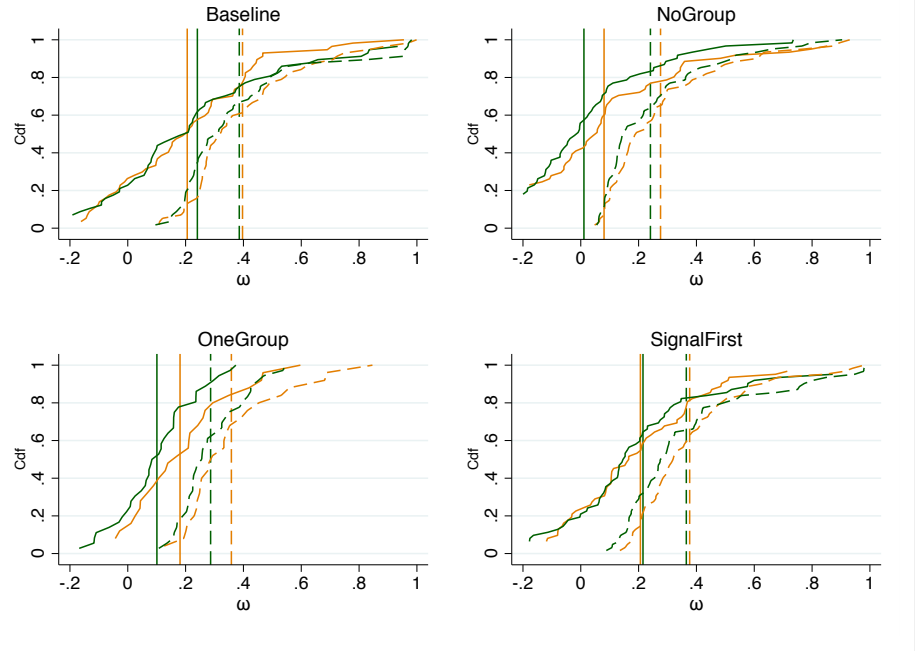


Figure 11: Distribution of Individual Estimates of Weight on Prior by Group and Treatment *Notes: Green (Orange) solid line represents estimated weight on prior for low-mean (high-mean) group. Green (Orange) dashed line represents Bayesian weight on prior for low-mean (high-mean) group. Vertical lines denote median individual values.*

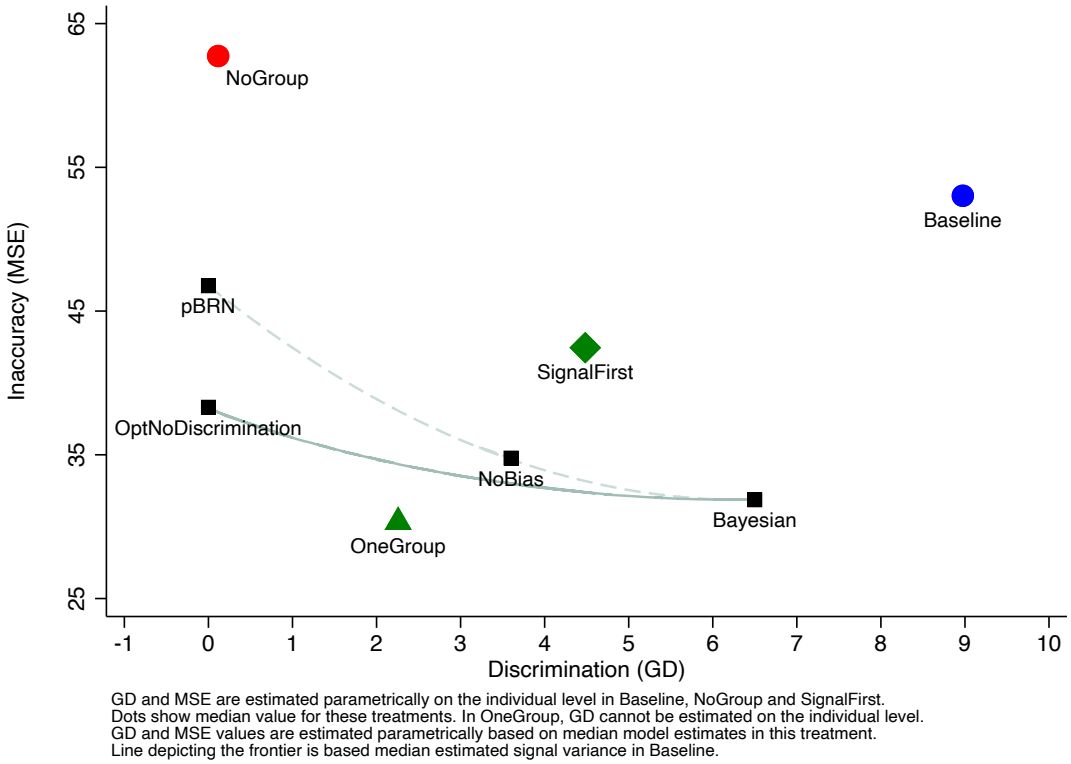


Figure 12: Accuracy-Discrimination Tradeoff (Using Median Values)

H Learning

Table 7: OLS estimation (Dependent Variable: Squared Error in Assessments), All Rounds

	(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 8: Model Estimates (All Rounds)

	Baseline	NoGroup	SignalFirst	OneGroup
Regression estimates:				
ω_l	0.184***	-0.00299	0.188***	0.0824***
	(0.0308)	(0.0298)	(0.0271)	(0.0217)
ω_h	0.175***	-0.0189	0.143***	0.115***
	(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)
Estimates derived from ω and B:				
Δ_l	-2.59***	-0.49	-1.08***	0.29
Δ_h	1.84***	-0.21	0.59	- 0.47
ξ_l^2	79***	73***	85***	53***
ξ_h^2	89***	79***	84***	60***
ω_l^{Bay}	0.44***	0.27***	0.46***	0.35***
ω_h^{Bay}	0.47***	0.29***	0.46***	0.38***
Tests:				
H_0: $\omega_l = \omega_h$	0.758	0.639	0.089	0.391
H_0: $B_l = B_h$	0.000	0.406	0.011	0.111
H_0: $\Delta_l = \Delta_h$	0.000	0.198	0.006	0.042
H_0: $\omega_l = \omega_l^{Bay}$	0.000	0.000	0.000	0.000
H_0: $\omega_h = \omega_h^{Bay}$	0.000	0.000	0.000	0.000

Notes: For each treatment and group $g \in \{l, h\}$, we estimate B_g and ω_g using OLS on the following specification: $y_{g,i} = B_g + \omega_g x_{g,i} + \varepsilon_{g,i}$, where i denotes each distinct observation, $y_{g,i} \equiv \hat{t}_{g,i} - t_{g,i}$, and $x_{g,i} \equiv \mu_g - t_{g,i}$. This specification is derived by subtracting $t_{g,i}$ from both sides of equation (4). Given estimates for B_g and ω_g, we back out Δ_g using equation (5) and estimate ξ_g^2 by identifying the error associated with the signal using $\varepsilon_{i,g} = (1-\omega_g)\epsilon_{i,g}$ and then taking the sample average of $\epsilon_{i,g}^2$. Given ξ_g^2, ω_g^{Bay} is derived from equation (2). See Sections 2 and 3 for further discussion. Standard errors (clustered at the subject level) are reported in parentheses. ***1%, **5%, *10% significance. Rows on tests report p-value associated with test of each hypothesis. Statistical assessments on estimates derived from ω and B use bootstrapping.

Table 9 reports changes in (ω_g, B_g) from early (1-37) to late (38-75) rounds. The table reveals, for example, ω_h in the Baseline was 0.192 in the early rounds and decreased by 0.0346 in the late rounds.

Table 9: Change in Model Estimates by Treatment

	Baseline	NoGroup	SignalFirst	OneGroup
ω_h	0.192***	-0.0580	0.135***	0.0655**
	(0.0392)	(0.0386)	(0.0353)	(0.0305)
ω_l	0.168***	-0.0267	0.188***	0.0746***
	(0.0326)	(0.0366)	(0.0397)	(0.0288)
B_h	1.222***	-0.363	0.750*	-0.632
	(0.332)	(0.380)	(0.445)	(0.532)
B_l	-2.529***	-0.702	-1.462***	0.120
	(0.436)	(0.563)	(0.369)	(0.336)
Change in ω_h in Rounds >37	-0.0346	0.0776	0.0165	0.0981***
	(0.0434)	(0.0513)	(0.0428)	(0.0372)
Change in ω_l in Rounds >37	0.0310	0.0460	-0.00414	0.0152
	(0.0418)	(0.0404)	(0.0445)	(0.0252)
Change in B_h in Rounds >37	0.613*	0.264	-0.504	0.476
	(0.348)	(0.605)	(0.380)	(0.705)
Change in B_l in Rounds >37	0.787**	0.413	1.134***	0.288
	(0.321)	(0.748)	(0.379)	(0.311)
Observations	4050	4350	4425	4500

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

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

Subjects with MSE less than or equal to 200.

I Additional Figures

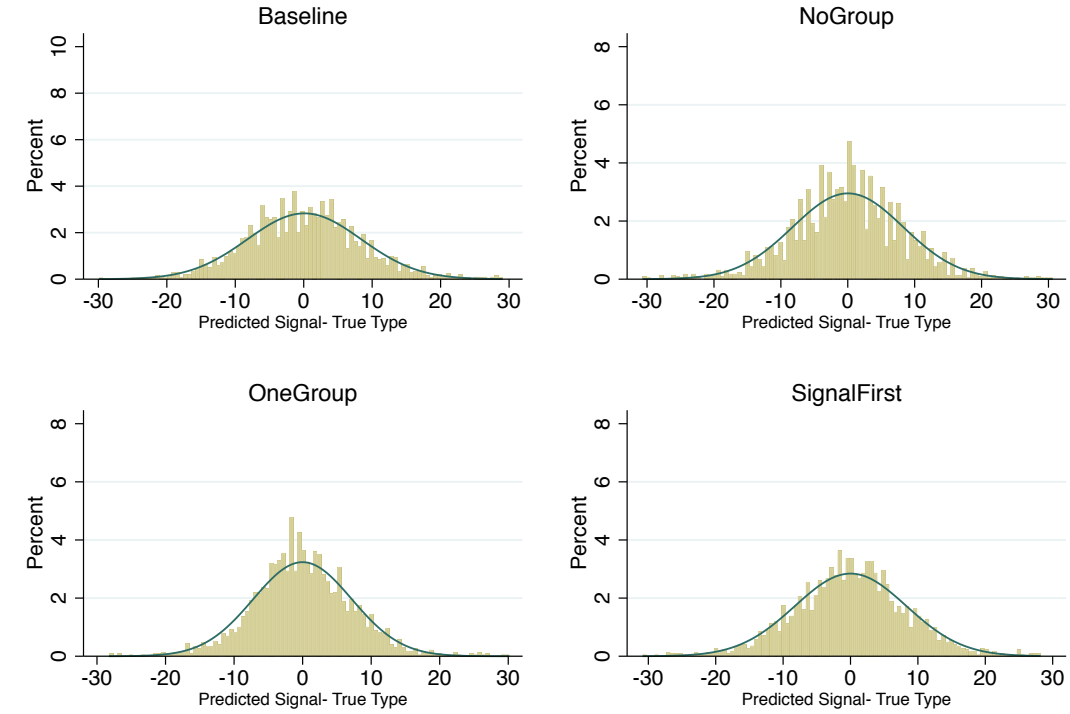
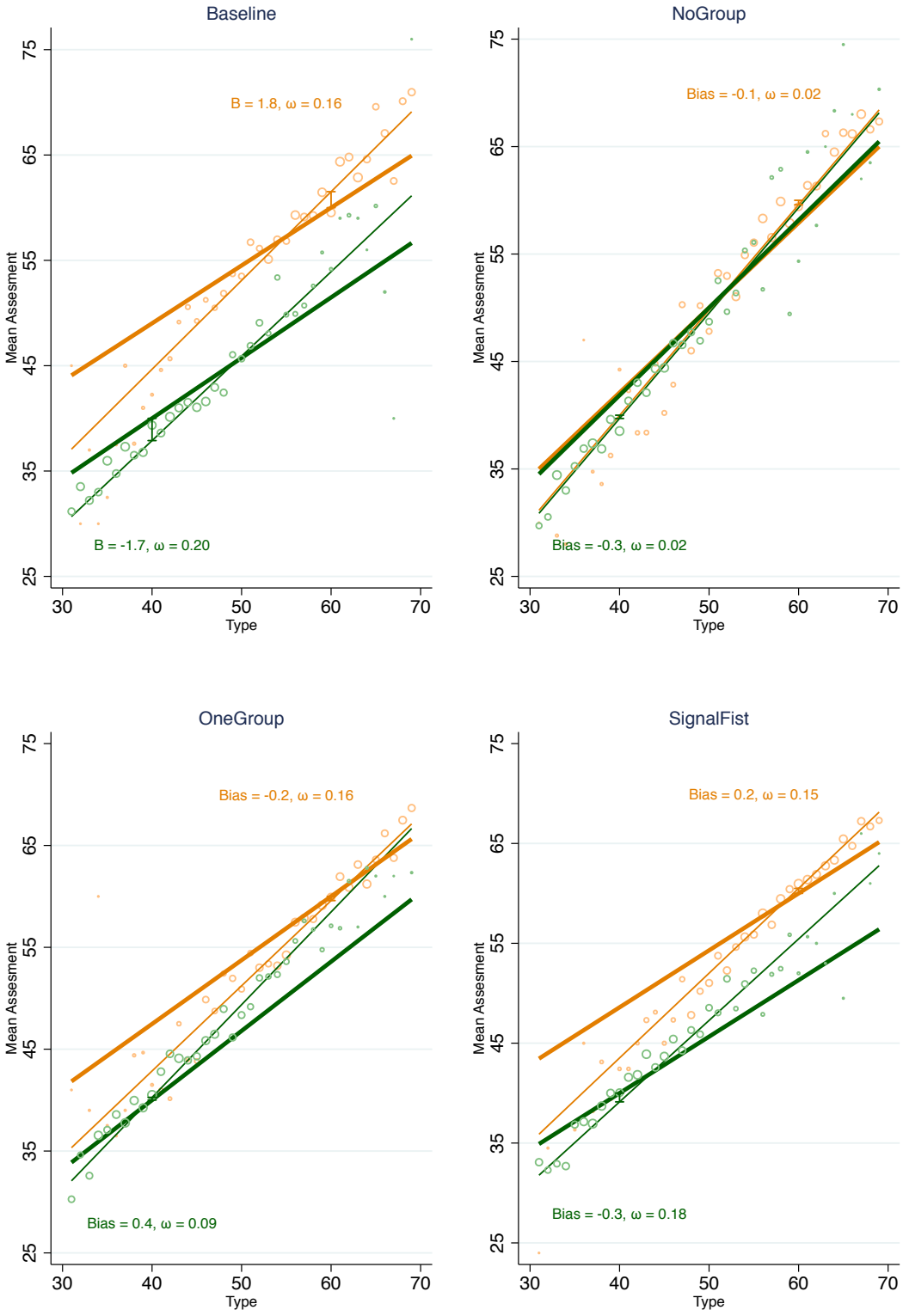


Figure 13: Distribution of Estimated Signal Error by Treatment



Green (Orange) dots are for low (high) mean group. Thinner Green and Orange lines depict best linear fit by group and treatment.
Thicker Green and Orange lines depict outcome of Bayesian inference strategy by group and treatment; gray line depicts 45 degree line. Subjects with MSE > 200 excluded.

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

J Using Model Estimates to Improve Outcomes

Our model suggests that an outside observer can construct a simple algorithm to correct for the errors we find: contrast-driven evaluation and base-rate neglect. 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 key parameters of the model $(\Delta_g, \omega_g, \xi_g^2)$. 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 $(\Delta_g, \omega_g, \xi_g^2)$, for each observation in the testing set, estimate a (debiased) subjective signal.
   - Each assessment $\hat{t}$ can be represented as follows: $\hat{t} = (1-\omega_g)\Delta_g + \omega_g\mu + (1-\omega_g)s_u$, where s_u is an unbiased signal.
2. The signal variance estimate ξ_g^2 implies an optimal weight ω_g for each of the two benchmarks.
   - For the OptNoDiscrimination benchmark, $\omega^{Ond} = \frac{\xi_g^2}{\xi_g^2+2\sigma^2}$[48]
   - For the Bayesian benchmark (as described in Section 2), $\omega^{Bay} = \frac{\xi_g^2}{\xi_g^2+\sigma^2}$.
3. Compute adjusted prediction using estimates for signal s_u and weights ω_g^{Ond} or ω_g^{Bay}.
   - Adjusted Bayesian prediction $\hat{t}^{Bay} = \omega_g^{Bay}\mu_g + (1-\omega_g^{Bay})s_u$.
   - Adjusted OptNoDiscrimination prediction $\hat{t}^{Ond} = \omega_g^{Ond}\left(\frac{\mu_l+\mu_h}{2}\right) + (1-\omega_g^{Ond})s_u$.

Table 10 reports results from 500 random repetitions of the procedure described above. We make two observations. First, adjusted predictions in the OptNoDiscrimination column are welfare increasing, generating lower inaccuracy and discrimination relative to the data. Second, adjusted

[48] Consistent with our linearity restriction in Section 3, ω_g is chosen to minimize expected squared error of predictions $\hat{t} = (1-\omega_g)s + \omega_g\left(\frac{\mu_l+\mu_h}{2}\right)$.

Table 10: Actual vs. Adjusted Predictions

	Data		OptNoDiscrimination		Bayesian
GD	7.98	$> (p = 0.000)$	0.81	$< (p = 0.002)$	9.40
MSE	59	$> (p = 0.088)$	56	$> (p = 0.000)$	44

Values represent mean estimates from 500 repetitions of the procedure.

Inequalities compare counterfactuals (OptNoDiscrimination and Bayesian) to data.

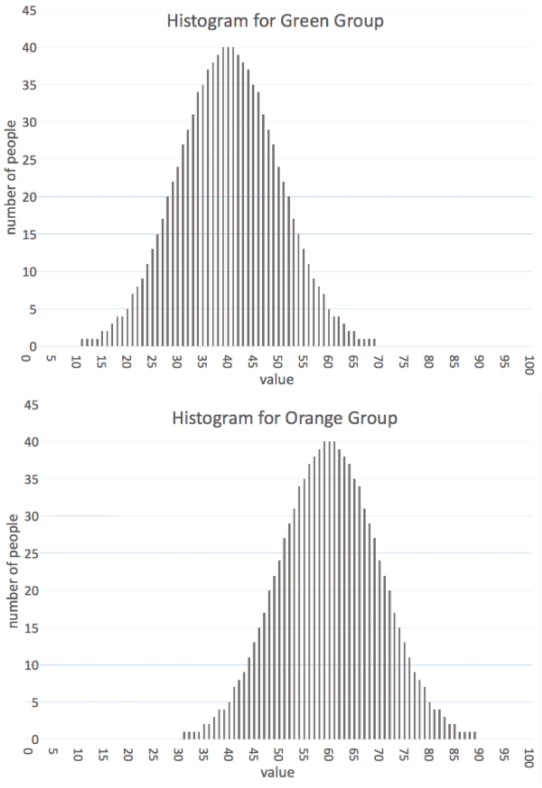
Reported p-values represent frequency of repetitions in which the inequality was violated.

predictions in the Bayesian column display higher discrimination, but much lower inaccuracy as expected from our earlier analysis. These results demonstrate that our model might be useful for correcting for biases like contrast-driven evaluation and base-rate neglect in applications in which the relevant data is available.

K Instructions for Baseline Treatment

The Data

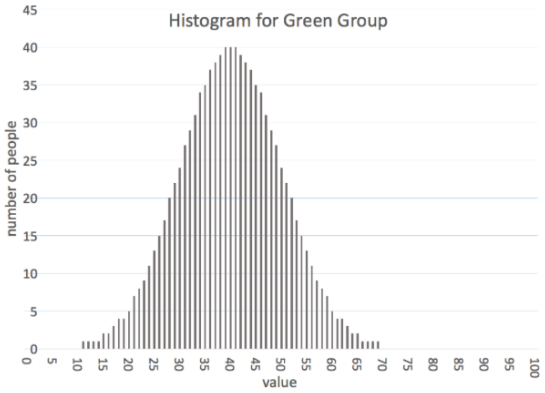
- We have **TWO GROUPS** of people. Each group consists of 1,000 people.
  - We'll refer to the first group as the Green group, and the second group as the Orange group.
- Each person is assigned a **NUMERICAL VALUE** from 0 to 100.
- The following figures, which are called histograms, depict the distribution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

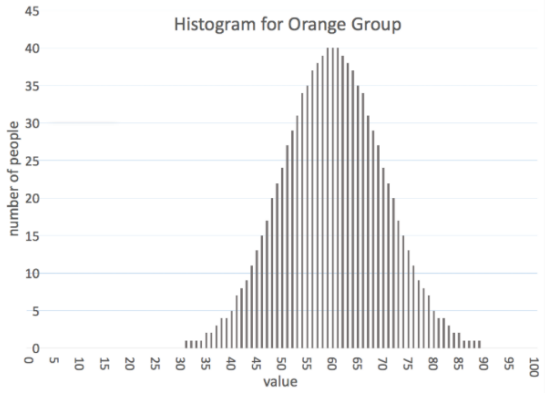


- These figures are read as follows. The horizontal axis shows each possible value, from 0 to 100 . For each possible value, the height of the corresponding bar depicts the number of people who have that value (as shown on the vertical axis).
- For example, there were 24 people in the Green group with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

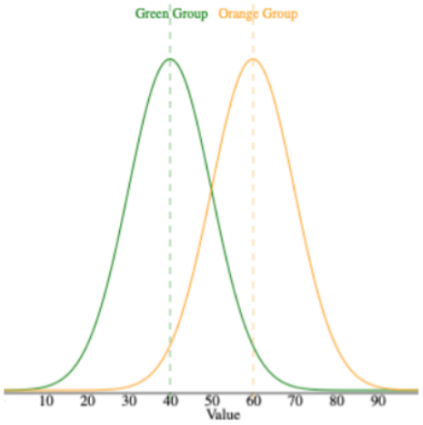
More on the Distributions

- 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:

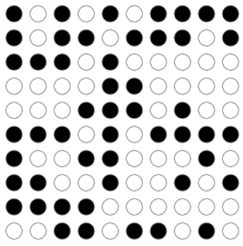
- **1.** The interface will **RANDOMLY** select, with equal chance, **ONE OF THE GROUPS**, either the Green group or the Orange group.

- **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.

Investment Task

In this last task, we will ask you an additional question. You will be paid one cent for every two **TOKEN** you earn in this task.

You have 85 tokens that you can keep or invest in a risky project. The points that you do not invest in the risky project are yours to keep.

The risky project has a 50% chance of success:

- If the project is successful, you will receive 2.5 times the number of tokens you
- invested.
- If the project is unsuccessful, you will lose the amount invested.

Use the slider to decide how many tokens you want to invest in the risky project? Note that you can pick any number between 0 and 85, including 0 or 85:

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85

Investment

Comprehension questions

Subjects had to answer these questions correctly to be able to begin the experiment.