NET WORK LISER

Discussion Paper Series

IZA DP No. 18654 May 2026

Grade Inflation and the Interpretation of Labor Market Signals

Zhizhong Pu ⓡ
Harvard University

Martin Abel ⓡ
Bowdoin College
and IZA@LISER

Jeffrey Carpenter ⓡ
Middlebury College
and IZA@LISER

IZA DP No. 18654 May 2026

Grade Inflation and the Interpretation of Labor Market Signals*

Abstract

We study how grading policies shape employers' interpretations of academic signals. In an experiment, managers observe letter grades from math tests and set wages across three schemes: (C- to A+), (B- to B+), and (A- to A+). As predicted, coarser grading leads managers to place less weight on grades and more on prior beliefs, reducing match efficiency. Departing from predictions, managers assign higher scores to inflated than to compressed grades, as if A-grade candidates are positively selected, and place greater decision weight on inflated As than compressed Bs. Moreover, coarser grading increases reliance on gendered priors, widening gender wage gaps.

JEL classification

J31, J71, D83, C91, I24

Keywords

grade inflation, signaling, hiring experiment, gender wage gap, statistical discrimination

Corresponding author

Jeffrey Carpenter
jcarpent@middlebury.edu

* Our experiment was pre-registered at AEARCTR-0013774 and approved by the Middlebury College Institutional Review Board (IRB ID: 348). The authors have no relevant conflicts or financial interests to disclose. For helpful comments, we thank Alex Chan, Christopher Campos, Katherine Coffman, Dan Stone along with audiences at various seminars. The use of ⓡ signals that the authors' names have been randomized using a uniform distribution.

1 Introduction

The prevalence of grade inflation has raised concerns that academic credentials are losing their informational value as labor market signals. A growing body of research suggests that employers respond strategically when signals become noisier, for example, by adjusting their reliance on alternative indicators of ability (Hansen et al., 2024) or shifting toward more subjective evaluation criteria (Schwager, 2012). However, little is known about how employers actually incorporate coarsened grades into their decision-making. In particular, it remains unclear whether behavioral biases influence how employers interpret inflated grades or whether they treat them in line with standard signaling theory. This distinction has implications for the efficiency and equity effects of grade inflation, since systematic deviations from rational updating may create wage advantages for inflated credentials and amplify the effect of stereotypes about job candidates in ways that persist even when grade inflation is common knowledge.

In this paper, we conduct a two-stage hiring experiment similar to Bohren et al. (2023). We first recruit a sample of 200 college students and recent graduates online to serve as job candidates. They complete a math test that is graded under three different schemes. In the *Control*, participants receive grades from one of nine "bins" ranging from C- to A+. In the *Compress* treatment, grades are compressed to just three bins from B– to B+, while in the *Inflate* treatment, grades are shifted upward to the three bins from A– to A+. The mean-preserving shift from Control to Compress identifies the effect of grade compression, and the variance-preserving shift from Compress to Inflate identifies the effect of grade inflation.

We then recruit (via Connect) 926 participants representative of the U.S. population to act as hiring managers. First, we elicit managers' beliefs about the math ability of the candidates, measured as math SAT scores. Managers are then randomly assigned to one of the three grading schemes and presented with nine profiles. Profiles vary only in the candidate's gender and assigned letter grade, with other characteristics held constant across profiles. Managers complete two incentivized tasks. In the first "signal extraction" stage, they estimate the grading thresholds used in their assigned scheme and each candidate's underlying test score.

In the second "wage-setting" stage, managers make wage offers for each candidate, earning a higher bonus the more closely their wage offers match each candidate's true math ability.

We first examine whether managers correctly perceive the information content of the different grading schemes. We find that managers report nearly identical grading thresholds for the Compress and Inflate treatments, suggesting that they correctly understand that the two schemes are equally informative and do not interpret the different sets of grades as the result of varying grading criteria. Despite this, managers assign significantly higher test scores to candidates in Inflate than in Compress during the signal extraction stage. Rather than attributing the difference to grading standards, managers appear to infer that candidates receiving inflated grades represent a positively selected pool of higher ability applicants, a pattern we refer to as a "positive selection effect". In the control, extracted signals are more dispersed, with top (bottom) performers assigned higher (lower) scores than in either treatment group, consistent with the greater informativeness of the nine-bin grading scheme.

Second, we explore the role that prior beliefs play in the signals that managers extract from coarse grades. Managers with more favorable priors systematically assign higher underlying scores to a given letter grade. Importantly, this signal-dependence on priors is more than twice as large when grades are coarse. This pattern is consistent with the signal loss induced by grade compression. When letter grades are less informative, managers *should* put less decision weight on them when inferring underlying ability; however, we find that managers are more responsive than expected – they adjust the decision weights they use as expected (see below) *and* they adjust the signals they extract from letter grades to be more in line with their priors.

Third, turning to wages, we find that the higher extracted signals in the Inflate treatment translate directly into higher wage offers relative to Compress, while wages in the Control are more dispersed and track underlying ability more closely, given the greater informativeness of the nine-bin grading scheme. To explore the mechanisms underlying these patterns, we estimate the decision weights managers assign to prior beliefs and extracted signals when setting wages, and compare them to a quasi-Bayesian benchmark, which assumes that man-

agers correctly infer the informativeness of grades and combine those signals with prior beliefs. We find that managers deviate substantially: rather than placing the majority of decision weight on prior beliefs as the quasi-Bayesian would, managers place approximately 75% of the weight on the extracted signal and only 25% on the prior, consistent with base rate neglect (Tversky and Kahneman, 1974). Moreover, even though managers report equal confidence in the signals extracted from the Compress and Inflate schemes, they place greater decision weight on inflated than on compressed grades. Taken together, candidates under grade inflation benefit from two reinforcing mechanisms: managers infer positive selection and assign higher extracted signals, and they then apply greater decision weight to those signals when setting wages. These results suggest there is "something special about an A" – it generates a wage premium that standard signaling theory would not predict, even when grade inflation is common knowledge.

Fourth, we examine whether these patterns lead to gendered outcomes. On average, managers believe male candidates score 12 points higher on the math SAT, consistent with the actual performance gap over the past decade. Our framework predicts that as grade signals become coarser and managers place greater weight on prior beliefs, any gender wage gap should widen in proportion to the strength of those gendered priors. While grade inflation and compression only lead to a modest (and insignificant) increase in the gender wage gap, overall, we indeed find substantial heterogeneity by prior beliefs. Among managers in the fourth quartile, with beliefs strongly favoring men, the gender wage gap in the Control is small and statistically insignificant. As signals become coarser, however, a sizable and significant gender gap emerges, reaching approximately 45 SAT points in Inflate treatments. The pattern is symmetric for managers in the first quartile, with beliefs favoring women. These managers assign a wage premium to female candidates that grows substantially as grading becomes coarser and is largest in the Inflate treatment. Further, and as an important comparison, managers in the second quartile, whose priors reflect no gender difference, display no wage gap across any of the three treatment arms, consistent with the predictions of our theoretical framework.

Fifth, in addition to implications for (gender) equity, signal coarsening can also impede firms'

ability to choose the most capable candidates, which can, in turn, reduce match efficiency and overall labor demand (Hsieh and Klenow, 2009). We indeed find that the gap between wage offers and "true" ability significantly widens when grading signals are less precise, as predicted by our quasi-Bayesian benchmark.

Sixth, we complement the manager results with survey evidence on how candidates perceive and respond to grade coarsening. Our candidates correctly anticipate that coarser signals reduce the decision weight managers place on grades, and we also find that higher-performing candidates are more likely to prefer precise grading schemes. However, candidates do not anticipate that managers assign higher decision weights to inflated grades. Turning to real-world signaling behavior, we find that female candidates are substantially less likely to submit precise signals in applications: they are 22 percentage points less likely to report submitting SAT scores and 11 percentage points less likely to disclose their GPA when doing so is optional, even after controlling for actual performance. This is particularly consequential given an additional finding from the manager experiment: when grade signals are coarse, managers rely more heavily on prior beliefs when evaluating female compared to male candidates. This means that women who withhold precise signals are disproportionately exposed to stereotype-based evaluation, which can amplify gender inequities in labor markets where managers hold gendered priors.

Our findings contribute to a small but growing literature that tests whether grade inflation distorts the informational value of academic credentials. Quasi-experimental evidence from grading reforms and discontinuities shows that inflated grades raise short-term earnings, but these effects fade as employers learn workers' true ability (Tan, 2023; Hansen et al., 2024). Indeed, Denning et al. (2025) show that inflating (mean) grades reduces earnings in the long run. We complement this work with one of the first controlled experiments isolating how grade regimes shape manager beliefs. Closely related, Moore et al. (2010) show in a hypothetical admission setting that students benefit from inflated grades because managers under-adjust for grading leniency. Our design separates the effects of grade compression and inflation, holding perceived grading standards constant, and allows us to estimate decision weights that measure how managers integrate prior beliefs and coarse signals. Finding that

managers react more favorably to inflated than compressed grades helps explain the earnings effects and strategic grading behavior observed in other studies.

These results also contribute to a broader literature on how employers interpret signals and how information frictions affect labor market matching. Research shows that reducing information frictions through skill testing (Bassi and Nansamba, 2022; Carranza et al., 2022), reference letters (Abel et al., 2020; Heller and Kessler, 2021), or public performance reviews (Pallais, 2014; Bohren et al., 2019) improves match quality, a pattern we replicate. Related work highlights how information frictions can reinforce inequities. For example, Agan and Starr (2018) find that "ban-the-box" policies that remove information about criminal background reduce hiring for Black applicants, while reducing frictions can disproportionately benefit women (Abel et al., 2020; Bohren et al., 2019). This is particularly relevant in STEM fields, where people hold biased beliefs (see Kahn and Ginther (2017) for a review). Our finding that managers rely more on priors when grades are coarsened for female candidates helps explain how gendered beliefs affect hiring decisions and complements evidence that signal ambiguity disadvantages women in medicine (Sarsons, 2017), finance (Abel et al., 2024), and academia (Sarsons et al., 2021).

Last, our paper relates to the theoretical work on how information design affects screening efficiency and equilibrium behavior. In Bayesian-persuasion and coarse-signaling models, senders may strategically withhold or compress information (Kamenica and Gentzkow, 2011; Boleslavsky and Cotton, 2015).[1] When academic signals become less precise, employers rationally place less weight on them and rely more on priors or institutional reputation, potentially reducing match efficiency and reinforcing inequities (Schwager, 2012). Our experiment provides evidence consistent with these predictions: grade compression and inflation lower the perceived informativeness of grades, increase managers' reliance on prior beliefs and reduce match efficiency.

[1]Related work documents that students respond to the returns to higher grades by selecting into more leniently graded courses (Bar et al., 2009; Tan, 2023). Along similar lines, Denning et al. (2025) find that, across the ability spectrum, grade inflation reduces test scores, suggesting that it lowers students' motivation to study. Similarly, Frankel and Kartik (2019) show that as grade signals become less precise, students have an incentive to invest efforts into "gaming the system," such as using standardized test preparation.

2 Theoretical framework

Our framework develops a quasi-Bayesian model of wage setting in which managers combine prior beliefs with information extracted from coarse letter grades. Because managers may differ in their prior beliefs, the same grade may generate different wage offers across managers. The framework posits two important predictions: coarser grading increases managers' reliance on their priors by reducing the precision of grade-based signals, and heterogeneity in prior beliefs translates into heterogeneity in wages both directly and indirectly through signal extraction. This structure helps organize the empirical analysis that follows.

2.1 Preliminaries

Suppose managers believe that candidate ability is normally distributed, $\theta \sim \mathcal{N}(\mu_p, \sigma_p^2)$, where the prior mean, μ_p, reflects the manager's baseline beliefs about productivity and σ_p^2 captures the manager's uncertainty about that prior. Managers may differ in their prior beliefs about candidate ability for a variety of reasons, including different expectations about the productivity of candidates from different demographic groups or educational backgrounds. We therefore treat μ_p as a parameter that varies across managers.

To learn about any individual candidate's ability, managers ask all candidates to take a test that generates a numerical score, s, which provides a noisy signal of ability. That is, $s = \theta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ captures the magnitude of "testing noise", and therefore $s \mid \theta \sim \mathcal{N}(\theta, \sigma_\varepsilon^2)$. Since θ and ε are independent, the marginal distribution of test scores is $s \sim \mathcal{N}(\mu_p, \sigma_p^2 + \sigma_\varepsilon^2)$, where the variance compounds the manager's uncertainty about ability and the noise in the testing technology. This marginal distribution varies across managers with different prior means, so two managers with different values of μ_p will form different expectations about where a given candidate's score falls in the grade distribution.

Consistent with the information available on college transcripts, the manager does not directly observe s, but instead observes the letter grade that each candidate received on the

test. Specifically, the test administrator (e.g., a college professor) maps the score into a discrete grade, $g = G(s)$, where $G(\cdot)$ partitions the real line into intervals $I_g = [b_g, t_g]$. The manager observes only g, that is, only that $s \in I_g$ and not the exact functional form of $G(\cdot)$. As a result, the manager must form a subjective impression of the grading scheme, $\hat{G}(s)$.

Regarding manager incentives, we assume that the firm is risk-neutral, labor productivity is equal to ability θ, and wages are set competitively. Hence, hiring a candidate of ability θ generates profit $\theta - w$. In the competitive equilibrium, the zero-profit condition implies that the manager sets the wage equal to the candidate's expected ability conditional on the observed grade: $w(g; \mu_p) = \mathbb{E}[\theta \mid g, \mu_p]$.

2.2 Coarse Signal Extraction and Wage Setting

Because the letter-grade signals observed by the manager are coarse, exact Bayesian updating under truncation yields nonlinear posteriors and wage-setting formulas. For tractability, we adopt a reduced-form Gaussian representation of the grading signal.[2] Specifically, upon observing grade g, the manager imputes a signal equal to the conditional expectation of s given that the score falls in interval I_g, evaluated under her subjective score distribution $s \sim \mathcal{N}(\mu_p, \sigma_p^2 + \sigma_\varepsilon^2)$:

$$\hat{s}(g; \mu_p) \equiv \mathbb{E}[s \mid s \in I_g, \mu_p].$$

[2]This approach has antecedents in several related literatures. First, when information processing is constrained, the optimal signal structure takes a Gaussian form (Sims, 2003), and tractable sparse representations of bounded cognition similarly feature Gaussian signal extraction (Gabaix, 2014). Second, the Normal–Normal updating architecture is standard in models of public information (Morris and Shin, 2002). Third, the broader literature on alternatives to Bayesian updating (Ortoleva, 2024), coarse thinking (Mullainathan et al., 2008), and coarse Bayesian updating (Jakobsen, 2025) provides foundations for agents who approximate exact posteriors with simplified representations, as our manager does here. In our setting, $\hat{s}(g; \mu_p)$ and $\sigma_g^2 \equiv \mathbb{E}[\mathrm{Var}(\theta \mid g, \mu_p)]$ are derived from the truncated-normal posterior induced by the true grading rule, and the Gaussian projection's tractability advantage lies in the downstream comparative statics and structural estimation.

Because this expectation is taken with respect to a distribution centered at μ_p, the imputed signal varies across managers: those with higher prior means interpret a given grade as corresponding to a higher underlying score. Computing $\hat{s}(g;\mu_p)$ and the associated posterior variance, $\sigma_g^2 \equiv \mathbb{E}[\mathrm{Var}(\theta \mid g, \mu_p)]$, requires solving the truncated-normal problem. However, rather than carrying the resulting non-Gaussian posterior into the wage-setting stage, we assume that the manager acts as if the posterior were Gaussian with the same mean and variance. She then sets wages using the standard Normal–Normal formula, treating $\hat{s}(g;\mu_p)$ as a noisy point estimate of ability with residual variance σ_g^2. This is the sense in which the model is quasi-Bayesian: the manager correctly extracts the information content of the grade, but sets wages as if the posterior were Gaussian rather than truncated normal.

By construction of the moment-matched Gaussian approximation, $\sigma_g^2 \equiv \mathbb{E}[\mathrm{Var}(\theta \mid g, \mu_p)]$ serves a dual role: it is both the posterior variance of ability after observing the grade, and the noise variance of the believed likelihood $\theta \mid g, \mu_p \sim \mathcal{N}(\hat{s}(g;\mu_p), \sigma_g^2)$. These coincide because the Gaussian approximation is moment-matched to the exact truncated-normal posterior.

The manager updates as if the believed posterior were $\theta \mid g, \mu_p \sim \mathcal{N}(\hat{s}(g;\mu_p), \sigma_g^2)$. In other words, observing a grade is equivalent, from the manager's perspective, to observing a noisy point estimate of ability, where $\hat{s}(g;\mu_p)$ is the estimate and σ_g^2 measures the residual uncertainty after observing the grade. The manager then combines this believed posterior with the prior, $\theta \sim \mathcal{N}(\mu_p, \sigma_p^2)$, using Normal–Normal updating to assign a wage to the observed letter grade according to

$$w(g;\mu_p) = \mathbb{E}[\theta \mid g, \mu_p] = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_g^2}\hat{s}(g;\mu_p) + \frac{\sigma_g^2}{\sigma_p^2 + \sigma_g^2}\mu_p \tag{1}$$

Since $\mathrm{Var}(\theta) = \sigma_p^2$ and $\sigma_g^2 \equiv \mathbb{E}[\mathrm{Var}(\theta \mid g, \mu_p)]$, the law of total variance implies

$$\mathrm{Var}(\mathbb{E}[\theta \mid g, \mu_p]) = \sigma_p^2 - \sigma_g^2.$$

Thus, more informative grading schemes narrow the grade intervals I_g, reducing residual uncertainty σ_g^2 and increasing the dispersion of posterior mean beliefs about ability across letter grades.

The wage equation (1) makes clear that μ_p affects wages through two distinct channels. First, it enters directly through the term $\frac{\sigma_g^2}{\sigma_p^2+\sigma_g^2}\mu_p$. Second, it enters indirectly through the imputed signal $\hat{s}(g;\mu_p)$, since managers with higher prior means extract higher signals from the same grade. As a result, two managers with different prior beliefs may assign different wages to otherwise identical candidates receiving the same grade, with the gap driven by both channels simultaneously.[3]

Since σ_g^2 serves as the noise variance of the believed likelihood for our quasi-Bayesian manager, we can decompose it into two sources of uncertainty. The first is the irreducible testing noise σ_ε^2, which reflects uncertainty about ability that would remain even if the manager could observe the true score directly. The second is the signal extraction uncertainty $\sigma_{s|g}^2$, which reflects the additional uncertainty introduced by having to infer the score from a coarse letter grade rather than observing it directly. Under the Gaussian approximation, these two sources combine additively:

$$\sigma_g^2 = \sigma_\varepsilon^2 + \sigma_{s|g}^2.$$

The first component is a property of the testing technology and is common across managers, while the second reflects how much uncertainty each manager faces in translating a coarse grade into an underlying score, and may therefore vary across managers and grading schemes. Coarser grading schemes widen the grade intervals I_g, increasing $\sigma_{s|g}^2$ and hence σ_g^2, which in turn shifts weight from the grade signal toward the prior in the wage equation.

In summary, the manager first forms a prior about the average ability of candidates, μ_p, and then observes the letter grade that an individual candidate receives on the test, g. Given

[3]This indirect channel dovetails nicely with Fryer et al. (2019) who develop a model in which agents interpret ambiguous signals through the lens of their current beliefs.

her beliefs about the coarse grading system used to map raw test scores into letter grades, she extracts a signal, $\hat{s}(g; \mu_p)$, that reflects both the grade itself and her prior expectations about where scores fall in the grade distribution. The decision weights the manager places on these two sources of information when forming a posterior belief, and hence a wage, depend on how noisy they are. When the prior is less precise, the weight placed on the extracted signal increases. Conversely, when observing a grade leaves substantial residual uncertainty about ability, i.e. when σ_g^2 is large, the manager relies more heavily on her prior in setting the wage. Because μ_p varies across managers, two managers observing the same grade will generally extract different signals and assign different wages, with the gap reflecting both the direct and indirect influence of their differing priors.

2.3 Implications of Grade Compression

The primary comparative static of our framework examines how managers respond to a change in the grading function $G(s)$. Suppose the grader compresses the scheme by collapsing a granular nine-bin system ($C-$ to $A+$) into a coarser three-bin system (e.g., $B-$ to $B+$ or $A-$ to $A+$). In the Blackwell sense, such a transition makes the grading system strictly less informative (Blackwell, 1953). The central empirical question is whether managers respond to this loss of precision by shifting weight from the grade signal toward their prior beliefs, and whether managers with different priors respond differently to the same grade under compression.

First, coarser grading reduces the informativeness of grades, leading managers to rely more heavily on their priors when setting wages. To see this, notice that merging grade bins widens the intervals I_g. By the law of total variance, the residual posterior variance $\sigma_g^2 \equiv \mathbb{E}[\text{Var}(\theta \mid g, \mu_p)]$ weakly increases under compression: if g_3 is a three-bin coarsening of the nine-bin g_9, then $\sigma_{g_3}^2 \geq \sigma_{g_9}^2$. Consequently, the weight placed on the grade signal falls, and the weight placed on the prior rises, so that posterior wages become less dispersed across grades and more reflective of managers' prior beliefs.

Second, beyond compressing wages across grades, coarser grading shifts the mechanism through which prior heterogeneity generates wage dispersion. While the total effect of a unit difference in μ_p on the wage gap between two managers is always unity, the share of that gap operating through the direct channel $\frac{\sigma_g^2}{\sigma_p^2+\sigma_g^2}\mu_p$ is strictly increasing in σ_g^2. Under finer grading, most of the between-manager wage gap is driven by differences in extracted signals $\hat{s}(g;\mu_p)$, which are at least partially anchored to the observed grade. Under coarser grading, a larger share operates through the direct prior term, which is unconstrained by the grade itself.

Third, notice that grade compression also reduces match efficiency — that is, the manager's ability to set wages that track true ability. Since the manager sets $w(g;\mu_p) = \mathbb{E}[\theta \mid g, \mu_p]$, a natural measure of how well wages track true ability is the mean squared error, $\mathbb{E}[(\theta - w(g;\mu_p))^2]$. Under Bayesian updating, $\mathbb{E}[(\theta - \mathbb{E}[\theta \mid g, \mu_p])^2] = \mathbb{E}[\mathrm{Var}(\theta \mid g, \mu_p)] = \sigma_g^2$, so that $\mathrm{MSE} = \sigma_g^2$. Since σ_g^2 weakly increases under compression, match efficiency declines as grading becomes coarser. With coarser grades, posterior means bunch together, managers rely more heavily on their priors, and candidates of both high and low ability are more likely to fall into the same bin and receive the same wage.

3 Methods

The experiment consists of two components: a candidate stage and a manager stage. First, we recruited "candidates" to report their Scholastic Aptitude Test (SAT) scores and complete a math test. Second, "managers" evaluated these candidate profiles and set incentive-compatible wages. In this design, math SAT scores serve as a proxy for underlying ability, while letter grades serve as signals of varying informativeness, given the randomly assigned grading scheme. Detailed protocols and instructions are available in Appendix B.4.

3.1 Candidate sample

In the first stage, we recruited a sample of 200 participants on the survey platform Connect to form our candidate pool.[4] Candidates were all college students or recent graduates aged 21–25 living in the U.S. The sample has a nearly equal gender split (49% female) and an average reported SAT score of 611, slightly above the national average of 605 (see Table A1 for additional descriptive statistics).

After passing ReCAPTCHA and attention checks, candidates completed a 20-question math test modeled after the SAT, administered in ten timed blocks to prevent external assistance. Following Bohren et al. (2023), we use a fixed-payment scheme rather than performance incentives. (A detailed description of the procedures and sample tasks is provided in Appendix B3.) Finally, we explained the hiring experiment to candidates, asked them to predict the behavior of managers, and elicited their preferences for using signals during the job search. On average, our candidates spent about 16 minutes on the experiment and earned $4.38.

We used this data to select a pool of 18 candidates that the managers evaluated in the second part of the experiment. After eliminating participants with implausible SAT scores, we sorted candidates into nine bins based on their performance on the test and randomly picked one male and one female candidate from each bin. This design ensures that variation across profiles is limited to the dimensions of interest: gender, math SAT scores (underlying ability), and test performance (the signal). The final pool of candidates is listed in appendix Table B1. The correlation between test performance and math SAT scores in this sample is 0.54 and is similar across genders.

[4] Gupta et al. (2021) find that the data quality is at least as high on Connect as on the other common online platforms.

3.2 Manager experiment

We recruited 926 participants on Connect to act as managers in our hiring experiment. The sample was designed to be representative of the U.S. population across age, gender, and race (see Table A1, Col. 4 for sample characteristics).[5] The managers are relatively well educated (59.2% completed a four-year college degree or more), and 49.6% have experience participating in the hiring process, including 46.9% who reviewed job applicants.

Managers were asked to imagine they were hiring for an engineering firm where productivity was associated with math ability. To this end, they were tasked with evaluating a common pool of candidates. Before being assigned to a grading treatment, managers reported their baseline beliefs regarding the math ability of the candidate pool. For both male and female candidates, managers estimated the average math SAT score on a scale of 200 to 800. To capture their uncertainty about these priors (σ_p^2), managers also reported the likelihood that the true average SAT score in the candidate sample fell within 10 points of their estimate. This confidence elicitation was bonus-eligible: managers received a 25-cent bonus if their reported likelihood was within 5 percentage points of the realized value (see Figure B10 for details).[6]

Managers were then randomized into one of three treatments: Control (using nine grade bins from C– to A+), Compress (using three bins: B–, B, and B+), or Inflate (using three bins: A–, A, and A+) (Figure B13). The design utilizes a mean-preserving shift (Control vs. Compress) to isolate the effect of grade compression and a variance-preserving shift (Compress vs. Inflate) to identify the effects of grade inflation. Differences in observable characteristics of managers across groups are small and not significant (including their priors), suggesting

[5]Table B2 confirms that our sample is roughly representative of the U.S. population according to the 2023 ACS, with the exception of those older than 75, a common problem with online samples.

[6]This bonus rule does not constitute a proper scoring rule for subjective probabilities. We use this simpler format because proper scoring rules are cognitively demanding and difficult to explain to participants in online survey settings with limited attention (Danz et al., 2022). The resulting responses should therefore be interpreted as an approximate proxy for confidence rather than a precisely elicited subjective likelihood. Because the same elicitation is used in all treatment arms, concerns about imperfect calibration are less likely to affect comparisons across grading regimes.

that randomization was successful (Table A1).

To ensure that the candidate pool was perceived as identical across conditions, we disclosed the three grading regimes used by the professors and that these schemes were applied to the same individuals. We further fixed expectations by informing managers that grades were distributed roughly equally across bins within each grading scheme. This ensures that, under the framework developed in Section 2, managers will form identical extracted scores across both three-bin schemes. Consequently, these two treatments were designed to be equally informative, though both offer a less informative signal than the nine-bin control in the Blackwell sense.

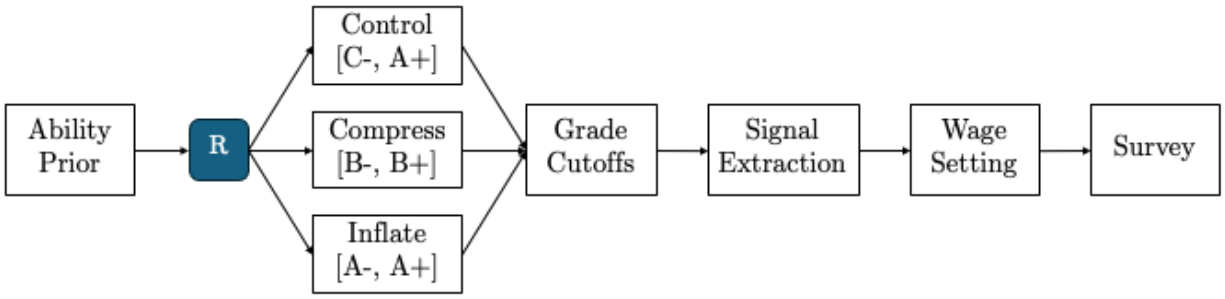


Figure 1: Experimental Design

Following treatment assignment, managers estimated both the candidate pool's test performance and the specific grading thresholds used. To provide context, managers first reviewed three sample math questions to assess difficulty before reporting their expectations for the range (minimum and maximum) and the mean of candidates' numerical scores (Figure B14). Subsequently, managers estimated the score cutoffs they believed the professor used to map these performances into the assigned letter grades (Figure B15). These elicitations provide a measure of the managers' perceived signal-generating function ($\hat{G}(s)$) within their specific grading regime.

Following the threshold estimation, managers completed the signal extraction task, wherein they estimated the numerical test scores for a randomly selected subset of nine candidates. For each candidate, presented sequentially and in random order, managers observed the

assigned letter grade and gender before reporting their estimate of the candidate's underlying performance on our test (in %) (Figure B18). To ensure incentive compatibility, we follow Bohren et al. (2023) and reward these assessments using the Becker–DeGroot–Marschak (BDM) mechanism.[7]

After completing the scoring for all nine candidates, managers provided a self-assessment of their accuracy on average. Specifically, they were asked to report the likelihood that their estimate for any given candidate was within one question (5 pp) of the true performance (Figure B19). Parallel to the elicitation of prior variance, this response is used to infer the manager's signal extraction error variance ($\sigma^2_{s|g}$). To preclude learning effects or adjustments of strategies across the sequence, managers received no feedback on their accuracy during the experiment.

In the final stage of the experiment, managers set wages for the same nine candidates evaluated during the signal extraction phase. The objective was to match each candidate's wage to their underlying ability, proxied by their math SAT score (Figure B20). For each candidate, presented sequentially and in random order, managers again observed the letter grade and gender. To assist in their evaluation, candidate profiles reminded managers of the 0.54 correlation between math test performance and SAT scores, as well as their own previously elicited priors regarding the gender-specific average ability in the candidate pool (Figure B21).

As in the signal extraction stage, elicitation was incentivized via the BDM mechanism to ensure that reported wages accurately reflected the managers' posterior beliefs about candidate ability. Specifically, they "hired" a candidate if their wage offer was at least as large as a randomly drawn "market wage" (both measured on the 200-800 point SAT scale). If hired, the manager's bonus was adjusted upward or downward by subtracting the market

[7]Specifically, for one randomly chosen candidate, we compared the manager's estimated test score to a randomly drawn number between 0 and 100. If the estimate was smaller than the random draw, the manager received their base bonus of 100 cents. If the estimate was equal or larger than the random draw, the base of 100 cents was adjusted by subtracting the random draw and adding the candidate's actual performance (see appendix Figures B16 and B17 for details). This procedure ensures that the manager's expected payoff is maximized by reporting their true belief.

wage and adding the candidate's actual ability. As with signal extraction, managers received no feedback on the accuracy of their wage-setting until the conclusion of the experiment.

Last, we asked the managers a series of demographic questions (gender, age, race, education). We also asked them about their hiring experience at work and about traditional gender roles in the workplace. Overall, the hiring experiment took an average of 16 minutes, for which the participants received a base payment of $2.75, plus the opportunity to earn a bonus. The average total compensation was $3.67, resulting in an average hourly wage of $13.76.

4 Results

This section presents our results, testing the quasi-Bayesian predictions developed in Section 2. Following the order of the manager's experiment (Figure 1), we first describe managers' baseline ability priors and their perceived grading thresholds. We then analyze signal extraction — how managers map letter grades to numerical scores — before evaluating how these beliefs translate into wage-setting decisions. Next, we estimate the decision weights that managers assign to priors and signals in their wage decisions and compare them to the quasi-Bayesian benchmark. We conclude by assessing the welfare implications of signal coarsening on match efficiency.

4.1 Ability priors

Manager ability priors (μ_p) should be independent of treatment assignment but reflect the demographics of the college-educated candidate pool. On average, managers estimated a mean math SAT score of 605.6, which is well-calibrated to the actual candidate pool average of 598 (Table B1) and notably higher than the national average of 508 (College-Board, 2025). As shown in Figure 2 (left panel), these priors are approximately normally distributed with substantial variation (standard deviation = 81.1). Importantly, regression analysis in Table

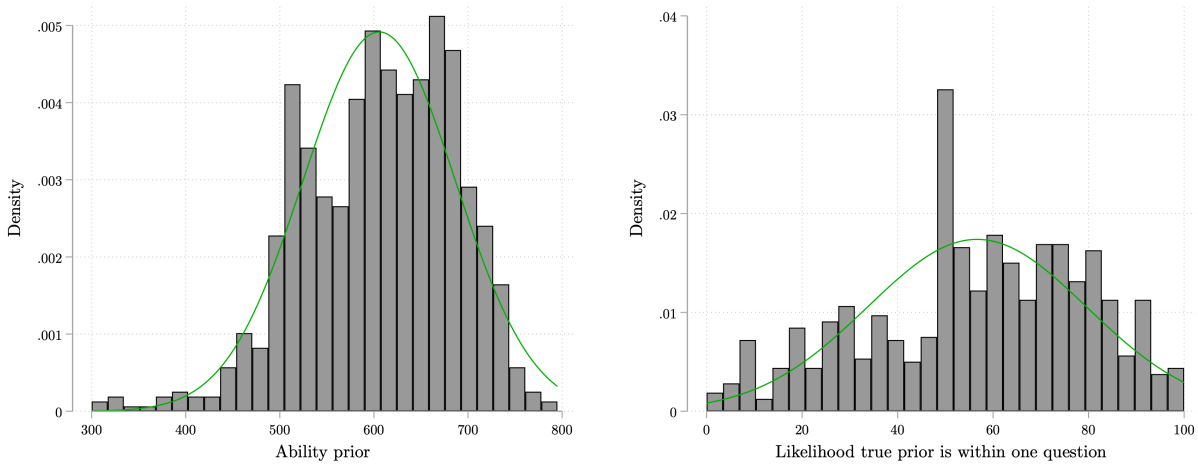


Figure 2: Manager priors about the unobserved ability of candidates.

Notes: The distribution of manager prior expectations about how well candidates did on the math SAT (left). Confidence that the manager's prior is within 10 SAT points of the actual average (right).

A2 confirms that these priors do not vary significantly across grading treatments, indicating that randomization was successful.[8]

The right panel of Figure 2 shows that managers' confidence in these estimates, measured as the perceived likelihood that the true mean in our candidate sample falls within 10 SAT points of their guess, ranges from 0 to 100% with a mean of 56.7% and a standard deviation of 22.9. As with mean beliefs, these confidence assessments do not differ by treatment (Table A2). In sum, managers hold well-calibrated priors that are balanced across experimental treatments.

4.2 Expectations about test performance and grading schemes

While managers were informed that test performance serves as a signal of underlying ability, our theoretical framework (Section 2) does not provide specific predictions regarding how

[8]The only significant demographic determinant of priors is manager gender, as male managers report priors approximately 15 SAT points lower than female managers ($p < 0.01$).

well managers expect candidates to perform on the test, other than the requirement that these expectations remain invariant across grading treatments. As summarized in Table A3, managers expected the lowest, mean, and highest test scores to be 51%, 74%, and 92%, respectively.[9] Interestingly, the expected mean performance of 74% is similar to the average prior SAT belief of 605 out of 800, suggesting some intensity matching and scale comparability, similar to Kahneman and Frederick (2002).

We next examine managers' expectations about the grade cutoffs used by professors. A Bayesian manager would recognize that, because grade bins were designed to contain roughly equal numbers of test scores, the two three-bin grading schemes are equally informative. Accordingly, the manager would map the three-bin schemes onto the nine-bin Control by aggregating the corresponding bins. In particular, the combined width of the three A-range bins in the Control should match the width of the single $B+$ bin in Compress and the $A+$ bin in Inflate, with the same logic applying to the B and C ranges. Under this mapping, the average cutoffs for the aggregated bins should be centered on the A, B, and C thresholds observed in the Control.

The empirical results for these thresholds, presented in Figure 3, largely confirm this reasoning. On average, the managers set cutoffs to maintain bins of equal width and, strikingly, the cutoffs for the two three-bin treatments, Compress and Inflate, are nearly identical.[10] When comparing these treatments to the Control, the average $B-$ in Compress and $A-$ in Inflate align closely with the average C in the Control. Similarly, the comparisons between Compress B, Inflate A, and Control B, as well as Compress $B+$, Inflate $A+$, and Control A, reveal small differences, which is broadly consistent with the mapping implied by a Bayesian interpretation of the three-bin schemes as coarsenings of the nine-bin control. These findings imply that managers correctly understand that grade differences across coarsened treatments arise from fewer grading bins rather than divergent grading standards. Finally, managers

[9]These beliefs were elicited before grading schemes were assigned and are balanced across experimental groups. The actual mean performance was 57% (Table B1), implying that managers systematically overestimate actual test performance. That said, the scores being balanced across treatments and the lack of feedback ensure this does not bias the estimates of treatment differences.

[10]No significant differences exist between these sets of cutoffs at the 5% level, and only one comparison, Control A versus Compress $B+$, is significant at the 10% level.

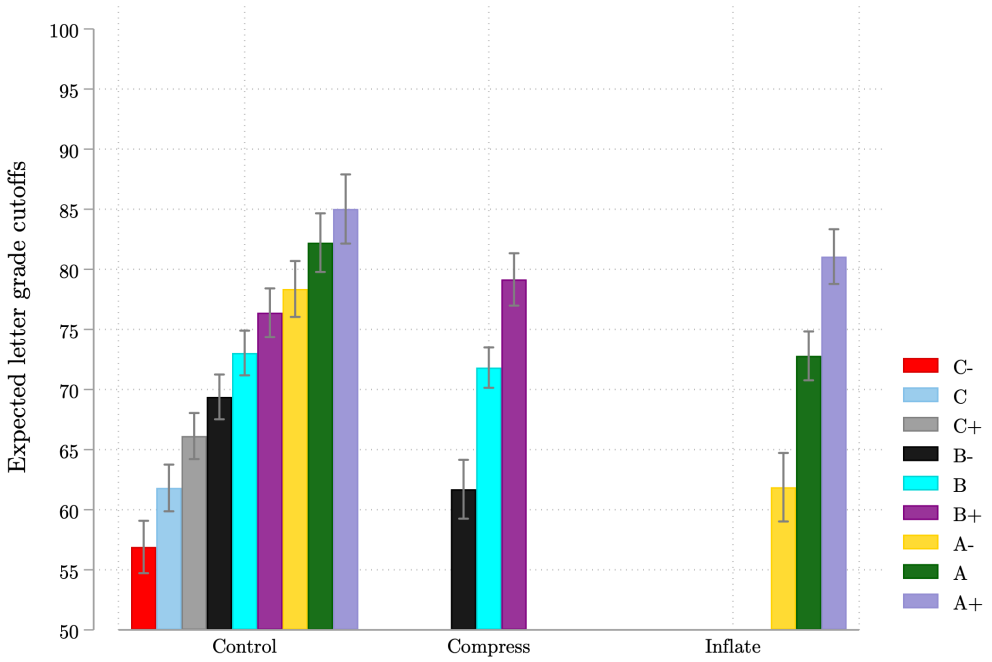


Figure 3: Manager expectations about test grading schemes.

Notes: Manager expectations about where the college professor will set letter grade cutoffs for the math test (by treatment).

appear to anchor their thresholds on their expectations of mean candidate performance: the cutoffs for the median grades (B in Control and Compress and A in Inflate) are set near 73, which is consistent with the reported mean performance expectation.

4.3 Signal extraction

Managers were next asked to map observed letter grades into numerical performance signals, $\hat{s}(g;\mu_p) \equiv \mathbb{E}[s \mid s \in I_g, \mu_p]$. Under the quasi-Bayesian benchmark, a manager should score each grade at the conditional mean of the underlying score distribution within the intervals (I_g) defined by their believed thresholds. Comparing the perceived thresholds in Figure 3 with the extracted scores in Figure 4 reveals strong internal consistency in manager

assessments. Specifically, the correspondence between a manager's perceived grading cutoffs and the scores they assign is strong, as is demonstrated by a pooled Spearman rank-order correlation of $\rho = 0.64$. This indicates that managers use their perceived grading bins to derive numerical signals from coarse letter grades.

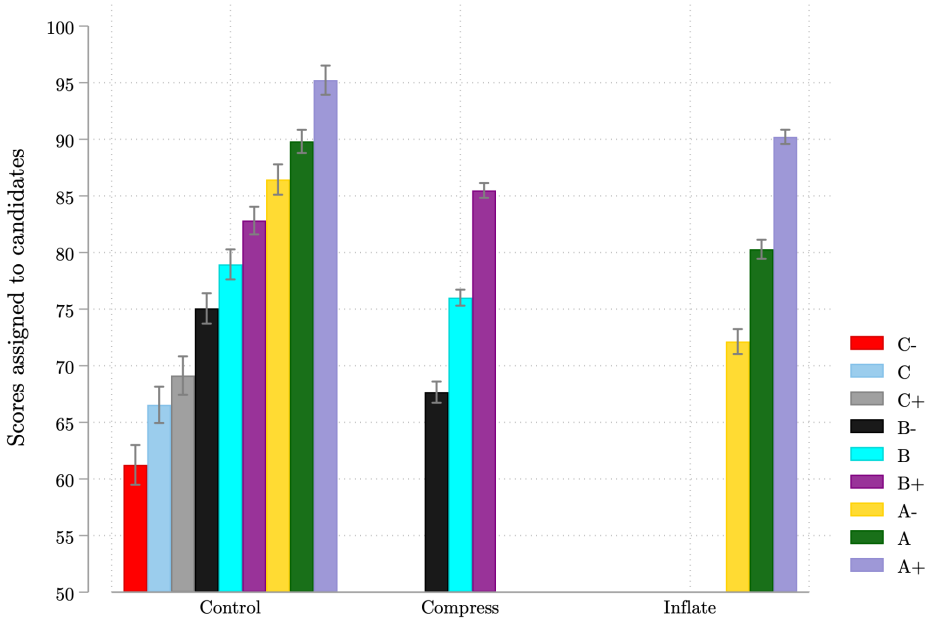


Figure 4: The scores (i.e., signals) managers extract from letter grades.

Notes: The scores that manager assigned to candidates based on the letter grade received on the math test (by treatment).

In Compress, managers' scoring closely aligns with the hypothesized signal extraction procedure. As shown in Figure 4, the average score assigned to a $B-$ is 67.7%, which falls nearly at the midpoint of the perceived interval defined by the 61.7% ($B-$) and 71.8% (B) cutoffs. This consistency extends to candidates receiving a B: given average perceived thresholds of 71.8% and 79.2%, the mean extracted score is 76%. Finally, managers score a $B+$ at an average of 85.5%, positioned roughly five percentage points above the perceived lower threshold for that category.

Results for the Inflate treatment diverge notably from the quasi-Bayesian midpoint bench-

mark. While managers set perceived thresholds for A-, A, and A+ at 61.9, 72.8, and 81.0, respectively, they extract signals at the extreme upper end of these intervals: 72.1 for an A-, 80.3 for an A, and 90.2 for an A+. Consequently, extracted scores in Inflate are 4.3 to 4.6 percentage points higher than those in Compress, a difference that is statistically significant ($p < 0.01$). One interpretation is that managers treat inflated grades as if they reflect positive candidate selection rather than merely lenient grading standards (similar to Moore et al. (2010)). Managers essentially behave as though they are evaluating a "lucky" draw of exceptionally capable candidates, a belief held in spite of instructions stating that candidate quality was held constant across treatments.

The positive selection effect is further supported by a comparison of terciles across treatments. In the Inflate treatment, the scores assigned to the top tercile ($A+$) remain statistically indistinguishable from the top third of the Control group ($p = 0.68$). However, the middle and bottom terciles in the Inflate group receive higher scores than their Control counterparts ($p = 0.12$ and $p < 0.01$, respectively). Conversely, signal coarsening in the Compress treatment penalizes the top two terciles relative to the Control ($p < 0.01$) while providing a modest benefit to the lowest tercile ($p = 0.06$).

Finally, we evaluate the role of managers' prior beliefs in the signal extraction process. As manager ability priors increase, the extracted value of a given letter grade should shift upward accordingly if the extracted signals are anchored on manager priors a la Section 2. Figure 5 illustrates this relationship by grouping manager priors into terciles, showing that managers with the highest priors assign systematically higher scores across all treatments. Regression analysis in Table A4 confirms this positive correlation ($p < 0.01$), which remains robust to the inclusion of letter grade fixed effects (Col. 1-2). Notably, the correlation between priors and extracted signals is significantly more pronounced in the coarse grading regimes – also predicted in Section 2. Estimates indicate that the reliance on priors in the three-bin treatments is nearly three times as strong as in the Control (Col. 3-4), which is consistent with the prediction that managers also rely on their priors to set signals when grades are less informative.

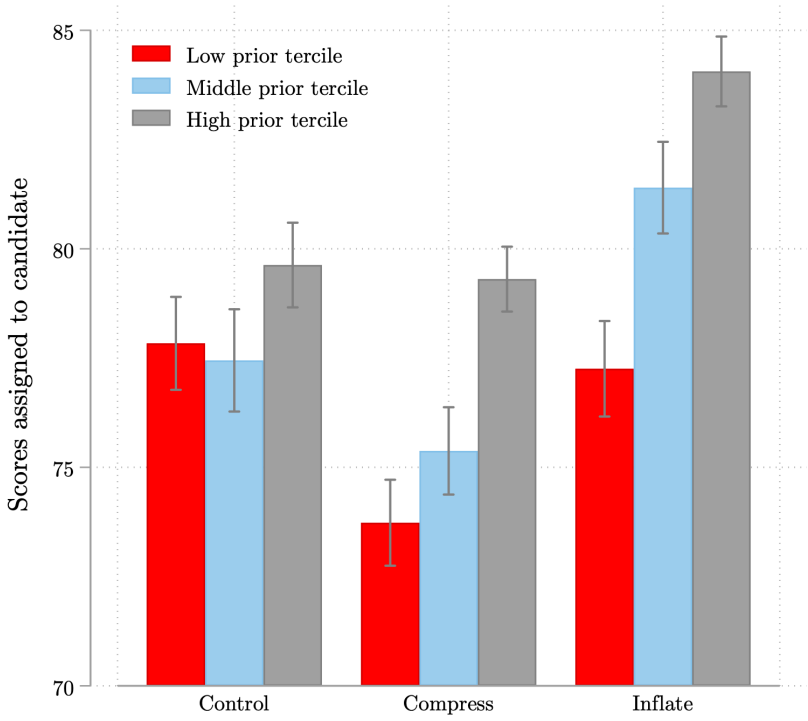


Figure 5: Manager ability priors affect the scores (i.e., signals) extracted from letter grades.
Notes: The scores that manager assigned to candidates based on the tercile of the ability prior distribution (by treatment).

Eliciting manager confidence in their signal extraction provides support for the prediction that signal coarsening reduces the perceived precision of the grading schemes. As established in Section 2, reducing the number of grade bins in Compress and Inflate presents a loss of information, which should lead managers to expect higher variance in their extraction errors. Figure 6 confirms this prediction, as participants in Control are significantly more confident that their extracted signals are within one correct response of the true score than those in either coarse signaling group ($p < 0.01$).[11]

[11]This disparity in informativeness is further substantiated in Table B3, where the adjusted R^2 for the Control is markedly higher than for the coarse treatments, indicating that a more granular grading scheme allows for a more precise mapping of candidate performance.

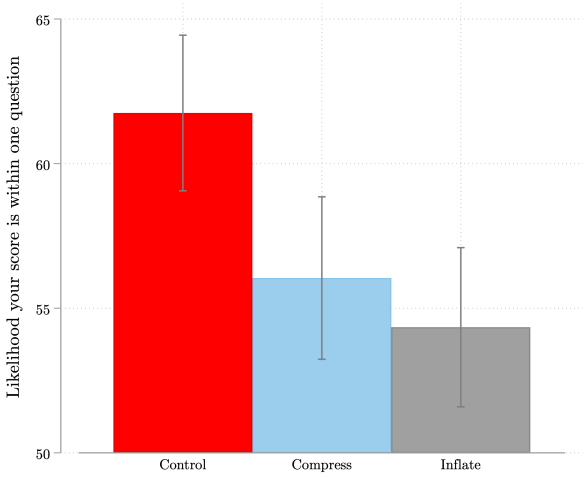


Figure 6: Manager predictions of signal extraction error.

Notes: The likelihood that managers think that their extracted signals are within one correct response of the true signal, on average (by treatment).

4.4 Wage Decisions

After the signal extraction task, managers were incentivized to match their wage offers to the underlying ability (SAT score) of the nine candidates they evaluated. Figure 7 illustrates the average wages assigned across treatment groups and observed letter grades. When pooling across all conditions, the average wage assigned is 594 SAT points, which is close to both the average manager prior of 605 and the true candidate average of 598. Notably, the "middle" letter grade of each treatment – B for Control and Compress and A for Inflate – receives a wage nearly identical to the average prior. This "anchored" wage does not significantly vary across treatments, implying that these median signals trigger minimal belief updating. Furthermore, this "anchor and adjust" heuristic (Tversky and Kahneman, 1974), which encompasses the Bayesian weights as a special case, appears to broadly characterize the wage-setting behavior observed among the managers.

Comparing across grading regimes in Figure 7, average wages in Inflate are 12 to 14 SAT points higher than those in Compress and Control ($p < 0.01$ for both comparisons). This

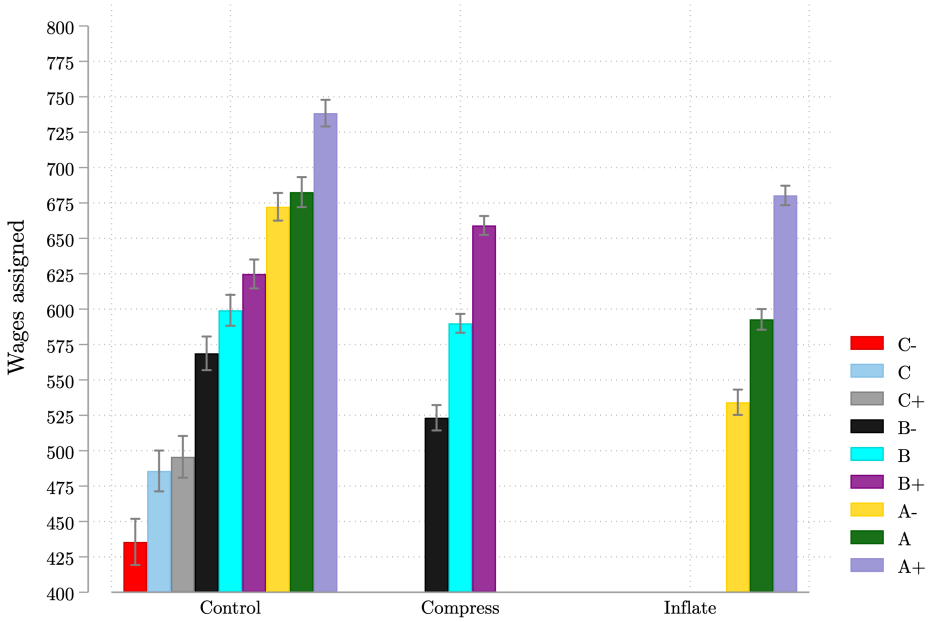


Figure 7: The wages assigned by managers (by candidate letter grade seen).
Notes: Wages are denominated in SAT points to match the units in which ability and priors are measured.

result is consistent with the signal extraction premium assigned to candidates in Inflate and suggests that the perceived positive selection of these applicants translates directly into higher wage offers. Distributional patterns similarly mirror the signal extraction results; signal coarsening penalizes the top of the grade distribution while benefiting the bottom: candidates in the top tercile receive 18 fewer SAT points in Inflate and 39 fewer points in Compress relative to the Control ($p < 0.01$ in both cases). Conversely, candidates in the bottom tercile who receive *C*s under the traditional scheme benefit significantly from coarsening, with average wage increases of 51 SAT points in Compress and 63 points in Inflate ($p < 0.01$ for both).

4.5 Decision weights

According to Section 2, the optimal wage is a weighted average of prior beliefs and extracted signals, where the weights are determined by the relative precision of the baseline prior and the perceived grading signal. We test whether managers set wages consistent with the quasi-Bayesian predictions.

To construct this benchmark, we determine how theoretical managers would set wages given the same information available to our experimental participants. Specifically, given their prior (μ_p) and the signal extracted for each candidate ($\hat{s}(g;\mu_p)$), Section 2.2 predicts that the decision weights will be ratios of the prior variance σ_p^2 and σ_g^2 to the sum of these variances. As established in Section 2.2, σ_g^2 decomposes into testing noise and signal extraction uncertainty: $\sigma_g^2 = \sigma_\epsilon^2 + \sigma_{s|g}^2$, where σ_ϵ^2 would be the only source of residual uncertainty in a standard Bayesian model with directly observed scores, and $\sigma_{s|g}^2$ captures the additional uncertainty introduced by having to infer scores from coarse letter grades. We explicitly elicit manager beliefs about σ_p^2 (Appendix Figure B12) and $\sigma_{s|g}^2$ (Appendix Figure B19 and Figure 6), so the only remaining unknown is σ_ϵ^2, which we recover from the design feature of telling managers that $Corr(\theta, s) = 0.54$.[12] These derived weights are then applied to the nine candidate profiles evaluated by each manager to generate the predicted posterior wages used for comparison.

To estimate these decision weights for our real managers structurally, we use the simple Gaussian functional form in Section 2.2 and OLS.[13] As shown in Panel A of Table 1, our quasi-Bayesian managers react to the two sources of uncertainty in our setting by placing approximately 75% of their decision weight on the prior and 25% on the signal across all treatments. In contrast, our experimental managers in Panel C exhibit the opposite behavior:

[12]Specifically, $Corr(\theta, s) = \frac{\sigma_p^2}{\sqrt{\sigma_p^2 \cdot (\sigma_p^2 + \sigma_\epsilon^2)}} = \frac{\sigma_p}{\sqrt{\sigma_p^2 + \sigma_\epsilon^2}}$, which we solve for σ_ϵ^2 given the elicited σ_p^2 and the known correlation of 0.54.

[13]We also rescale the extracted signals, so that both priors and signals are in SAT points. Because SAT scores are not a linear transformation of the number of correct answers, we use the estimated relationship in our candidate pool and note that it is very similar to the relationship estimated using College Board practice test conversion tables.

Table 1: Decision Weight Analysis

	Combined (1)	Control (2)	Compress (3)	Inflate (4)
Panel A: Quasi-Bayesian Decision Weights				
Prior	0.752	0.748	0.753	0.754
Signal	0.249	0.253	0.248	0.246
Observations	8275	2809	2719	2747
Panel B: Naive-Bayesian Decision Weights				
Prior	0.287	0.258	0.287	0.315
Signal	0.716	0.744	0.717	0.686
Observations	8275	2809	2719	2747
Panel C: Actual Decision Weights				
Prior	0.245***	0.071	0.421***	0.276***
	(0.033)	(0.048)	(0.059)	(0.060)
Signal	0.750***	0.919***	0.566***	0.727***
	(0.033)	(0.047)	(0.060)	(0.061)
Observations	8221	2783	2704	2734
R square	0.961	0.955	0.967	0.964

Notes: Average predicted wages in Panels A and B. In Panel C the dependent variable is the wage assigned by managers. Extracted signals are scaled to SAT scores using a non-linear transformation similar to that used by the College Board. Estimation is done with OLS, constants are suppressed. Clustered standard errors are clustered at the manager level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

they place 75% of the decision weight on the signal and only 25% on the prior. This reversal is consistent with base rate neglect, a well-documented phenomenon where managers overweight new information at the expense of established priors (Tversky and Kahneman, 1974; Griffin and Tversky, 1992; Benjamin, 2019). To investigate this result further, Panel B

of Table 1 hints at a more specific form of base rate neglect. In this "Naive-Bayesian" benchmark, we set $\sigma_\epsilon^2 = 0$ and see that the decision weights of the Naive-Bayesian flip and, in the combined data, look very similar to the weights used by the real managers. This suggests that our real managers more specifically neglect the fact that the scores candidates earn on the test are not perfect signals of ability.[14]

The treatment differences in manager confidence documented in Figure 6 correspond to the varying decision weights reported in Table 1. Qualitatively, the behavior of actual managers aligns with the quasi-Bayesian benchmark, as both exhibit a reduced reliance on signals as the perceived variance of extraction error increases. Quantitatively, however, the observed shifts are significantly more pronounced than the benchmarks predict. Control managers place 92% of the decision weight on the signal and largely ignore their priors, whereas this weight falls to 57% in the Compress treatment. Notably, and in contrast to the benchmarks, managers in the Inflate treatment place an intermediate weight of 73% on the signal. While the slight weight variations in the quasi-Bayesian model are not statistically significant, the larger treatment differences among actual managers are highly significant, which demonstrates that managers are particularly sensitive to signal coarsening.[15]

In summary, while both the quasi-Bayesian benchmark and actual managers reduce their reliance on signals as grades become coarser, actual managers are more responsive to shifts in information precision. This sensitivity is accompanied by a notable behavioral asymmetry: managers place greater decision weight on signals in the Inflate treatment than in the Compress treatment, *despite* reporting similar levels of confidence in those signals. Consistent with Moore et al. (2010), this "something special about an A" effect persists even when controlling for extracted signal values: inflated As carry more weight in wage-setting than equally informative Bs (Table 1, Panel C).

[14]It is also important to note that, while these weights differ from the quasi-Bayesian benchmarks, aggregate manager wages remain relatively well-calibrated; real manager offers are generally within 3 percentage points of the quasi-Bayesian predictions, particularly in the Inflate treatment. See Online Appendix Figure A1 for a summary of these average normalized wage deviations.

[15]For managers, control weights differ from those in the Compress and Inflate treatments at the $p \leq 0.01$ level, and the difference between the Compress and Inflate weights is significant at the $p \leq 0.10$ level

Ultimately, the wage premium for candidates with inflated grades results from two reinforcing mechanisms. First, managers extract higher numerical signals from inflated grades, behaving as if they represent a "positive selection" of high-ability candidates (Section 4.3). Second, managers place greater decision weights on these signals when formulating posterior beliefs. Together, these results suggest that grade inflation provides labor market advantages even when its presence is common knowledge, creating strategic incentives for educational institutions to adopt more lenient grading standards to benefit their graduates (Ostrovsky and Schwarz, 2010).

4.6 Welfare: Match efficiency

As a final assessment of how closely our results align with the theoretical framework, we explore the welfare implications of our grading interventions. Recalling the predictions of Section 2, we expect that grade compression/inflation will lead managers to "bunch" their assessments of candidates, resulting in a narrower range of wage offers that fail to accurately reflect the true productivity of both high- and low-ability candidates. As a result, grade coarsening should lead to a systematic decline in match efficiency.

As predicted, when we examine the mean-squared error of the wages set by both real and quasi-Bayesian managers, we see that both types of managers are less accurate when letter grades are compressed or inflated ($p = 0.07$ for real managers and $p < 0.01$ for quasi-Bayesians). However, the absolute deviations between the wage assigned and a candidate's true ability (SAT) are easier to visualize. On the left of Figure 8, we plot the normalized average absolute deviations. The average deviation in Control of 22% increases by around 1.5 pp (7%) when signals are coarsened through compression or inflation ($p < 0.05$ for both comparisons). As shown in the right panel, overall match quality would have been higher (i.e., the deviations would be lower) if managers had acted more "Bayesian." However, even the quasi-Bayesian managers perform better with the more informative Control grading scheme than in either the Compress or Inflate treatments ($p < 0.01$ for all comparisons).[16]

[16]These findings align with recent empirical evidence showing that reducing information frictions, e.g.,

While the reduction in match efficiency represents a welfare cost primarily affecting firms, informational frictions may also have distributional consequences for labor market equity.

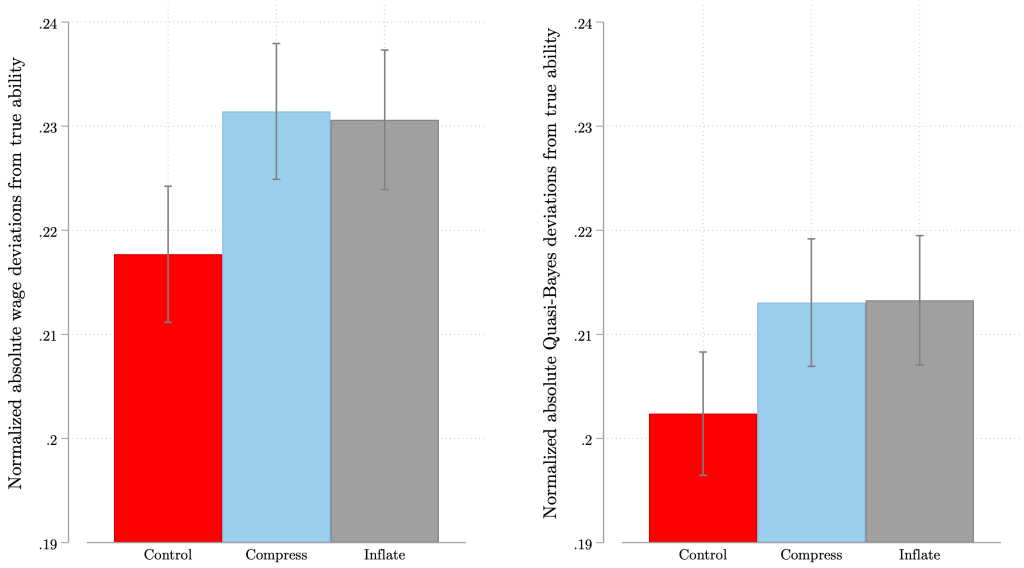


Figure 8: Average absolute deviations from true candidate ability (by treatment).

Notes: Welfare losses are measured as absolute wage (left) or quasi-Bayesian prediction (right) deviations from candidate true ability (i.e., math SAT score). These deviations are normalized to the average manager wage, overall.

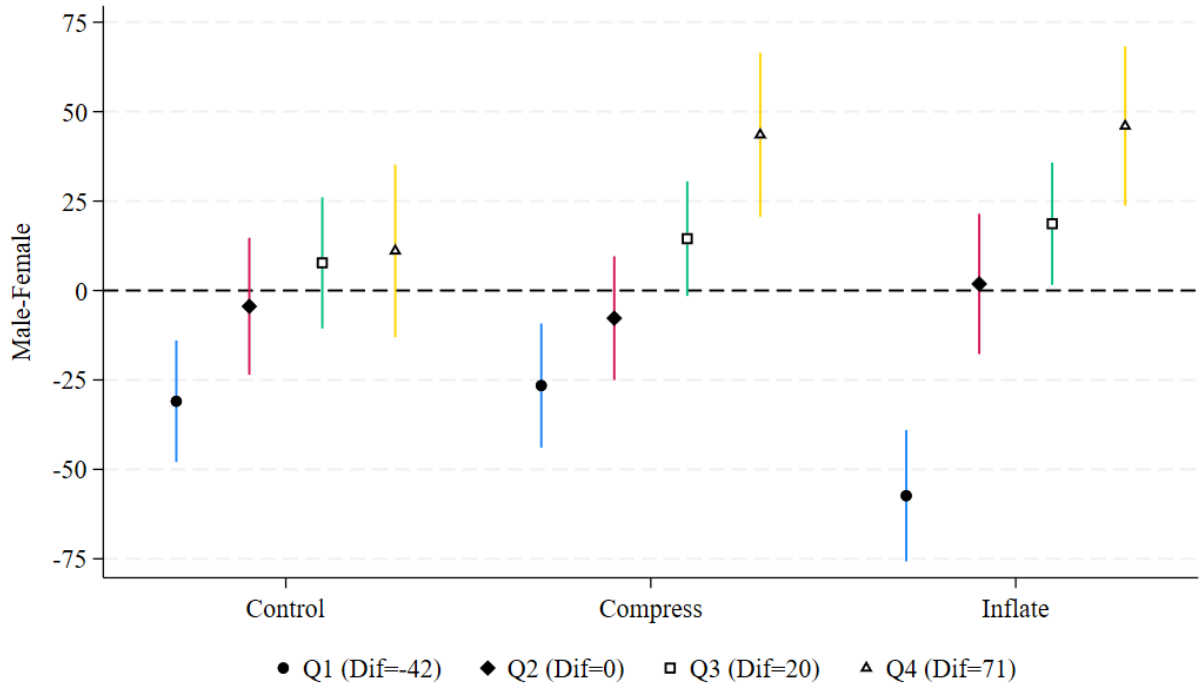
4.7 Implications for Gender Equity

We begin by describing the beliefs managers hold regarding ability by candidate gender. On average, managers report an ability prior (μ_p) of 611 points for men and 599 points for women. This 12-point difference is statistically significant ($p < 0.01$) and consistent with national gender performance gaps in SAT data over the last decade (College-Board, 2025). However, Figure B1 illustrates that this mean difference masks substantial heterogeneity: the median participant in the first quartile of gender differences in beliefs has priors favoring

through reference letters (Abel et al., 2020) or skill testing (Carranza et al., 2022), can increase match quality.

women by 42 points, while the median participant in the fourth quartile favors men by 71 points.

Figure 9: Effect of Grade Coarsening on Gender Wage Gap



Notes: Legend indicates gender difference in ability priors ($\mu_p^{Male} - \mu_p^{Female}$), split into quartiles. Wages measured in SAT score units (200-800 scale). Median difference in prior belief per bin shown in parentheses. 95% error bars are presented. Number of wage decisions per bin: Q1=1884, Q2=1907, Q3=1843, Q4=1843

Within our theoretical framework, information coarsening reduces the precision of the grade signal and leads managers to place greater weight on their prior beliefs. When these priors are gendered, this shift in decision weights can systematically alter the wages assigned to female and male candidates with identical grades. We explore this hypothesis graphically in Figure 9, which plots the gender wage gap in posterior wage beliefs across treatment groups as a function of manager prior differences ($\mu_p^{Male} - \mu_p^{Female}$).

In the Control, the relationship between prior beliefs and the gender wage gap is relatively

flat, with three of the four quartile estimates being indistinguishable from zero. However, as grading signals become coarser in the Compress and Inflate treatments, the relationship between priors and wage offers becomes stronger. Managers with priors favoring women assign (significant) wage premia to female candidates, while those with priors favoring men produce a sizable and significant wage gap in favor of male candidates. Further, these disparities are most pronounced in the Inflate treatment, where even larger gender gaps emerge among managers with the most gendered priors. Anchoring our results, it is noticeable that in the second quartile, where manager priors do not differ by gender, we observe no significant wage gap in any treatment, accordant with the predictions of Section 2.

Regression analysis using a continuous measure of the gender gap in prior beliefs confirms these patterns (Table A6). While no significant aggregate gender wage gap exists when pooling across all managers (Col. 1), the influence of prior beliefs on wage setting becomes significantly more pronounced as signals are coarsened (Col. 2). In the Control group, each point of gender difference in priors translates into only a 0.32-point difference in wage offers. In the Compress group, the role of priors increases by 0.23 points, representing a 72% increase in the marginal effect of beliefs on wages ($p = 0.051$ for the change in slope). The effect is most striking in the Inflate group, where the role of priors effectively doubles to 0.64 compared to the Control group ($p = 0.015$), demonstrating how signal imprecision amplifies the impact of subjective beliefs on gendered labor market outcomes.

Consistent with our framework (Section 2), these wage dynamics appear to operate partially through the signal extraction process. In the Control, prior beliefs do not systematically bias the scores assigned to female versus male candidates (Table A6, Col. 4). As grading becomes coarser, however, extracted scores begin to shift in the direction of managers' prior beliefs. These patterns are most pronounced and statistically significant in the Inflate treatment ($p < 0.01$).

Table A5 reveals an important asymmetry underlying these wage gaps. While coarser grading increases reliance on prior beliefs on average (Col. 2), this effect is concentrated entirely among evaluations of female candidates. For male candidates, the interaction between prior

beliefs and treatment assignment is small and statistically insignificant across both the Compress and Inflate treatments (Col. 4). For female candidates, by contrast, signal coarsening produces a substantial and highly significant increase in prior reliance. This is especially pronounced for inflated grades, where the role of prior beliefs is nearly three times the corresponding estimate for men ($p = 0.06$ for a test of equal coefficient across candidate gender). Importantly, this asymmetry is not driven solely by managers with pro-male priors. Columns 5 and 6 restrict the sample to managers whose prior beliefs favor women, and the pattern is remarkably consistent.

This asymmetry suggests that male candidates are evaluated primarily on the basis of their grade signal, regardless of its coarseness, while female candidates are increasingly judged through the lens of prior beliefs as signals become less informative. Grade inflation and compression, therefore, do not simply reduce information equally for all candidates but specifically increase the degree to which women are evaluated by stereotype rather than performance. This pattern is consistent with evidence that signal ambiguity disproportionately disadvantages women in stereotypically male-associated professional settings, such as surgery, finance, or economics (Sarsons, 2017; Sarsons et al., 2021; Abel et al., 2024).

5 Do Candidates Respond Optimally to Grade Inflation?

Do candidates understand how managers will respond to grade inflation? The broader grade inflation literature largely assumes that students and job seekers correctly anticipate how coarser signals affect employer evaluations, which in turn shapes their strategic behavior, including course selection (Tan, 2023), study effort (Denning et al., 2025), and test preparation (Frankel and Kartik, 2019). Yet direct evidence on the accuracy of these beliefs is scarce, and it remains unclear to what extent beliefs differ across students, which could imply that some adjust optimally to grade inflation while others do not. To get a sense of this, we explained to the candidates the protocol of the hiring experiment and asked them to predict

how much weight managers would place on grades under each of the three grading regimes and which grading scheme they would benefit from the most.[17]

The results suggest that candidates are neither fully naive nor fully sophisticated. On the one hand, they correctly anticipate the direction of the informativeness effect: candidates expect Control managers to assign an average decision weight of 70 points to grades but expect substantially lower weights of 53 and 50 points for compressed and inflated grades, respectively (Table A8). This is consistent with the intuition that coarser signals should carry less evidential value. On the other hand, candidates expect the two three-bin schemes to be treated roughly symmetrically, a prediction that aligns with Bayesian reasoning but overlooks that managers actually place greater weight on inflated As than on compressed Bs, despite reporting similar confidence in both (Section 4.5).

We also find that higher-performing candidates correctly anticipate that they benefit most from precise signals (Table A9). Each additional question a candidate believes they answered correctly on the test they took increases the likelihood of preferring precise grades by 1.6 percentage points ($p < 0.05$, Table A9, Col. 1), suggesting that at least some candidates understand the distributional consequences of grading regimes for their own labor market outcomes. However, this relationship masks heterogeneity by gender. Women are 4.3 percentage points less likely than men to prefer precise grades (Col. 2), and while the relationship between perceived performance and preference for precise grades is strongly positive for men, it is close to zero and statistically insignificant for women (Col. 4).

Turning to behavior outside the experimental setting, we also document gender differences in signaling decisions in real-world applications. For applications in which submitting standardized test scores is optional, women report being 22 pp (39.3%) less likely to submit their scores than men (Table A10, Col. 1). This difference remains large and statistically significant after controlling for actual SAT performance and other covariates (Cols. 2-3), ruling out that it simply reflects differences in underlying ability. We find a similar pattern for GPA disclosure: conditional on actual GPA, women are approximately 11 pp (15%) less

[17]To incentivize responses, participants could earn a 20 cent bonus if they picked the grading scheme they would benefit from the most.

likely than men to voluntarily reveal their GPA ($p < 0.1$, Cols. 5-6). In sum, these patterns indicate that women are less likely to leverage performance signals, the implications of which we will discuss in the next section.

6 Discussion

Our results paint a consistent picture of how grade coarsening shapes managerial behavior. When grading signals are precise, managers place substantial weight on the information they convey. As signals become coarser, managers shift weight away from grades and toward prior beliefs, reducing match efficiency in a manner broadly consistent with our quasi-Bayesian benchmark. Two findings, however, go beyond what the rational benchmark predicts. First, managers in the Inflate treatment extract systematically higher signals than those in the Compress treatment, despite being informed that the two grading schemes are equally informative, behaving as if candidates with inflated grades represent a positively selected pool of higher ability applicants. Second, even after accounting for extracted signals, managers place greater decision weight on inflated As than on compressed Bs, suggesting that the label of an A has a behavioral influence beyond its informational content.

Together, these two mechanisms create a compounding advantage for candidates under grade inflation: they receive both higher extracted signals and greater decision weight, even when inflation is common knowledge. These behavioral departures from the rational benchmark suggest that grade inflation may provide labor market advantages through channels that go beyond what standard signaling theory would predict, creating strategic incentives for institutions to adopt more lenient grading standards even in a world where their consequences are well understood.

While these distortions affect match efficiency broadly, with wage deviations from true ability increasing significantly under coarser grading schemes, the shift toward prior-based evaluation has particularly consequential effects when those priors are themselves shaped by demographic stereotypes. A striking result is that the increased reliance on prior beliefs

under coarser grading is not symmetric across candidate gender, but is concentrated among evaluations of female candidates, regardless of whether the manager's prior favors or disfavors women.[18] When a grade signal is precise, managers engage in individual evaluation and take the signal at face value. When the signal is coarse and ambiguous, however, managers may fall back on heuristics. In a math context where male is the implicit reference category, female gender is a salient characteristic that distinguishes female candidates from the reference group and may therefore play a larger role in evaluation when grade signals are weak or coarse, consistent with representativeness-based accounts of stereotyping (Bordalo et al., 2016). This interpretation is further consistent with dual-process theories of social judgment, in which ambiguity shifts evaluation away from individual-level signals toward category-based beliefs (Fiske and Neuberg, 1990).

This asymmetry in how coarse signals are processed can have direct consequences for the gender wage gap. Our results show that managers place more weight on prior beliefs when assessing coarse signals from female candidates, and that these prior beliefs translate directly into wage offers. Importantly, when grades are precise, the gender gap disappears even among those with strongly gendered priors. This pattern implies that women benefit disproportionately from sending precise signals, particularly in environments where managers hold gendered priors. However, our results also suggest that women may not fully recognize this dynamic: despite the potential advantage of precise signals, they are less likely to report submitting scores from standardized tests in real-world applications.[19] This reluctance cannot be explained by a belief that managers will evaluate applicants fairly. When asked what information should be omitted from resumes, 81.4% of women mention gender, compared to 59.2% of men ($p < 0.001$), indicating that women are more likely to anticipate discrimination by managers.

In settings where gender is visible to employers, the precision of performance signals matters

[18] One interpretation of this pattern is that it reflects the cognitive salience of gender as a heuristic specifically in math-related evaluations, where implicit associations between gender and math ability are well documented (Nosek et al., 2002).

[19] This also has important implications for investment incentives: if women expect their labor market signals to be discounted, they may be less willing to invest in acquiring them (Coate and Loury, 1993).

for how candidates are evaluated. Our results suggest that when individual performance signals are less precise, evaluation shifts toward group-level priors rather than away from them. This dynamic is not unique to our setting. Agan and Starr (2018) document a similar pattern in the context of race: ban-the-box policies that remove criminal background information from job applications were motivated in part to reduce racial disparities in hiring, but instead led employers to rely more heavily on race as a proxy, ultimately reducing the hiring of Black applicants.

These findings have direct implications for ongoing debates about grade inflation and the role of standardized signals in hiring and admissions. Our results suggest that precise, verifiable signals function as an equalizer with respect to stereotype-based evaluation: when grade information is informative, managers evaluate candidates on individual performance and are less influenced by (gendered) priors. Distortions such as grade inflation, pass/fail grading, or test-optional admissions that reduce signal precision risk shifting evaluation away from individual performance and toward group-level priors, disadvantaging candidates from groups subject to negative stereotypes. Reducing informational ambiguity in credentialing may therefore be one of the more effective tools for limiting the role of stereotypes in labor market outcomes, precisely because it operates by modifying the evaluator's information environment rather than requiring a change in their underlying beliefs.

These findings also speak to a broader principle in information design. In Bayesian persuasion models, information structures are typically analyzed for their effects on aggregate efficiency or sender payoffs (Boleslavsky and Cotton, 2015; Kamenica and Gentzkow, 2011). Our results suggest that the distributional consequences of information design deserve equal attention but may not be fully anticipated by the policymakers and institutions that shape them.

References

Abel, M., E. Bomfim, I. Cisneros, J. Coyle, S. Eraou, M. Gebeyehu, G. Hernandez, J. Juantorena, L. Kaplan, D. Marquez, et al. (2024): "Are women

blamed more for giving incorrect financial advice?" *Journal of Economic Behavior & Organization*, 228, 106781.

Abel, M., R. Burger, and P. Piraino (2020): "The value of reference letters: Experimental Evidence from South Africa," *American Economic Journal: Applied Economics*, 12, 40–71.

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

Bar, T., V. Kadiyali, and A. Zussman (2009): "Grade information and grade inflation: The Cornell experiment," *Journal of Economic Perspectives*, 23, 93–108.

Bassi, V. and A. Nansamba (2022): "Screening and signalling non-cognitive skills: experimental evidence from Uganda," *The Economic Journal*, 132, 471–511.

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

Blackwell, D. (1953): "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 24, 265–272.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2023): "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 2023, 1–45.

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

Boleslavsky, R. and C. Cotton (2015): "Grading standards and education quality," *American Economic Journal: Microeconomics*, 7, 248–279.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *Quarterly Journal of Economics*, 131, 1753–1794.

Carranza, E., R. Garlick, K. Orkin, and N. Rankin (2022): "Job search and hiring with limited information about workseekers' skills," *American Economic Review*, 112, 3547–3583.

Coate, S. and G. C. Loury (1993): "Will affirmative-action policies eliminate negative stereotypes?" *The American Economic Review*, 1220–1240.

College-Board (2025): "SAT suite of assesments annual report," Tech. rep., College Board.

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief Elicitation and Belief Manipulation in Surveys," *American Economic Review*, 112, 3WW4–3169.

Denning, J. T., R. Nesbit, N. Pope, and M. Warnick (2025): "Easy A's, Less Pay: The Long-Term Effects of Grade Inflation," .

Fiske, S. T. and S. L. Neuberg (1990): "A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation," *Advances in Experimental Social Psychology*, 23, 1–74.

Frankel, A. and N. Kartik (2019): "Muddled information," *Journal of Political Economy*, 127, 1739–1776.

Fryer, R. G., P. Harms, and M. O. Jackson (2019): "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization," *Journal of the European Economic Association*, 17, 1470–1501.

Gabaix, X. (2014): "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 129, 1661–1710.

Griffin, D. and A. Tversky (1992): "The weighing of evidence and the determinants of confidence," *Cognitive psychology*, 24, 411–435.

Gupta, N., L. Rigotti, and A. Wilson (2021): "The experimenters' dilemma: inferential preferences over populations," *arXiv preprint arXiv:2107.05064*.

Hansen, A. T., U. Hvidman, and H. H. Sievertsen (2024): "Grades and employer learning," *Journal of Labor Economics*, 42, 659–682.

Heller, S. B. and J. B. Kessler (2021): *The Effects of Letters of Recommendation in the Youth Labor Market*, National Bureau of Economic Research.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 124, 1403–1448.

Jakobsen, A. M. (2025): "Coarse Bayesian Updating," *Review of Economic Studies*, forthcoming.

Kahn, S. and D. Ginther (2017): "Women and STEM," Working Paper 23525, National Bureau of Economic Research.

Kahneman, D. and S. Frederick (2002): "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," in *Heuristics and Biases: The Psychology of Intuitive Judgment*, ed. by T. Gilovich, D. Griffin, and D. Kahneman, New York: Cambridge University Press, 49–81.

Kamenica, E. and M. Gentzkow (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

Moore, D. A., S. A. Swift, Z. S. Sharek, and F. Gino (2010): "Correspondence bias in performance evaluation: Why grade inflation works," *Personality and Social Psychology Bulletin*, 36, 843–852.

Morris, S. and H. S. Shin (2002): "Social Value of Public Information," *American Economic Review*, 92, 1521–1534.

Mullainathan, S., J. Schwartzstein, and A. Shleifer (2008): "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, 123, 577–619.

Nosek, B. A., M. R. Banaji, and A. G. Greenwald (2002): "Math = Male, Me = Female, Therefore Math $\neq$ Me," *Journal of Personality and Social Psychology*, 83, 44–59.

Ortoleva, P. (2024): "Alternatives to Bayesian Updating," *Annual Review of Economics*, 16, 545–570.

Ostrovsky, M. and M. Schwarz (2010): "Information disclosure and unraveling in matching markets," *American Economic Journal: Microeconomics*, 2, 34–63.

Pallais, A. (2014): "Inefficient hiring in entry-level labor markets," *American Economic Review*, 104, 3565–3599.

Rabin, M. (2000): "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, 68, 1281–1292.

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

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021): "Gender differences in recognition for group work," *Journal of Political economy*, 129, 101–147.

Schwager, R. (2012): "Grade inflation, social background, and labour market matching," *Journal of Economic Behavior & Organization*, 82, 56–66.

Sims, C. A. (2003): "Implications of Rational Inattention," *Journal of Monetary Economics*, 50, 665–690.

Tan, B. J. (2023): "The consequences of letter grades for labor market outcomes and student behavior," *Journal of Labor Economics*, 41, 565–588.

Tversky, A. and D. Kahneman (1974): "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185, 1124–1131.

A Appendix

A.1 Tables

Table A1: Summary Statistics and Balance of Managers, by Treatment Assignment

	(1) Control		(2) Compress		(3) Inflate		(4) Total		T-test Difference		
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(3)	(2)-(3)
Male	315	0.467	304	0.510	307	0.502	926	0.492	-0.043	-0.035	0.008
		(0.028)		(0.029)		(0.029)		(0.016)			
Age	315	46.376	304	46.316	307	45.844	926	46.180	0.060	0.533	0.472
		(0.914)		(0.880)		(0.855)		(0.510)			
White	315	0.781	304	0.780	307	0.785	926	0.782	0.001	-0.004	-0.005
		(0.023)		(0.024)		(0.023)		(0.014)			
Black	315	0.105	304	0.128	307	0.143	926	0.125	-0.024	-0.039	-0.015
		(0.017)		(0.019)		(0.020)		(0.011)			
Asian	315	0.070	304	0.059	307	0.052	926	0.060	0.011	0.018	0.007
		(0.014)		(0.014)		(0.013)		(0.008)			
College	315	0.622	304	0.582	307	0.570	926	0.592	0.040	0.052	0.012
		(0.027)		(0.028)		(0.028)		(0.016)			
Hiring Experience	315	0.527	304	0.467	307	0.492	926	0.496	0.060	0.035	-0.025
		(0.028)		(0.029)		(0.029)		(0.016)			
Review Applicants	315	0.489	304	0.451	307	0.466	926	0.469	0.038	0.023	-0.015
		(0.028)		(0.029)		(0.029)		(0.016)			
Prior Math (M-F)	315	596.057	304	605.885	307	598.404	926	600.062	-9.828	-2.347	7.481
		(5.102)		(5.057)		(5.026)		(2.923)			
F-test of joint significance (p-value)									0.237	0.537	0.947

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A2: The determinants of manager ability priors.

	(1) Prior	(2) Prior	(3) Likelihood	(4) Likelihood
Compress	6.072	7.884	2.278	2.832
	(6.458)	(6.392)	(1.876)	(1.871)
Inflate	0.884	1.777	2.311	2.784
	(6.544)	(6.504)	(1.832)	(1.812)
Male		-15.738***		-5.055***
		(5.371)		(1.510)
White		5.943		-2.871*
		(6.368)		(1.691)
Less than Bachelors degree		-4.613		0.432
		(5.803)		(1.680)
Advanced degree		-7.920		1.588
		(7.391)		(2.064)
Sexist		-6.822		-0.390
		(6.900)		(2.048)
Participate in hiring		9.643		7.868*
		(15.766)		(4.222)
Review resumes		1.816		-2.809
		(15.647)		(4.178)
Constant	603.297***	604.678***	55.135***	56.412***
	(4.523)	(8.139)	(1.327)	(2.198)
Observations	926	926	926	926
Adjusted R^2	-0.001	0.010	0.000	0.018

Notes: Dependent variable is ability prior or likelihood prior is with one question of truth. OLS (robust standard errors). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Expected candidate test performance.

		(1)	(2)	(3)	$T-test$ (p $value$)		
	$Overall$	$Control$	$Compress$	$Inflate$	$(1)v(2)$	$(1)v(3)$	$(2)v(3)$
Minimum score	51.17	50.46	51.78	51.30	0.41	0.61	0.77
Mean score	74.04	73.48	74.82	73.83	0.21	0.75	0.34
Maximum score	92.32	93.00	92.24	91.70	0.47	0.22	0.64

Notes: average responses (overall and by treatment) to questions asking experimental managers to predict how well the candidates did on the math test. Responses were the percentage of correct answers.

Table A4: Effect of Priors on Signal Extraction

	Signal Extraction (0-100)			
	(1)	(2)	(3)	(4)
Prior (0-100)	0.141***	0.143***	0.069**	0.075**
	(0.024)	(0.024)	(0.033)	(0.032)
Prior × Compress			0.107**	0.099*
			(0.052)	(0.051)
Prior × Inflate			0.119**	0.108*
			(0.060)	(0.060)
Compress			-2.082**	-2.807***
			(0.817)	(0.816)
Inflate			2.625***	-9.558***
			(0.882)	(0.867)
Observations	8275	8275	8275	8275
Control Mean	78.3	78.3	78.3	78.3
Control SD	16.7	16.7	16.7	16.7
Adj R-Square	0.02	0.30	0.03	0.33
Grade F.E.	No	Yes	No	Yes

Notes: The dependent variable is the score assigned by the manager in the signal-extraction stage. The main explanatory variable is the manager's prior belief about candidate ability, measured as the expected average math SAT score of the candidate pool. Compress and Inflate are treatment indicators, with Control omitted. Columns 1–2 include letter-grade fixed effects; Columns 3–4 interact priors with coarse-grade treatments. The unit of observation is a manager–candidate evaluation. Standard errors are clustered at the manager level.. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Effect of Priors on Signal Extraction

	Signal Extraction (0-100)					
	(1)	(2)	(3)	(4)	(5)	(6)
Prior (0-100)	0.132***	0.056	0.045	0.082*	-0.008	0.122
	(0.025)	(0.036)	(0.039)	(0.046)	(0.056)	(0.076)
Prior × Compress		0.096*	0.125**	0.049	0.142	0.051
		(0.052)	(0.055)	(0.066)	(0.096)	(0.121)
Prior × Inflate		0.141**	0.203***	0.069	0.171*	0.023
		(0.066)	(0.069)	(0.081)	(0.100)	(0.137)
Compress		-1.360	-2.383***	-0.250	-2.690*	-0.089
		(0.859)	(0.901)	(0.936)	(1.497)	(1.380)
Inflate		2.618***	2.012**	3.368***	3.670**	3.136**
		(0.946)	(1.003)	(1.020)	(1.514)	(1.487)
Observations	7472	7472	3752	3720	1176	1207
Control Mean	78.2	78.2	78.2	78.2	79.2	79.2
Control SD	16.9	16.9	16.9	16.9	14.8	14.8
Adj R-Square	0.01	0.03	0.04	0.02	0.05	0.03
Candidate Gender	Pooled	Pooled	Women	Men	Women	Men
Prior Beliefs	Pooled	Pooled	Pooled	Pooled	F>M	F>M

Notes: The dependent variable is the score assigned by managers to candidates during the signal extraction stage, measured on a 0-100 scale. The key explanatory variable is the manager's prior belief about candidate ability, measured as the expected average math SAT score of the candidate pool. Compress and Inflate are treatment indicators, with Control omitted. Columns vary by candidate gender and manager subsample, as indicated in the table. The unit of observation is a manager–candidate evaluation. Standard errors are clustered at the manager level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Effect of Gender Priors on Signal Extraction and Wages

	Wage Offer		Signal Extraction	
	(1)	(2)	(3)	(4)
Female	3.852	8.717*	1.424***	1.389***
	(0.450)	(0.073)	(0.005)	(0.007)
Compress	6.485	3.662	-0.102	-0.352
	(0.409)	(0.644)	(0.914)	(0.699)
Inflate	19.464**	16.505**	3.411***	3.146***
	(0.015)	(0.043)	(0.001)	(0.002)
Fem x Compress	-7.963	-9.088	-2.182***	-2.109***
	(0.261)	(0.166)	(0.001)	(0.002)
Fem x Inflate	-6.904	-2.917	-1.630**	-1.006
	(0.360)	(0.675)	(0.029)	(0.163)
Prior: Male-Fem		-0.095		-0.020*
		(0.195)		(0.067)
Female x Prior Diff		-0.320***		0.002
		(0.000)		(0.831)
Prior × Compress		0.233**		0.013
		(0.036)		(0.372)
Prior × Inflate		0.225		0.015
		(0.156)		(0.361)
Fem x Prior Diff × Compress		-0.230*		-0.012
		(0.051)		(0.281)
Fem x Prior Diff × Inflate		-0.319**		-0.041***
		(0.015)		(0.002)
Observations	7477	7477	7472	7472
Control Mean	587.5	587.5	78.2	78.2
Control SD	146.7	146.7	16.9	16.9
Adj R-Square	0.00	0.02	0.01	0.02

Notes: In Col. 1-2, the outcome is the wage assigned by the manager to the candidate, measured in math SAT points; in Col. 3-4, the outcome is the manager's estimated test score for the candidate (in percent). The key explanatory variable is the manager-level gender gap in prior beliefs, measured as the difference between the expected average math SAT score for male and female candidates. Compress and Inflate are treatment indicators, with Control omitted. The unit of observation is a manager–candidate evaluation. Standard errors are clustered at the manager level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A7: Gender Signal and Wage Gap

	Signal Extraction				Wage Offer			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.17	1.42***	1.70**	0.91	-1.04	3.85	-10.21	28.67***
	(0.29)	(0.51)	(0.70)	(0.83)	(3.00)	(5.10)	(7.04)	(6.95)
Compress	-1.20	-0.10	-0.22	-0.18	2.47	6.48	11.77	0.51
	(0.87)	(0.94)	(1.35)	(1.39)	(7.22)	(7.86)	(10.59)	(12.71)
Inflate	2.59***	3.41***	4.12***	2.75*	15.99**	19.46**	15.39	22.42
	(0.97)	(1.02)	(1.45)	(1.52)	(7.35)	(8.00)	(10.20)	(14.56)
Female × Compress		-2.18***	-2.76***	-1.61		-7.96	-14.09	-7.64
		(0.66)	(0.90)	(1.04)		(7.07)	(9.64)	(10.41)
Female × Inflate		-1.63**	-3.63***	1.65		-6.90	-19.61**	14.09
		(0.74)	(0.99)	(1.21)		(7.55)	(9.63)	(11.70)
Observations	7472	7472	4363	2383	7477	7477	4369	2383
Control Mean	78.3	78.3	78.3	78.3	588.5	588.5	588.5	588.5
Control SD	16.7	16.7	16.7	16.7	146.8	146.8	146.8	146.8
Adj R-Square	0.01	0.01	0.01	0.02	0.00	0.00	0.01	0.03
Prior	Pooled	Pooled	M > F	F > M	Pooled	Pooled	M > F	F > M

Notes: The dependent variable is the manager's extracted score in Columns 1–4 and the manager's wage offer in Columns 5–8. Extracted scores are measured as estimated candidate test performance (in percent); wage offers are measured in math SAT points. Compress and Inflate are treatment indicators, with Control omitted. Columns 3 and 7 restrict the sample to managers whose priors favor male candidates ($M > F$); Columns 4 and 8 restrict the sample to managers whose priors favor female candidates ($F > M$). The unit of observation is a manager–candidate evaluation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A8: Summary Statistics

	Mean	SD	Min	Max	N
Panel A: Who Benefits Most					
Control	1.09	3.63	-10	10	200
Compress	1.12	3.18	-10	10	200
Inflate	0.87	4.07	-10	10	200
Panel B: Weights on Grades					
Control	70.11	20.51	8	100	200
Compress	52.97	20.51	3	94	200
Inflate	49.91	27.83	0	100	200

Notes: Panel A reports whether workers believe women or men benefit more from different grading regimes on a scale from -10 (women) to 10 (men). Panel B reports average scores of the perceived weight that employers put on grades on a scale from 0 to 100.

Table A9: Who Prefers Precise Signals?

	1 = Prefer Precise Grades			
	(1)	(2)	(3)	(4)
Perceived Performance	0.016**		0.015**	0.023**
	(0.007)		(0.007)	(0.011)
Female		-0.071	-0.039	-0.037
		(0.071)	(0.072)	(0.072)
Performance x Female				-0.014
				(0.015)
Observations	200	200	200	200
Mean Male	0.55	0.55	0.55	0.55
Control SD	0.50	0.50	0.50	0.50
Adj R-Square	0.02	0.00	0.02	0.01

Notes: The sample are 200 participants who completed the math test and serve as workers in our experiment. The dependent variables are binary variables measuring whether workers would rank the grading scheme with precise grades (control) first. Perceived performance is measuring how many questions (out of 20) participants believe they answered correctly. The variable is demeaned for this analysis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A10: Do Workers Choose to Reveal Signals?

	1 = Reveal SAT			1 = Reveal GPA		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.211***	-0.234***	-0.217***	-0.026	-0.117**	-0.126*
	(0.069)	(0.067)	(0.072)	(0.065)	(0.057)	(0.064)
SAT (std.)		0.122***	0.124***			
		(0.047)	(0.045)			
GPA					0.685***	0.682***
					(0.068)	(0.077)
Observations	200	200	200	200	200	200
Mean Male	0.55	0.55	0.55	0.71	0.71	0.71
Control SD	0.50	0.50	0.50	0.45	0.45	0.45
Adj R-Square	0.04	0.09	0.10	-0.00	0.27	0.25
Control Var	No	Yes	Yes	No	No	Yes

Notes: The sample are 200 participants who completed the math test and serve as workers in our experiment. The dependent variables are binary variables measuring whether workers would submit their SAT score (Col. 1-3) and grade point average (Col. 4-6) in an application if it was optional. SAT scores are standardized. GPA is measured on a 0 to 4 scale. Control variables include age, race, and college major. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2 Figures

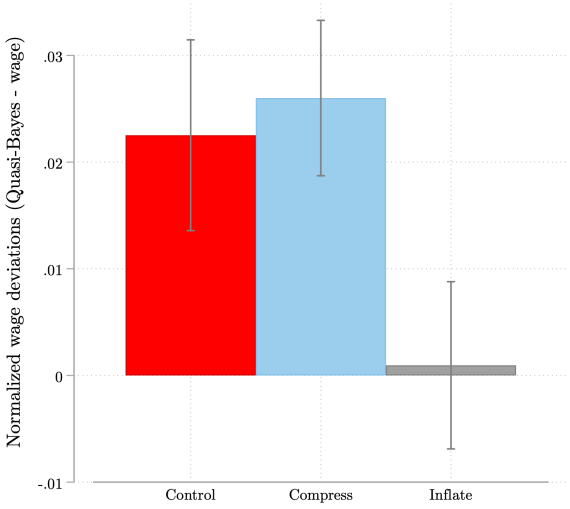


Figure A1: Average normalized wage deviations from the predicted benchmarks (by treatment).

Notes: Wages are denominated in SAT points to match the units in which ability and priors are measured. The deviation is *Quasi-Bayes prediction* − *wage* and normalized to the average real manager wage.

B Online Appendix

B.1 Tables

Table B1: Candidates and grading schemes for the manager experiment.

Bin	**Female**		**Male**		**Grade Scheme**		
	Test Score	Math SAT	Test Score	Math SAT	Control	Compress	Inflate
1	3	400	3	620	C–	B–	A–
2	5	600	6	250	C	B–	A–
3	7	550	8	620	C+	B–	A–
4	9	750	10	560	B–	B	A
5	11	500	11	650	B	B	A
6	13	400	14	720	B+	B	A
7	15	600	16	520	A–	B+	A+
8	18	780	17	650	A	B+	A+
9	19	800	19	800	A+	B+	A+

Notes: the mean test score for female candidates was 11.1 and it was 11.6 for male candidates. The mean math SAT scores were 597.8 and 598.9 for female and male candidates, respectively. The overall correlation between test scores and math SAT scores is 0.54.

Table B2: Manager characteristics.

	ACS	*Control*		*Compress*		*Inflate*	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Age (18 - 34)	0.29	0.28	0.45	0.30	0.46	0.28	0.45
Age (35 - 54)	0.33	0.35	0.48	0.36	0.48	0.42	0.48
Age (55 - 74)	0.29	0.34	0.47	0.32	0.47	0.28	0.45
Age (75+)	0.09	0.02	0.15	0.02	0.13	0.01	0.11
Female	0.51	0.53	0.50	0.49	0.50	0.51	0.50
White	0.71	0.78	0.45	0.78	0.44	0.79	0.44
Participants	-	315		304		307	

Notes: means and standard deviations of participant demographic characteristics matched to the U.S. population. Characteristics are shown separately for each treatment. The column ACS contains values of these demographics as reported in the 2023 American Community Survey. The ACS age figures reflect that fact that our participants must be at least 18.

Table B3: The scoring (and information content) of letter grades.

	(1) Control	(2) Compress	(3) Inflate
Saw A+	33.984***		18.082***
	(1.105)		(0.726)
Saw A	28.575***		8.153***
	(1.086)		(0.448)
Saw A-	25.206***		
	(1.159)		
Saw B+	21.582***	17.817***	
	(0.947)	(0.743)	
Saw B	17.711***	8.345***	
	(1.008)	(0.443)	
Saw B-	13.821***		
	(0.925)		
Saw C+	7.891***		
	(1.053)		
Saw C	5.304***		
	(1.045)		
Constant	61.240***	67.667***	72.134***
	(1.083)	(0.797)	(0.937)
Observations	2809	2719	2747
Adjusted R^2	0.413	0.272	0.229

Dependent variable is score; OLS (clustered standard errors) on manager. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Figures

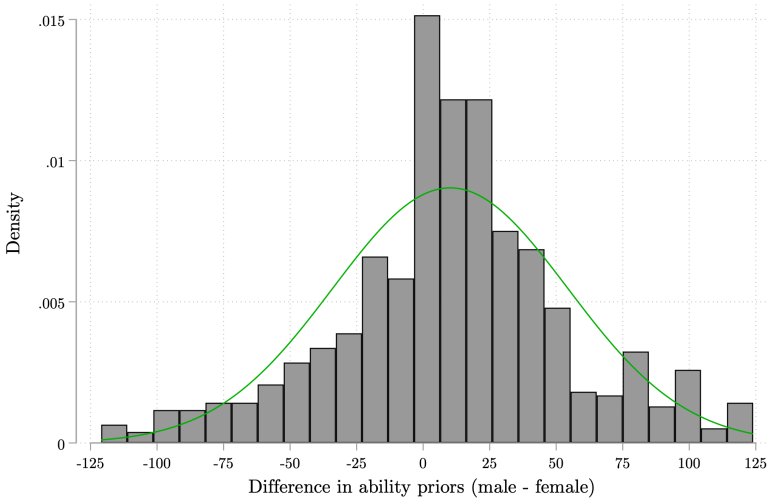


Figure B1: The distribution of prior differences ($\mu_{male} - \mu_{female}$).
Notes: Male prior minus female prior for each manager.

B.3 Candidate Experimental: Protocol and Instructions

The Connect solicitation for the candidate experiment indicated that the study was for people in the U.S. between 21 and 25 who recently graduated from college (or were in college) and took the SAT exam. This information was confirmed in an initial screening question (Figure B2). Candidates were then told about the math test they would take (Figure B3) and how they would be paid for that part of the experiment.

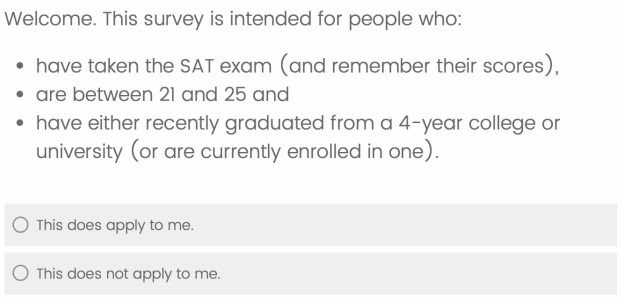


Figure B2: Candidate screening question.

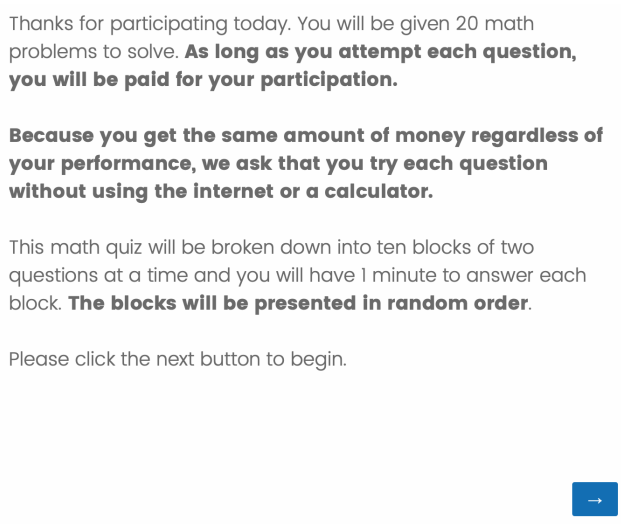


Figure B3: Candidate test instructions.

The test was based on the SAT math test. It was 20 multiple choice questions presented in 10 randomly-ordered two-question blocks. The questions from one of the blocks are presented in Figure B4. Directly after completing the test, candidates were asked to assess their performance (Figure B5).

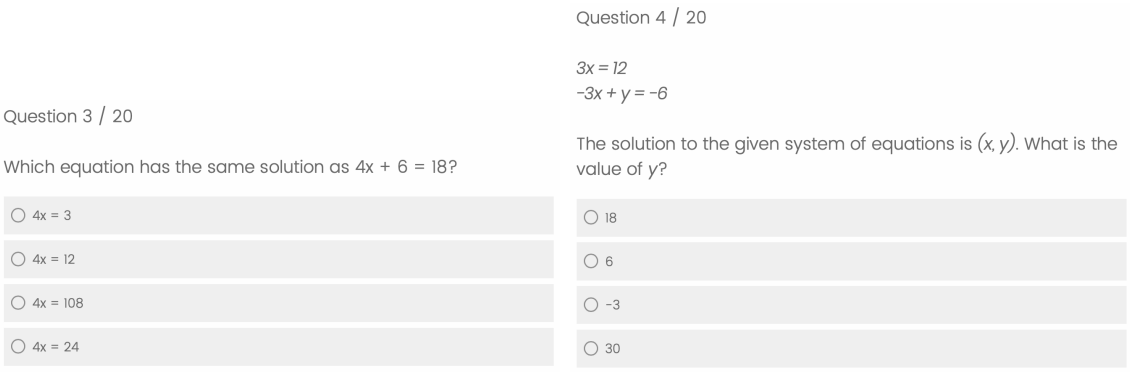


Figure B4: Candidate sample test questions.

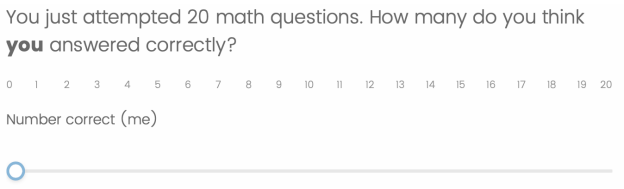


Figure B5: Candidate expected performance.

In the second part of the candidate experiment, we asked participants to try to anticipate the results of the manager experiment. They were told how the manager experiment would work, what the treatments would be and that managers would have an incentive to match wages to underlying ability (i.e., math SAT performance). With this background, we asked the candidates to predict the decision weight that the managers would put on the letter grades they would be shown, by treatment (Figure B6). They were then asked to predict whether men or women candidates who did equally well on the test would be assigned higher or lower wages by the managers (Figure B7). Lastly, we incentivized candidates to accurately predict

the grading treatment in which they would receive the highest wage from the managers (Figure B8).

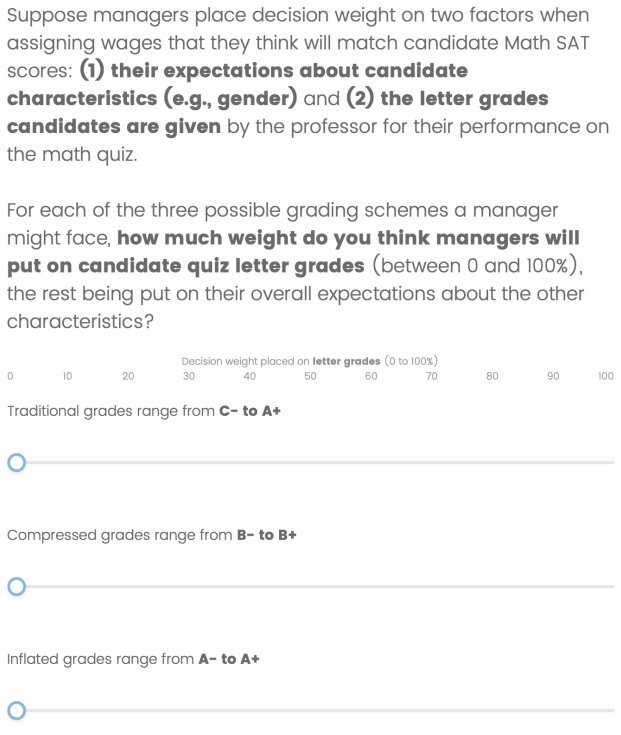


Figure B6: Candidate expected manager decision weights.

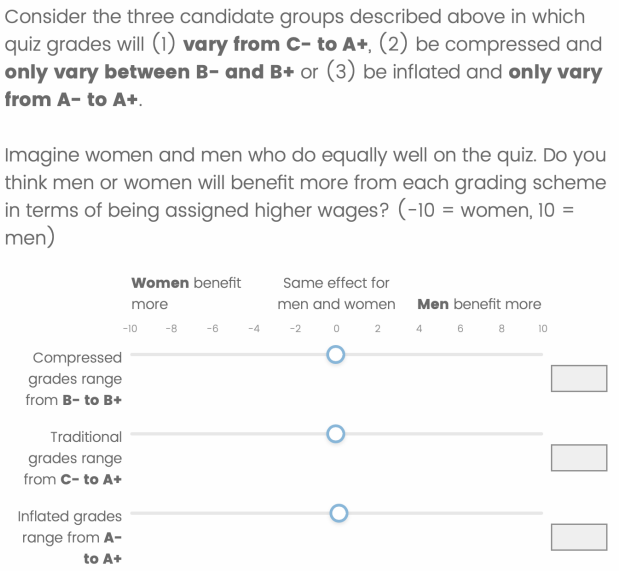


Figure B7: Candidate expected grading scheme beneficiaries.

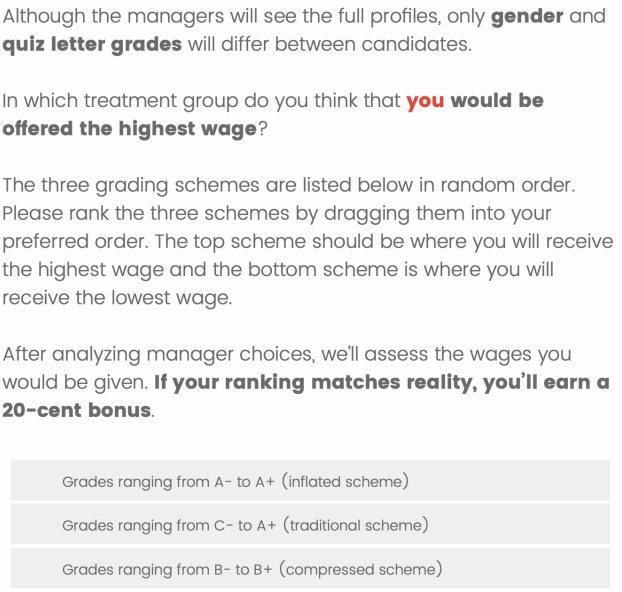


Figure B8: Candidate preferred treatment.

In the last part of the candidate experiment, we asked participants to give their priors about what the average math SAT score would be in the pool of candidates – by gender and what the managers would think these averages would be. We then asked a few questions about their general labor market experience (interviewing, recommendations letters, what should be revealed on resumes) and work experience.

B.4 Manager Experimental: Protocol and Instructions

After consenting to participate, respondents in the manager experiment had to pass both a recaptcha puzzle and and attention check (Figure B9). Those who passed both checks, were then introduced to the experiment with the instructions in Figure B10, which provided context for the decision making task, explained how payments and bonuses would be made, introduced the treatments and explained what the managers would be doing.

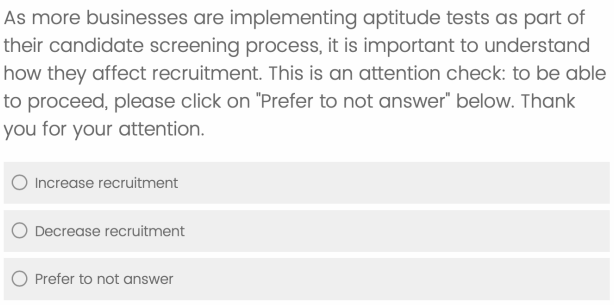


Figure B9: Manager attention check.

The managers were then asked for their priors about the average "ability" (i.e., math SAT score) of both men and women (Figure B12 shows the elicitation for the male candidates). Directly after the ability prior question, managers were asked how likely it was that their prior was close to the true average ability (the responses to which we use to create a proxy for the prior variance).

The next stage of the experiment was designed to elicit manager beliefs about the test that

candidates took and how it was graded. This was explained using the text in Figures B13 and B14, the second of which also shows the prompts asking managers to report how well they thought candidates did on the test (min score, max score and average score). In Figure B15, we show how managers were asked to report (in the Compress treatment) the letter grading cutoffs they thought were used by the professor grading the exams.

The experiment then transitioned to the signal extraction part. Here managers were told that they would be asked to predict the underlying score that nine randomly picked candidates got on the test, seeing just their gender and letter grade (see Figure B16). They were also introduced to the B-D-M procedure used to incentivize their assessments.[20] To get a sense of how well the managers understood the incentives, they were given a comprehension check (Figure B17), along with an explanation of the correct answer. In Figure B18, we provide an example of the elicitation used for signal extraction. After assessing nine of the candidates (matched at random), managers were asked to reflect on the signal extraction part of the experiment and give us some sense of their subjective beliefs about the signal extraction error variance (Figure B19).

[20]Interpreting elicited scores and wages as posterior means assumes risk neutrality over the bonus lottery. In our setting, however, the stakes are small, so explaining the observed patterns through treatment-varying risk premia would require implausibly strong local risk aversion (Rabin, 2000).

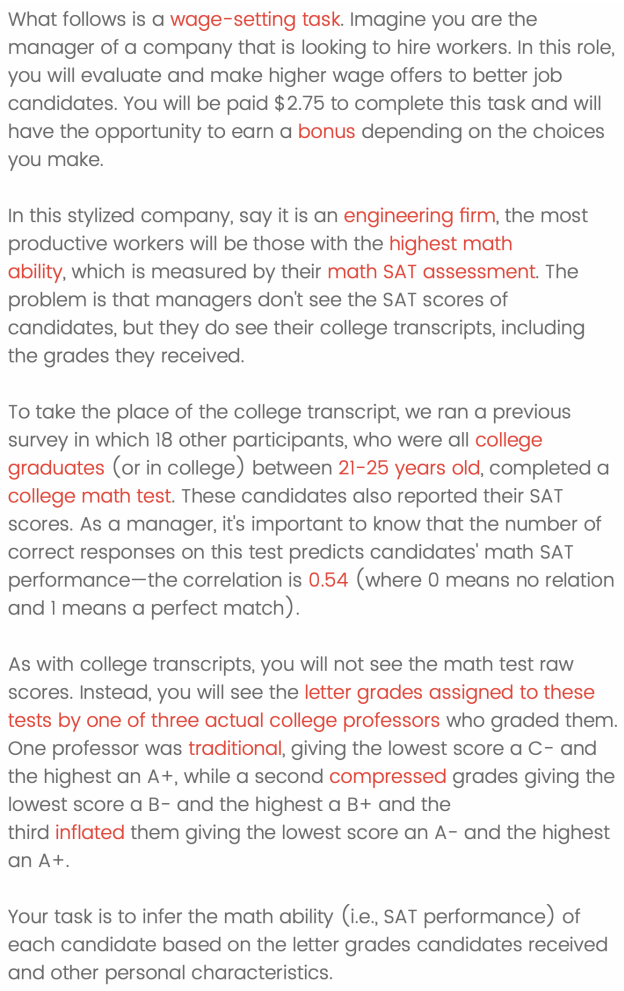


Figure B10: Manager initial instructions.

The final stage of the manager experiment asked respondents to set incentive compatible wages that match the underlying ability of the same nine candidate, assessed (again) one at a time. The instructions in Figure B20 informed managers that the wages would use the same B-D-M procedure as the signal extraction part of the experiment. Figure B21 provides an example of the information managers saw when setting wages.

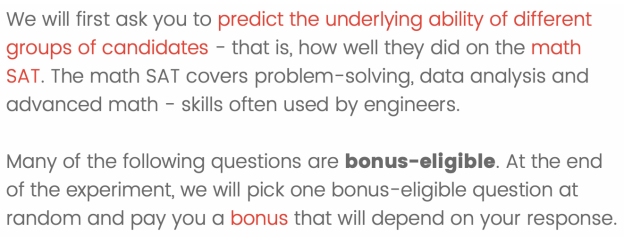


Figure B11: Manager prior and bonus instructions.

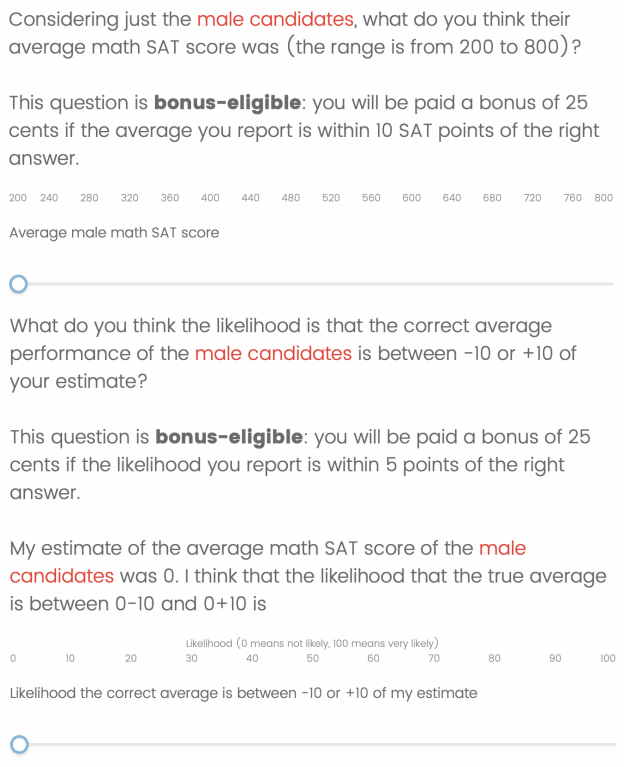


Figure B12: Manager prior elicitation.

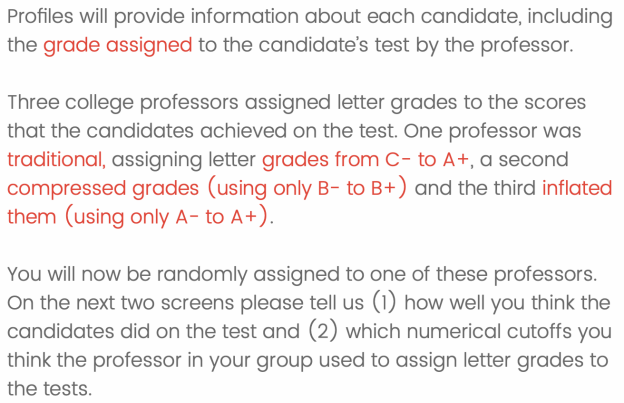
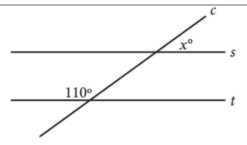


Figure B13: Manager grade scheme instructions.

Here are three randomly selected questions from the candidate math test to give you a sense of the content and its difficulty. Please read through them and then respond to the questions below.

The total cost, in dollars, to rent a surfboard consists of a $25 service fee and a $10 per hour rental fee. A person rents a surfboard for t hours and intends to spend a maximum of $75 to rent the surfboard. Which inequality represents this situation?

- $10t \leq 75$
- $10 + 25t \leq 75$
- $25t \leq 75$
- $25 + 10t \leq 75$



Note: Figure not drawn to scale.

In the figure shown, line c intersects parallel lines s and t. What is the value of x ?

Each face of a fair 14-sided die is labeled with a number from 1 through 14, with a different number appearing on each face. If the die is rolled one time, what is the probability of rolling a number greater than 10?

- 1/7
- 1/14
- 2/7
- 3/14

Reminder: the 18 candidates were between 21 and 25 and have completed college (or are currently enrolled in college). Given the sample questions above, what do you think are the lowest, the highest, and the average scores (0-100) the candidates achieved on the test?

This question is **bonus-eligible**: you will be paid a bonus of 25 cents if your responses are within 3 percentage points of the right answers.

Candidate test score out of 100 percent

0 10 20 30 40 50 60 70 80 90 100

Lowest score achieved

Highest score achieved

Average score achieved

Figure B14: Manager's description of the candidate test and performance beliefs.

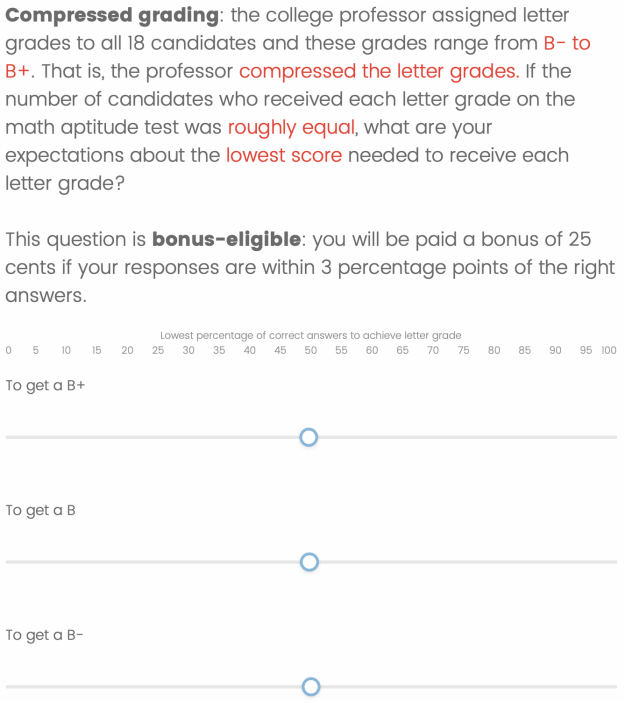


Figure B15: Manager expected grading cutoffs (Compress treatment).

Now we will ask you to review the profiles of 9 of the 18 candidates (selected at random) and assign them **scores** and **wages** in two stages.

Testing stage: in the first stage, you will see what letter grades the professor assigned each candidates' math test and you will guess the actual scores (between 0 and 100) that the candidates achieved. You can earn a **bonus** (details explained next) that will depend on how well your guesses match the actual scores of the candidates.

After assigning scores, you will be eligible to receive a bonus, the amount of which will depend on three things:

- the score you guess for one randomly selected candidate,
- how that candidate actually performed on the test,
- and a randomly generated "threshold" between 0 and 100.

Specifically, we will first give you a budget of 100 cents to hire a candidate and

- if the score you guess is less than the threshold, you do not hire the candidate and your bonus equals your budget of 100 cents.
- if the score you guess is greater than or equal to the threshold, you hire the candidate and your bonus is your budget of 100 cents minus the threshold, plus the score the candidate received on the test.

Here are two examples.

- Example 1: say you think the score is 85, the threshold is 50 and the candidate correctly answered 80 percent of the test questions. Because 85 > 50, the worker is hired and your bonus is 100 - 50 + 80 = 130 cents.
- Example 2: say you think the score is 60, the threshold is 50 and the candidate correctly answered 40 percent of the test questions. Because 60 > 50, the worker is hired and your bonus is 100 - 50 + 40 = 90 cents.

Most importantly, this procedure gives you an incentive to guess scores equal to how well the candidates actually did on the test.

Figure B16: Manager signal extraction B-D-M instructions.

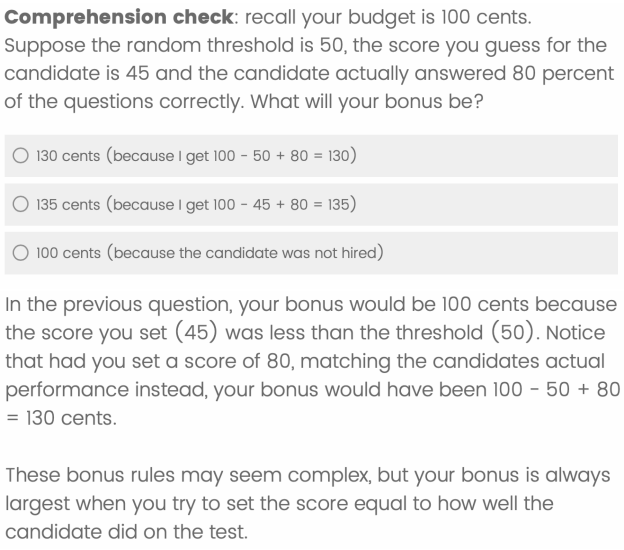


Figure B17: Manager signal extraction B-D-M understanding check.

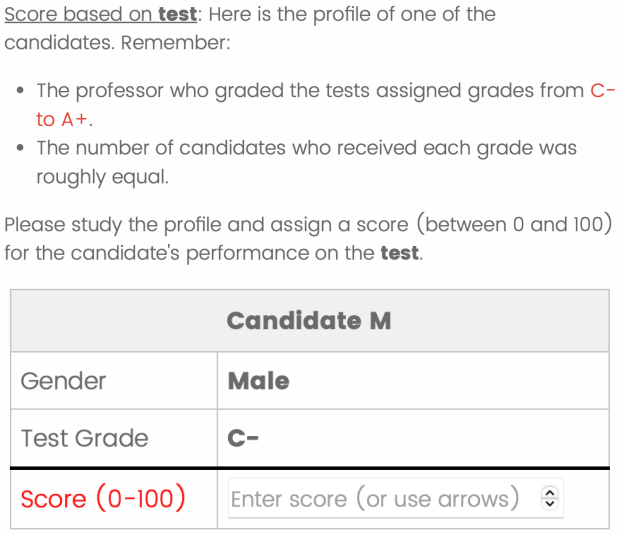


Figure B18: Manager signal extraction elicitation.

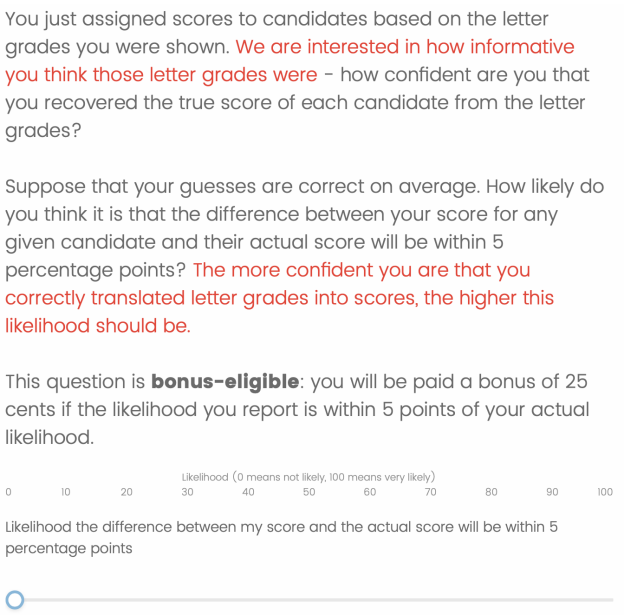


Figure B19: Manager signal extraction error variance elicitation.

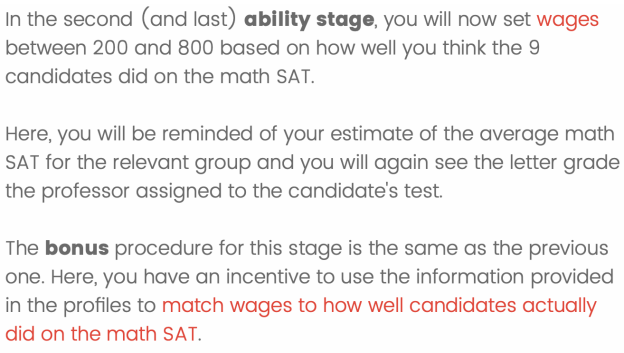


Figure B20: Manager wage instructions.

Wage based on **math SAT** (ability): Here is the profile of one of the candidates. Remember and consider:

- The professor assigned grades from C- to A+.
- Each grade was given to a similar number of candidates.
- Your estimate of how well men did on the math SAT.
- The correlation between test scores and the math SAT.

Please study the expanded profile, and assign a wage (between 200 and 800) for the candidate's ability (i.e., **math SAT**).

Candidate M	
Correlation between test and SAT	**0.54**
Test Grade (between C- and A+)	**C-**
Gender	**Male**
Your estimate of Male Math SAT	**${q://QID393/TotalSum}**
Wage (200-800)	Enter wage (or use arrows)

Figure B21: Manager wage elicitation.